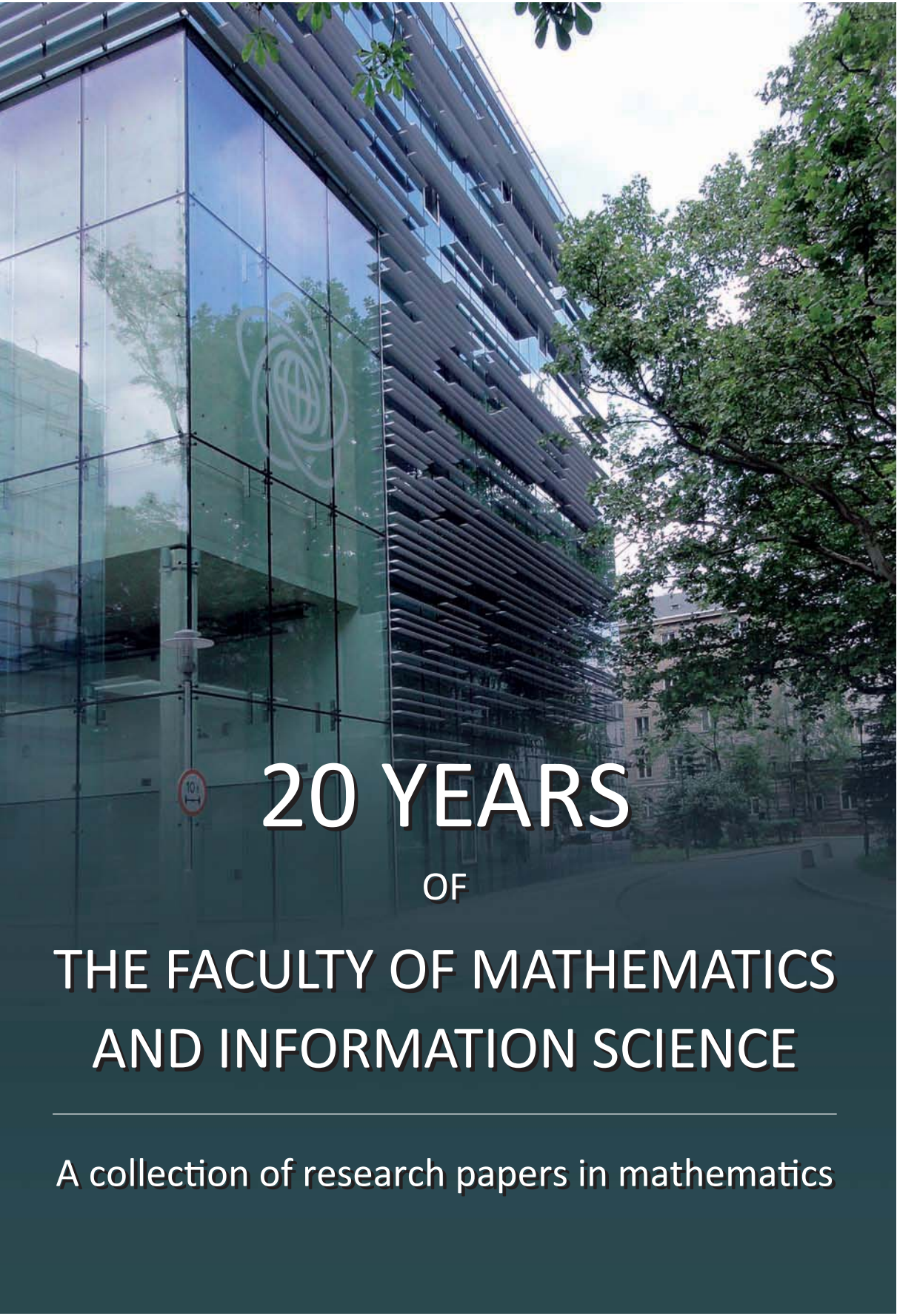


20 YEARS

OF

THE FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

A collection of research papers in mathematics

20 YEARS

OF

THE FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

A collection of research papers in mathematics

20 YEARS

OF

THE FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

A collection of research papers in mathematics

WARSAW 2020

Reviewers:

Krzysztof Chełmiński
Konstanty Junosza-Szaniawski
Bogusława Karpińska
Janina Kotus
Wojciech Matysiak
Tomasz Miller
Mariusz Niewęgłowski
Marek Rutkowski
Ewa Zadrzyńska-Piętka
Michał Ziembowski

Editor:

Janina Kotus

Language Editor:

Tomasz Traczyk

Typesetting:

Łukasz Błaszczyk

Cover design by

Danuta Czudek-Puchalska

Cover photograph — *Anna Agata Wagner*

Publisher: Warsaw University of Technology Press
(Oficyna Wydawnicza Politechniki Warszawskiej – UIW 48800)
Polna 50, 00-644 Warsaw, Poland, phone (48) 22 234 70 83

Internet bookstore of the Warsaw University of Technology Press (OWPW):
http://www.wydawnictwopw.pl; e-mail: oficyna@pw.edu.pl
phone (48) 22 234 75 03; fax (48) 22 234 70 60

ISBN 978-83-8156-156-3

Warsaw University of Technology Press (Oficyna Wydawnicza Politechniki Warszawskiej)
Polna 50, 00-644 Warsaw, Poland, phone (48) 22 234 70 83
Printed and bounded by OWPW, phone (48) 22 234 70 30, first edition, order no 223/2020

Contents

Preface

The history of mathematics at the Warsaw University of Technology goes back to 1826 when the Preparatory School for the Polytechnic Institute was founded thanks to the efforts of Stanisław Staszic. Its first director became Kajetan Garbiński, a professor of mathematics. The school was closed in 1831.

The Warsaw Polytechnic Institute named after Tsar Nicolas II was establihed in 1898. Classes were conducted in Russian untill the outbreak of World War I. The Warsaw University of Technology started on its own in 1915. It was the first Polish technical university. All this time at faculties of engineering there were divisions of mathematics which employed famous professors including Georgij Voronoj, Kazimierz Żorawski, Witold Pogorzelski, Stanisław Saks, Antoni Zygmund, Franciszek Leja, Władysław Nikliborc, Stefan Straszewicz and Roman Sikorski.

In 1963 all the divisions of mathematics were joined together in order to establish the Institute of Mathematics, which in 1975 became a part of the Faculty of Technical Physics and Applied Mathematics. In 1999 the institute was transformed into the Faculty of Mathematics and Information Sciences.

The aim of this monograph is to celebrate 20 years of the Faculty of Mathematics and Information Science. We present a collection of research papers written by mathematicians representing various generations, from assistant professors to full professors, currently employed at the faculty. They cover many areas of mathematics including algebraic structures, analysis on manifolds, control theory, differential geometry, dynamical systems, general geometry, graph theory, mathematical statistics, numerical analysis, partial differential equations and stochastic analysis.

Konstancja Bobecka[1], **Jacek Wesołowski**[1,2]

[1] Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
[2] Central Statistical Office, Warsaw, Poland

NON-ADMISSIBILITY OF THE RUBIN ESTIMATOR OF THE VARIANCE IN MULTIPLE IMPUTATION IN THE BAYESIAN GAUSSIAN MODEL

Manuscript received: 18 June 2020
Manuscript accepted: 22 July 2020

Abstract: Multiple imputation is nowadays a generally accepted approach to statistical inference based on incomplete data sets. Within this methodology it is standard to assess the quality of the estimation by the Rubin estimator of the variance, which, when based on m imputations, has the form $\bar{U}_m + (1+1/m)B_m$. Here $\bar{U}_m$ is the average of imputation estimators of variance and B_m is the empirical variance of imputation estimators. We consider the problem of estimation of variance of multiple imputation estimator in the Bayesian Gaussian model with the Gaussian mean. We show that the Rubin estimator is inadmissible in the class of estimators of the form $v^2(\alpha,\beta) = \alpha\bar{U}_m + \beta B_m,\ \alpha,\beta \in \mathbb{R}$. We derive the optimal weights α_* and β_*, i.e. such that $v^2(\alpha_*,\beta_*)$ has the smallest MSE in this class of estimators. Since α_* and β_* are defined through complicated expressions we also derive approximate optimal estimators with simple weights $\alpha_{**} = \frac{1}{f}$, $\beta_{**} = -\frac{f}{n(1-f)}$, where f is the response rate and n is the original size of the sample. These estimators outperform the Rubin estimator with respect to both the bias and the MSE. We also consider the case of a non-informative prior. Then the Rubin estimator is unbiased, though it remains inadmissible. Numerical experiments show that the performance of the optimal and the approximate optimal estimators is rather similar, therefore we recommend to use simplified approximate weights.

Keywords: multiple imputation, Rubin estimator, Bayesian Gaussian model
Mathematics Subject Classification (2020): 62D10, 62F15

1. INTRODUCTION

The methodology of multiple imputation proposed in Rubin (1987) is nowadays one of the most frequently used approaches to missing data problems. The basic idea lies in creating, instead of one imputation sample, a larger number m of imputation samples. For each of such samples an imputation estimator $\hat{\theta}_{Imp}^{(\ell)}$ is designed according to the same rule, $\ell = 1,\ldots,m$. The final estimator is the average $\hat{\theta}_{MImp} = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}_{Imp}^{(\ell)}$. Typically the variance of $\hat{\theta}_{MImp}$ is estimated by the Rubin estimator:

$$v_{Rub}^2 = \bar{U}_m + \left(1+\tfrac{1}{m}\right) B_m,$$

where $\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{V}_{Imp}^{(\ell)}$ is the average of the imputation estimators $\hat{V}_{Imp}^{(\ell)}$ of the variance of $\hat{\theta}_{Imp}^{(\ell)}$, $\ell = 1,\ldots,m$, and $B_m = \frac{1}{m-1}\sum_{\ell=1}^{m}\left(\hat{\theta}_{Imp}^{(\ell)} - \hat{\theta}_{MImp}\right)^2$ is the empirical variance of the single imputation estimators. Though v_{Rub}^2 was introduced in the Bayesian context, it is widely used in applications for all kinds of data. In this aspect there is some criticism of the Rubin estimator in the literature mostly concerned with analysis of its bias, see e.g. Fay (1992), Kim (2004), Kim, Brick, Fuller, Kalton (2006), von Hippel (2013), Wang and Robins (1998), Robins and Wang (2000), Nielsen (2003), von Hippel (2007), Hughes, Sterne and Tilling (2016), as well as with the optimal choice of the number of imputations, see e.g. von Hippel (2005), Graham, Olchowski, Gilreath (2007), Bodner (2008).

We analyze variance estimation when the procedure of multiple imputation is applied to the mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the standard estimator of its variance, $\frac{S^2}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$. We consider the Bayesian Gausian model with the Gaussian distribution for the mean and unknown variance (the case of non-informative prior is also studied). We design a natural imputation scheme based on conditional distribution of $\mathbf{X}_{R^c}|\mathbf{X}_R$, where R and R^c are respectively, observed and unobserved part of the sample $\mathbf{X}$ of size n. In this scheme we introduce a class of the Rubin-type estimators of variance and investigate its properties. In particular, we derive the optimal estimator within this class.

Multiple imputation for different Gaussian models have been already considered in the literature, see e.g. von Hippel (2013a, b), Di Zio and Guarnera (2008). However, to the

best of our knowledge, no results on optimality of the variance estimation are available. In general, it may not be feasible since it involves expressions for moments of the fourth order which typically are hard to handle. But in some special models, as the Bayesian Gaussian model with random Gaussian mean, we analyze here, such formulas are available. In this model we study the optimal estimator of the variance of the multiple imputation estimator in the class $\mathfrak{R} = \{\alpha\bar{U}_m + \beta B_m,\ \alpha,\beta \in \mathbb{R}\}$ of the Rubin-type estimators. We derive optimal coefficients α and β and show that the Rubin estimator is not only biased but also inadmissible. Precise expressions for optimal α and β, we derive, are quite complicated (though explicit) functions of the number of imputations m, the original sample size n and the response rate f. Therefore we also propose a simplified version of optimal coefficients of the form $\alpha_{**} = \frac{1}{f}$ and $\beta_{**} = -\frac{f}{n(1-f)}$ (for large n and $m \to \infty$). We also compare asymptotic properties as ($m \to \infty$ and n is arbitrary) of the optimal estimator and the Rubin estimator.

The paper is organized as follows: In Section 2 basic properties of single imputation in the Bayesian Gaussian scheme are derived. This gives a base for analyzing, in Section 3, multiple imputation in this model. Section 4 is devoted to study properties of the Rubin-type variance estimators. In particular, it contains our main results in which we give the optimal and approximate optimal estimators both for informative and non-informative priors. We also obtain optimal unbiased estimators of the Rubin-type in the model with non-informative prior. Additionally, in this section we analyze properties of these estimators when number of imputations is large. Section 5 is for conclusions. All proofs are in the Appendix.

2. SINGLE IMPUTATION

Let $(X_1,\ldots,X_n,M)$ be a random vector with conditional distribution of $\mathbf{X} = (X_1,\ldots,X_n)$ given M of the form

$$\mathbf{X}|M = \left(\mathrm{N}(M,\sigma^2)\right)^{\otimes n},$$

that is, conditionally on M the components of $\mathbf{X}$ are iid normal with the mean M and (unknown) variance σ^2. Moreover, the distribution of M is normal $\mathrm{N}(\mu, \kappa\sigma^2)$, where $\mu \in \mathbb{R}$ and $\kappa > 0$ are (known) hyperparameters. We refer to this model by $\mathrm{GmG}(\mu, \sigma^2, \kappa)$, the "Gaussian-mean-Gaussian" model with parameters μ, σ^2, κ.

Alternatively,

$$X_i = M + \sigma Z_i, \quad i = 1, \ldots, n,$$

where $Z_1, \ldots, Z_n$ are iid standard normal random variables and $(Z_1, \ldots, Z_n)$ and M (defined above) are independent.

Let $R \subset \{1, \ldots, n\}$, $\#(R) = r$, be the set of labels of those X_i's which are observed, that is $\mathbf{X}_R = (X_i, i \in R)$ is the observed and $\mathbf{X}_{R^c} = (X_i, i \in R^c)$ is the missing part of the sample $\mathbf{X}$. For future reference by $f = r/n$ we denote the response rate. Missing variables are replaced by imputed ones $\tilde{X}_i$, $i \in R^c$. Thus, the sample after imputation, $\mathbf{X}_{Imp} = (\tilde{X}_1, \ldots, \tilde{X}_n)$, has the form

$$\tilde{X}_i = \begin{cases} X_i, & i \in R, \\ \tilde{X}_i, & i \in R^c. \end{cases}$$

Hence the imputation versions of estimators $\bar{X}$ and S^2 are

$$\bar{X}_{Imp} = \tfrac{1}{n} \sum_{i=1}^{n} \tilde{X}_i, \qquad S^2_{Imp} = \tfrac{1}{n-1} \sum_{i=1}^{n} (\tilde{X}_i - \bar{X}_{Imp})^2.$$

It is well known that in $\mathrm{GmG}(\mu, \sigma^2, \kappa)$ model the conditional distribution of unobserved $\mathbf{X}_{R^c}$ given observed $\mathbf{X}_R$ is $(n-r)$-dimensional Gaussian

$$\mathbf{X}_{R^c} | \mathbf{X}_R \sim \mathrm{N}\left(\tfrac{r\kappa\bar{X}_R + \mu}{r\kappa + 1} \mathbf{1}_{R^c}, \, \sigma^2 (\mathbf{I}_{R^c} + \tfrac{\kappa}{r\kappa+1} \mathbf{1}_{R^c} \mathbf{1}_{R^c}^T) \right),$$

where $\bar{X}_R = \frac{1}{r} \sum_{i \in R} X_i$, $\mathbf{1}_{R^c} \in \mathbb{R}^{n-r}$ is a vector of 1's and $\mathbf{I}_{R^c}$ is an $(n-r) \times (n-r)$ identity matrix.

Consequently, $\mathbf{X}_{R^c}$ has the representation

$$\mathbf{X}_{R^c} = \tfrac{r\kappa\bar{X}_R + \mu}{r\kappa + 1} \mathbf{1}_{R^c} + \sigma \mathbf{W},$$

where

$$\mathbf{W} = (W_i, i \in R^c) = \mathbf{Z} + \sqrt{\tfrac{\kappa}{r\kappa+1}} U \mathbf{1}_{R^c},$$

$\mathbf{Z} = (Z_i, i \in R^c)$ is a vector of iid standard normal random variables, U is a standard normal random variable and $(\mathbf{Z}, U, \mathbf{X}_R)$ are jointly independent.

Since the standard unbiased estimator of σ^2 based on the observed part of the sample is $S_R^2 = \frac{1}{r-1}\sum_{i\in R}(X_i - \bar{X}_R)^2$, provided $r > 1$, it is natural to impute missing values by

$$\tilde{X}_j = \frac{r\kappa\bar{X}_R+\mu}{r\kappa+1} + S_R W_j, \quad j \in R^c. \tag{1}$$

Consequently, the imputed sample has the form

$$\mathbf{X}_{Imp} = \left(X_i, i \in R,\ \frac{r\kappa\bar{X}_R+\mu}{r\kappa+1} + S_R W_j,\ j \in R^c\right).$$

Theorem 1. *In* $\mathrm{GmG}(\mu, \sigma^2, \kappa)$ *model with imputed values defined in* (1) *the imputation version of the sample mean is*

$$\bar{X}_{Imp} = f\,\frac{n\kappa+1}{r\kappa+1}\,\bar{X}_R + (1-f)\left(\frac{1}{r\kappa+1}\,\mu + S_R\bar{W}\right), \tag{2}$$

where $\bar{W} = \bar{Z} + \sqrt{\frac{\kappa}{r\kappa+1}}\,U$ *and* $\bar{Z} = \frac{1}{n-r}\sum_{i\in R^c} Z_i$.

The estimator $\bar{X}_{Imp}$ *is unbiased, i.e.* $\mathbb{E}\bar{X}_{Imp} = \mathbb{E}M = \mu$. *Its variance is*

$$\mathbb{V}\mathrm{ar}\bar{X}_{Imp} = \frac{\sigma^2}{n}(n\kappa+1) \tag{3}$$

and its MSE has the form

$$\mathrm{MSE}\bar{X}_{Imp} = \mathbb{E}(\bar{X}_{Imp} - M)^2 = \frac{\sigma^2}{n}\left(1 + 2\frac{(n-r)\kappa}{r\kappa+1}\right). \tag{4}$$

Now we consider the imputation version of $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

Theorem 2. *In* $\mathrm{GmG}(\mu, \sigma^2, \kappa)$ *model with imputed values defined in* (1) *the imputation version of the sample variance is*

$$S^2_{Imp} = \frac{1}{n-1}\left\{S_R^2\left(r-1+(n-r-1)S_Z^2\right) + r(1-f)\left(\frac{\bar{X}_R-\mu}{r\kappa+1} - S_R\bar{W}\right)^2\right\}. \tag{5}$$

It is an unbiased estimator of σ^2 *and*

$$\mathbb{V}\mathrm{ar}S^2_{Imp} = \frac{2\sigma^4}{(n-1)^2(r-1)}\left\{(r-1)(n-r-2) + n(n-2) + \tau^2\left(2(n-2)+3\tau^2\right)\right\}, \tag{6}$$

where $\tau^2 = \frac{n\kappa+1}{r\kappa+1}f$.

THE CASE OF NON-INFORMATIVE PRIOR

Consider now the special situation of non-informative prior distribution of M. This is formally realized by taking the limit $\kappa \to \infty$ in the previous considerations.

Therefore in the case of non-informative prior we impute the missing variables according to the formula

$$\tilde{X}_j = \bar{X}_R + S_R\left(Z_j + \frac{U}{\sqrt{r}}\right), \quad j \in R^c,$$

and thus the imputed sample has the form

$$\mathbf{X}_{Imp} = \left(X_i,\, i \in R,\, \bar{X}_R + S_R\left(Z_j + \frac{U}{\sqrt{r}}\right),\, j \in R^c\right).$$

Consequently, the imputation version of the sample mean, see (2), is

$$\bar{X}_{Imp} = \bar{X}_R + (1-f)S_R\left(\bar{Z} + \frac{U}{\sqrt{r}}\right).$$

The MSE of $\bar{X}_{Imp}$ has the form, see (4),

$$\mathrm{MSE}\bar{X}_{Imp} = \mathbb{E}(\bar{X}_{Imp} - M)^2 = \frac{\sigma^2}{r}(2-f).$$

Note that $\lim_{\kappa\to\infty} \tau^2 = 1$. Therefore, (5) yields

$$S^2_{Imp} = \frac{1}{n-1}S^2_R\left\{r - 1 + (n-r-1)S^2_Z + r(1-f)\left(\bar{Z} + \frac{U}{\sqrt{r}}\right)^2\right\}$$

and (6) implies

$$\mathbb{V}\mathrm{ar}\, S^2_{Imp} = \frac{2\sigma^4}{(n-1)^2(r-1)}\{(r+1)(n-r) + (n-1)^2\}.$$

3. MULTIPLE IMPUTATION

In multiple imputation several, say m, imputed samples $\mathbf{X}^{(\ell)}_{Imp} = (\tilde{X}^{(\ell)}_i, i = 1,\ldots,n)$, $l = 1,\ldots,m$, are created in such a way that random vectors $\tilde{\mathbf{X}}^{(\ell)}_{R^c} = (\tilde{X}^{(\ell)}_i, i \in R^c)$, $\ell = 1,\ldots,m$,

are conditionally independent given the observed part of the sample $\mathbf{X}_R = (X_i, i \in R)$ which is common in all imputed samples. Having these samples defined we consider respective imputation estimators of the sample mean, $\bar{X}_{Imp}^{(\ell)}$ and of the sample variance, $(S_{Imp}^{(\ell)})^2$, $\ell = 1, \ldots, m$. The multiple imputation estimator of the mean is

$$\bar{X}_{MImp} = \frac{1}{m} \sum_{\ell=1}^{m} \bar{X}_{Imp}^{(\ell)}. \tag{7}$$

Let us emphasize that this is the case of the proper multiple imputation procedure since any Bayesian multiple imputation is proper if only the complete data estimator is the MLE - see Nielsen (2003) - and this is the case of the empirical mean in the Gaussian model.

The Rubin estimator of the variance of $\bar{X}_{MImp}$ is defined as

$$v_{Rub}^2 = \bar{U}_m + \frac{m+1}{m} B_m,$$

where

$$\bar{U}_m = \frac{1}{mn} \sum_{\ell=1}^{m} \left(S_{Imp}^{(\ell)} \right)^2 \tag{8}$$

and

$$B_m = \frac{1}{m-1} \sum_{\ell=1}^{m} \left(\bar{X}_{Imp}^{(\ell)} - \bar{X}_{MImp} \right)^2. \tag{9}$$

In GmG(μ, σ^2, κ) model imputed samples have the form

$$\mathbf{X}_{Imp}^{(\ell)} = \left(X_i, i \in R, \ \frac{r\kappa \bar{X}_R + \mu}{r\kappa + 1} + S_R W_j^{(\ell)}, \ j \in R^c \right), \quad \ell = 1, \ldots, m, \tag{10}$$

with

$$W_j^{(\ell)} = Z_j^{(\ell)} + \sqrt{\frac{\kappa}{r\kappa+1}} U^{(\ell)}, \quad j \in R^c,$$

where $Z_j^{(\ell)}$, $j \in R^c$, $U^{(\ell)}$, $\ell = 1, \ldots, m$, are iid standard normal random variables.

It is easy to see that

$$\bar{W}^{(\ell)} = \bar{Z}^{(\ell)} + \sqrt{\frac{\kappa}{r\kappa+1}} U^{(\ell)}, \quad \ell = 1, \ldots, m,$$

are iid normal random variables with zero mean and variance $\tilde{\tau}^2 = \frac{n\kappa+1}{(r\kappa+1)(n-r)}$.

Theorem 3. *In* $\mathrm{GmG}(\mu,\sigma^2,\kappa)$ *model with imputation defined in* (10)*, the multiple imputation estimator of M has the form*

$$\bar{X}_{MImp} = f\,\tfrac{n\kappa+1}{r\kappa+1}\bar{X}_R + (1-f)\left(\tfrac{1}{r\kappa+1}\mu + S_R\bar{\bar{W}}\right), \tag{11}$$

where $\bar{\bar{W}} = \frac{1}{m}\sum_{\ell=1}^{m}\bar{W}^{(\ell)}$.

Statistics B_m *and* $\bar{U}_m$ *defined in* (9) *and* (8)*, respectively, assume the form:*

$$B_m = (1-f)^2 S_R^2 S_{\bar{W}}^2 \tag{12}$$

and

$$\bar{U}_m = \tfrac{1}{n(n-1)}\left\{S_R^2\left[r-1+(n-r-1)\bar{S}_Z^2\right] + r(1-f)\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1} - S_R\bar{\bar{W}}\right)^2 + \tfrac{r}{1-f}\tfrac{m-1}{m}B_m\right\}, \tag{13}$$

where

$$\bar{S}_Z^2 = \tfrac{1}{m}\sum_{\ell=1}^{m} S_{Z^{(\ell)}}^2 \quad \textit{and} \quad S_{\bar{W}}^2 = \tfrac{1}{m-1}\sum_{\ell=1}^{m}\left(\bar{W}^{(\ell)} - \bar{\bar{W}}\right)^2.$$

Theorem 4. *The estimator* $\bar{X}_{MImp}$ *is ubiased for M and its MSE has the form*

$$\mathrm{MSE}\,\bar{X}_{MImp} = \mathbb{E}\,(\bar{X}_{MImp}-M)^2 = \left(\tfrac{n\kappa+f}{n\kappa+1} + \tfrac{1-f}{m}\right)\tfrac{\tau^2\sigma^2}{r}. \tag{14}$$

Moreover,

$$\mathbb{E}\,\bar{U}_m = \frac{\sigma^2}{n}, \tag{15}$$

and

$$\mathbb{E}\,B_m = (1-f)\,\frac{\tau^2\sigma^2}{r} \tag{16}$$

and the Rubin estimator v_{Rub}^2 *is biased with the bias*

$$\mathbb{B}\,v_{Rub}^2 = \mathbb{E}\,v_{Rub}^2 - \mathrm{MSE}\,\bar{X}_{MImp} = \tfrac{2(1-f)}{n\kappa+1}\,\tfrac{\tau^2\sigma^2}{r}. \tag{17}$$

Remark 5. *Note that the relative bias of the Rubin estimator has the form*

$$\frac{\mathbb{B}\,v_{Rub}^2}{\mathrm{MSE}\,\bar{X}_{MImp}} = \frac{2(1-f)}{n\kappa+f+\frac{1}{m}(1-f)(n\kappa+1)} < \tfrac{2(1-f)}{n\kappa+f}. \tag{18}$$

Therefore, in the case of non-informative prior, that is when $\kappa\to\infty$*, we see that the Rubin estimate* v_{Rub}^2 *is unbiased for* $\mathrm{MSE}\,\bar{X}_{MImp}$ *which in this case (i.e. when* $\kappa\to\infty$*) assumes the form*

$$\mathrm{MSE}\,\bar{X}_{MImp} = \mathbb{E}\,(\bar{X}_{MImp}-M)^2 = \tfrac{\sigma^2}{r}\left(1+\tfrac{1-f}{m}\right).$$

Note also that $\frac{2(1-f)}{n\kappa+f}$*, the right-hand side of* (18)*, is the supremum over m of the relative bias for n, f and* κ *given (actually, it is its limit when* $m \to \infty$*). For* $\delta < \frac{2(1-f)}{n\kappa+f}$ *it follows that if the number of imputed samples m satisfies*

$$m < \frac{\delta(1-f)(n\kappa+1)}{2(1-f)-\delta(n\kappa+f)}$$

then the relative bias remains below the level δ*.*

4. OPTIMAL RUBIN-TYPE ESTIMATOR OF THE VARIANCE OF MULTIPLE IMPUTATION ESTIMATOR

In this section we consider estimator of the MSE of $\bar{X}_{MImp}$ in the class $\mathfrak{R}$ of generalized Rubin estimators of the form

$$v^2(\alpha,\beta) = \alpha\bar{U}_m + \beta B_m, \qquad \alpha,\beta \in \mathbb{R}. \tag{19}$$

Note that the Rubin estimator v^2_{Rub} belongs to class $\mathfrak{R}$ with $\alpha = 1$ and $\beta = 1 + \frac{1}{m}$.

Observe that the coefficients of the Rubin estimator do not depend on the response rate f. As it will be shown, the optimal coefficients do depend on f. In this context it is worth to mention that Bjørnstad (2007) (accompanied by a discussion in Skinner (2007)) suggested a modification of the Rubin estimator v^2_{Rub} by incorporating f in the coefficient β of B_m as follows: $v^2_{Bjo} = v^2(1, \frac{1}{1-f} + \frac{1}{m})$. Actually, a more general form $\beta = k + \frac{1}{m}$ was considered and then the approximate condition $\mathbb{V}ar\,\hat{\theta}_{MImp} \approx \mathbb{E}\bar{U}_m + (k + \frac{1}{m})\mathbb{E}B_m$ allowed to conclude that $k = \frac{1}{1-f}$. Nevertheless, the optimality of v^2_{Bjo} was not analyzed there. For a comparison of v^2_{Bjo} with the Rubin estimator see Laaksonen (2016a,b).

The aim of this section is to find optimal weights α,β, i.e. such that the estimator (19) has the smallest MSE in the class $\mathfrak{R}$. We will also compare the optimal estimator in the class $\mathfrak{R}$ with the Rubin estimator v^2_{Rub} and the Bjørnstad estimator v^2_{Bjo}.

The basic auxiliary characteristics for this kind of study are variances and covariances of $\bar{U}_m$ and B_m.

Proposition 6. *In* $\mathrm{GmG}(\mu,\sigma^2,\kappa)$ *model*

$$\mathbb{V}\mathrm{ar}\,\bar{U}_m = \frac{2\sigma^4}{(r-1)n^2(n-1)^2}\left[(r-1)\left(1-\tau^2\right)(1+\tfrac{2-m}{m}\tau^2)+\left(n-2+\tau^2\right)^2+\frac{(r+1)\left(n-r-1+\tau^4\right)}{m}\right], \tag{20}$$

$$\mathbb{V}\mathrm{ar}\,B_m = \frac{2\sigma^4\tau^4(1-f)^2}{r^2(r-1)}\left(1+\tfrac{r+1}{m-1}\right), \tag{21}$$

and

$$\mathbb{C}ov(\bar{U}_m,B_m) = \frac{2\sigma^4\tau^2(1-f)}{r(r-1)n(n-1)}\left[n-2+\tau^2\left(1+\tfrac{r+1}{m}\right)\right]. \tag{22}$$

Remark 7. *From* (20) - (22) *and* (15), (16) *it follows that*

$$\begin{bmatrix}\mathbb{E}\,\bar{U}_m^2\\ \mathbb{E}\,B_m^2\\ \mathbb{E}\,\bar{U}_m B_m\end{bmatrix} = \frac{\sigma^4(r+1)}{(r-1)n^2}\begin{bmatrix}a\\ b\\ c\end{bmatrix}, \tag{23}$$

where

$$\begin{aligned}
a &= 1+\tfrac{2n(1-f)}{(n-1)^2m}+\frac{2(1-\tau^2)\left(2\left(r+1-n-\frac{r-1}{m}\right)+\left(\frac{r-3}{m}-r\right)(1+\tau^2)\right)}{(r+1)(n-1)^2},\\
b &= \frac{\tau^4(1-f)^2}{f^2}\,\tfrac{m+1}{m-1},\\
c &= \frac{\tau^2(1-f)}{f}\left(1+\tfrac{2\tau^2}{m(n-1)}+\tfrac{2(\tau^2-1)}{(r+1)(n-1)}\right).
\end{aligned}$$

Theorem 8. *Let*

$$\alpha_* = \tfrac{2n(r-1)}{r(r+1)}\tfrac{A_2A_4}{A_1} \qquad \textit{and} \qquad \beta_* = \tfrac{2(r-1)}{(n-1)^2(r+1)(1-f)\tau^2}\tfrac{A_3A_4}{A_1}, \tag{24}$$

where

$$A_1 = a\,\tfrac{m+1}{m-1}-\left(1+\tfrac{2\tau^2}{m(n-1)}+\tfrac{2(\tau^2-1)}{(r+1)(n-1)}\right)^2, \tag{25}$$

$$A_2 = \tfrac{1}{m-1}-\tfrac{\tau^2}{m(n-1)}+\tfrac{1-\tau^2}{(r+1)(n-1)}, \tag{26}$$

$$A_3 = \tfrac{\tau^2-r}{m}+(1-\tau^2)\left(\tfrac{n}{m}+\frac{2r+1-n-2\frac{r-1}{m}+\left(\frac{r-3}{m}-r\right)(1+\tau^2)}{r+1}\right), \tag{27}$$

$$A_4 = \tau^2+\tfrac{(1-f)\tau^2}{m}-(1-\tau^2)f. \tag{28}$$

Then $v^2(\alpha_*,\beta_*)$ *has the smallest MSE among the estimators of the MSE of* $\bar{X}_{MImp}$ *from the class* $\mathfrak{R}$.

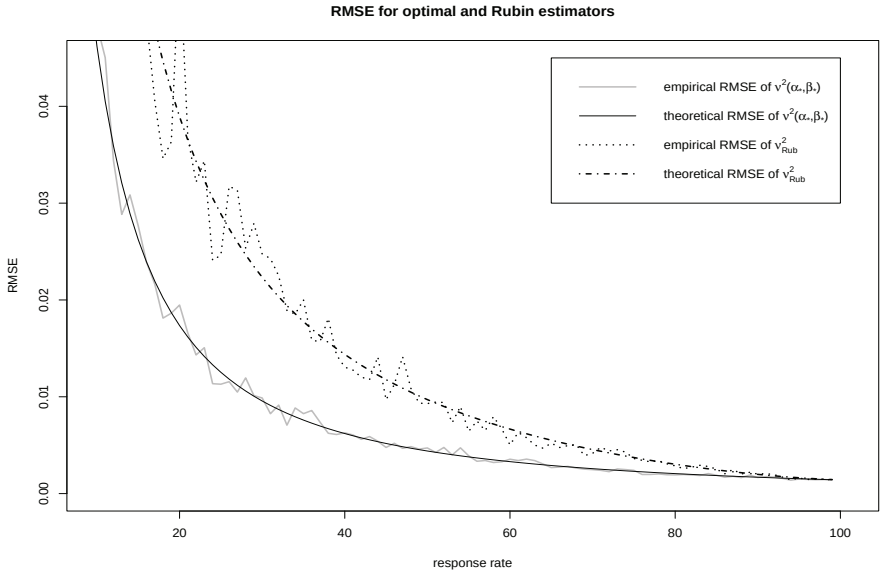


Fig. 1. Theoretical and empirical RMSE of the optimal estimator $v^2(\alpha_*, \beta_*)$ and the Rubin estimator v^2_{Rub}. Here $m = 5$, $n = 100$, $\sigma^2 = 1$, $\mu = 0$ and $\kappa = 1$. The empirical versions are computed from 100 repetitions

The optimal MSE is

$$\operatorname{MSE} v^2(\alpha_*, \beta_*) = \frac{\sigma^4}{r^2} A_4 \left(A_4 - \alpha_* f - \beta_* (1-f) \tau^2 \right). \tag{29}$$

Remark 9. *A comparison between the RMSE's (root MSE) of the optimal estimator $v^2(\alpha_*, \beta_*)$ and the Rubin estimator v^2_{Rub} is illustrated in Fig. 1 (with a close-up for high response rates in Fig. 2). The difference is larger for smaller response rates.*

4.1. THE CASE OF NON-INFORMATIVE PRIOR

The non-informative prior is the case when $\kappa \to \infty$ (which is equivalent to $\tau^2 \to 1$). The model we consider is denoted as $\mathrm{GmG}(\mu, \sigma^2, \infty)$. Then the optimal coefficients $\alpha_{*,\infty}$ and

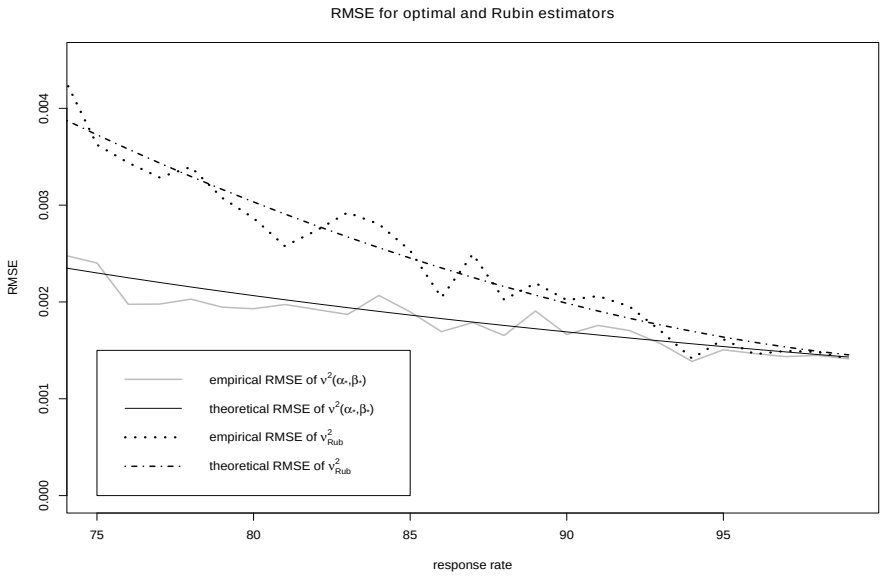


Fig. 2. Close-up of Fig. 1

$\beta_{*,\infty}$ of the estimator $v^2(\alpha,\beta) \in \mathfrak{R}$ are obtained by taking respective limits of α_* and β_* defined in (24)–(28) of Theorem 8.

Theorem 10. *Consider* $\mathrm{GmG}(\mu,\sigma^2,\infty)$ *model. Let*

$$\alpha_{*,\infty} = \frac{nm-2m+1}{f(m-1)}K \quad \textit{and} \quad \beta_{*,\infty} = -\frac{r-1}{(1-f)(n-1)}K, \tag{30}$$

where

$$K = \frac{2(r-1)\left(1+\frac{1-f}{m}\right)}{m(n-1)(r+1)A_{1,\infty}}$$

and

$$A_{1,\infty} = \lim_{K\to\infty} A_1 = \left(1+\tfrac{2(n-r)}{m(n-1)^2}\right)\tfrac{m+1}{m-1} - \left(1+\tfrac{2}{m(n-1)}\right)^2$$

Then, in the case of non-informative prior, $v^2(\alpha_{*,\infty},\beta_{*,\infty})$ *is the optimal estimator of the MSE of* $\bar{X}_{MImp}$ *in the class* $\mathfrak{R}$*. The MSE of this estimator is*

$$\mathrm{MSE}\, v^2(\alpha_{*,\infty},\beta_{*,\infty}) = \frac{\sigma^4}{r^2}\left(1+\frac{1-f}{m}\right)\left(1+\frac{1-f}{m}-\alpha_{*,\infty}f-\beta_{*,\infty}(1-f)\right). \tag{31}$$

Remark 11. *For $n \to \infty$ in such a way that the response rate f remains constant we obtain*

$$\lim_{n\to\infty} \alpha_{*,\infty} = \frac{1+\frac{1-f}{m}}{f},$$
$$\lim_{n\to\infty} n\beta_{*,\infty} = -\frac{(m-1)f\left(1+\frac{1-f}{m}\right)}{m(1-f)}.$$

Thus for large sample size n and small m we can use an approximate optimal version of the estimator of the MSE of the form

$$v^2(\alpha_{*,m},\, \beta_{*,m}), \tag{32}$$

where approximate (for large n) values of $\alpha_{*,\infty}$ and $\beta_{*,\infty}$ are

$$\alpha_{*,m} = \frac{1+\frac{1-f}{m}}{f} \quad \textit{and} \quad \beta_{*,m} = -\frac{(m-1)f\left(1+\frac{1-f}{m}\right)}{nm(1-f)}.$$

Taking $m \to \infty$ in $\alpha_{*,m}$ and $\beta_{*,m}$ we get

$$\alpha_{**} = \frac{1}{f} \qquad \textit{and} \qquad \beta_{**} = -\frac{f}{n(1-f)}.$$

Thus, if additionally number of imputations m is large one may use a simplified version of the optimal estimator of the form

$$v^2(\alpha_{**},\, \beta_{**}). \tag{33}$$

As it is seen in Fig. 3 and Fig. 4 below, both approximate estimators of the MSE, $v^2(\alpha_{*,m},\, \beta_{*,m})$ and $v^2(\alpha_{**},\, \beta_{**})$, are close to the optimal one and perform much better than Rubin's estimator $v^2_{Rub} = v^2(1, 1+1/m)$. The same holds true for the estimator $v^2(\alpha_1, \beta_1)$, where $\alpha_1 = 1/f$ and $\beta_1 = (\frac{1}{f} - \frac{1}{1-f})\frac{1}{n}$ designed as simplified approximate optimal for the ordinary (non-Bayesian) Gaussian model in Wesołowski (2017). As emphasized in Van Buuren (2018), p. 72, the $\frac{1}{m}$ part of the β coefficient in v^2_{Rub} "is critical to make multiple imputation work at low levels of m". However, as we see in Fig. 3, performance of v^2_{Rub} for low m (Fig.3 is for $m = 5$) is much worse than that of the optimal or approximate optimal estimators. Actually, performance of v^2_{Rub} for low m is even worse than that for high m (Fig. 4 is for $m = 100$). The Bjørnstad estimator, v^2_{Bjo}, except of very small response rates, performs poorly for both low and high m.

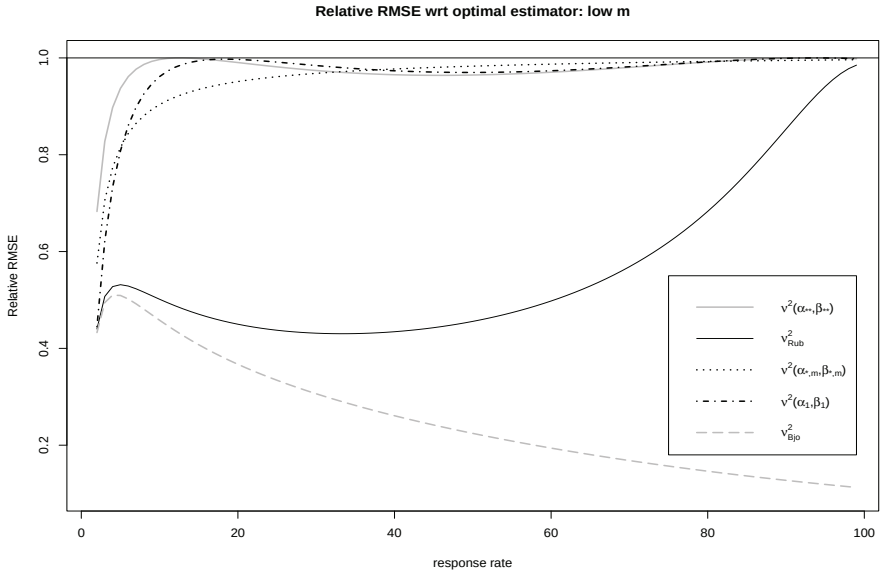


Fig. 3. Ratios (the case of low value of m): the RMSE of the optimal estimator $v^2(\alpha_*,\beta_*)$ divided by the RMSE of $v^2(\alpha_{**},\beta_{**})$, v^2_{Rub}, $v^2(\alpha_{*,m},\beta_{*,m})$, $v^2(\alpha_1,\beta_1)$ and v^2_{Bjo}, respectively. The computations were done for $m=5$, $n=100$, $\sigma^2=1$, $\kappa=\infty$

4.2. UNBIASED ESTIMATORS FOR NON-INFORMATIVE PRIOR

Note that it follows from the formula for the bias of Rubin's estimator, see (17), that if $\kappa\to\infty$, that is in the non-informative case, v^2_{Rub} is unbiased. We have already seen that this estimator is non-admissible in the class $\mathfrak{R}$.

Now we address a natural question of optimality of the Rubin estimator among unbiased estimators of the class $\mathfrak{R}$, i.e. we are interested in the class

$$\mathfrak{R}_u=\{\alpha\bar{U}_m+\beta B_m:\text{ such that }\ \alpha\mathbb{E}\bar{U}_m+\beta\mathbb{E}B_m=\mathrm{MSE}\bar{X}_{MImp}\}\subset\mathfrak{R}.$$

As it is shown below, Rubin's estimator v^2_{Rub} is non-admissible also in $\mathfrak{R}_u$.

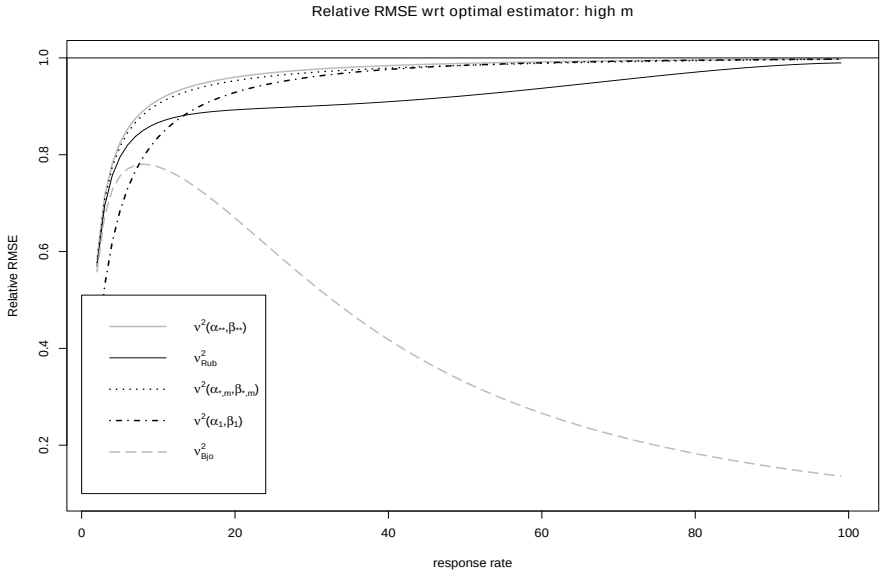


Fig. 4. Ratios (the case of high value of m): the RMSE of the optimal estimator $v^2(\alpha_*,\beta_*)$ divided by the RMSE of $v^2(\alpha_{**},\beta_{**})$, v^2_{Rub}, $v^2(\alpha_{*,m},\beta_{*,m})$, $v^2(\alpha_1,\beta_1)$ and v^2_{Bjo}, respectively. The computations were done for $m=100$, $n=100$, $\sigma^2=1$, $\kappa=\infty$

Theorem 12. *We consider the model* $\mathrm{GmG}(\mu,\sigma^2,\infty)$. *Let*

$$\alpha_{*,u}=\frac{1}{f}\left(1+\frac{1-f}{m}\right)\frac{(m(n-2)+1)(n-1)}{m(n-1)^2-(m-1)(n+r-2)} \tag{34}$$

and

$$\beta_{*,u}=-\frac{1}{1-f}\left(1+\frac{1-f}{m}\right)\frac{(r-1)(m-1)}{m(n-1)^2-(m-1)(n+r-2)}. \tag{35}$$

Then $v^2(\alpha_{*,u},\beta_{*,u})$ *is optimal estimator of the MSE of the* $\bar{X}_{MImp}$ *in the class* $\mathfrak{R}_u$.

Remark 13. *Note that in the case of unbiased estimators, simplified versions of* $v^2(\alpha_{*,u},\beta_{*,u})$ *for large n and large both n and m are exactly the same as the estimators given in* (32) *and* (33)*, respectively. It follows from the fact that the limits of* $\alpha_{*,u}$ *and* $n\beta_{*,u}$ *as* $n\to\infty$ *and then as also* $m\to\infty$ *are exactly the same as in Remark 11.*

Proposition 14. *The* MSE *of the Rubin estimator* v^2_{Rub} *has the form*

$$\mathrm{MSE}\, v^2_{Rub}=\frac{2\sigma^4}{r^2(r-1)}\left(\left(1+\frac{1-f}{m}\right)^2+A\right), \tag{36}$$

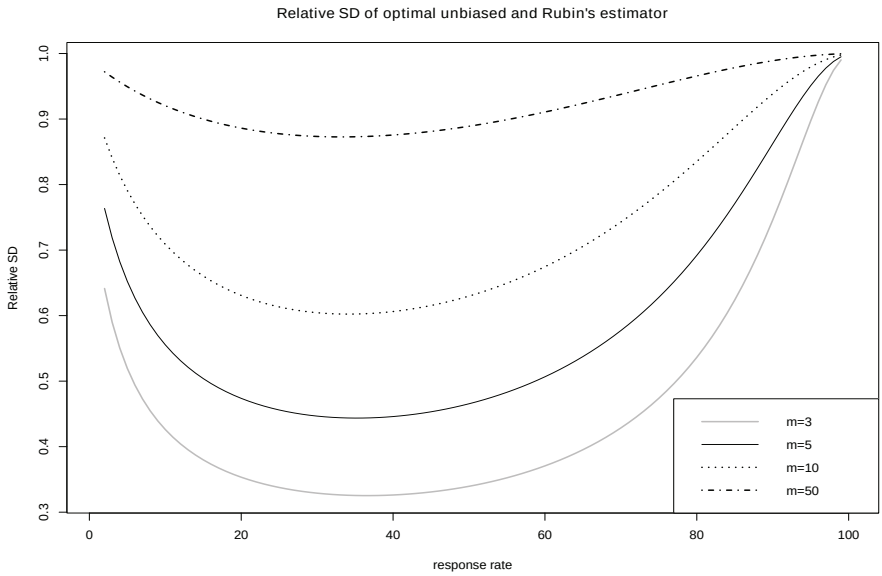


Fig. 5. Ratios: standard deviation of optimal unbiased estimator $v^2(\alpha_{*,u},\beta_{*,u})$ divided by that of the Rubin estimator v^2_{Rub} for different choices of $m = 3,5,10,50$. The computations were done for $n = 100$, $\sigma^2 = 1$

where

$$A = \frac{(r+1)(1-f)}{m}\left[\frac{rf}{(n-1)^2} + \frac{(m+1)^2(1-f)}{m(m-1)} + 2\frac{(m+1)f}{m(n-1)}\right].$$

In Fig.5 we compare standard deviations of v^2_{Rub} and $v^2(\alpha_{*,u},\beta_{*,u})$ for traditional choices for m, that is $m = 3,5,10$ and the higher one, $m = 50$. We see that the larger m gets, the closer standard deviation of the Rubin estimator to the one of the optimal unbiased estimator. Actually, as it is proved in the next result, asymptotically (as $m \to \infty$) they are identical. Nevertheless, the optimal estimator $v^2(\alpha_{*,\infty},\beta_{*,\infty})$ (the one without the unbiasedness constraint) is asymptotically strictly more efficient than v^2_{Rub}.

Theorem 15. *If $r > 1$ then*

$$\lim_{m\to\infty} \mathbb{V}\mathrm{ar}\, v^2(\alpha_{*,u},\beta_{*,u}) = \lim_{m\to\infty} \mathbb{V}\mathrm{ar}\, v^2_{Rub} = \frac{2\sigma^4}{r^2(r-1)}. \tag{37}$$

Moreover,

$$\lim_{m\to\infty} \frac{\mathrm{MSE}\, v^2(\alpha_{*,\infty},\beta_{*,\infty})}{\mathbb{V}\mathrm{ar}\, v^2_{Rub}} = \frac{r-1}{r+1} < 1. \tag{38}$$

5. CONCLUSIONS

In this paper we analyzed the multiple imputation methodology in the Bayesian Gaussian model with the Gaussian mean $\mathrm{GmG}(\mu,\sigma^2,\kappa)$. We derived optimal weights α_* and β_* such that the estimator $v^2(\alpha_*,\beta_*)$ of $\mathrm{MSE}\bar{X}_{MImp}$ in the class $\mathfrak{R}$ of Rubin-type estimators of the form

$$v^2(\alpha,\beta) = \alpha\bar{U}_m + \beta B_m, \qquad \alpha,\beta\in\mathbb{R},$$

is optimal, i.e. it has the minimal MSE. This estimator outperforms the popular Rubin estimator, $v^2_{Rub} = v^2(1,(m+1)/m)$, with respect to both the bias and the MSE. Since the Rubin estimator is widely used in practice it is worth to emphasize that, in view of the obtained results, this estimator is inadmissible (at least in the Bayesian Gaussian $\mathrm{GmG}(\mu,\sigma^2,\kappa)$-models). Similar situation holds for optimal unbiased estimators for non-informative prior, that is in $\mathrm{GmG}(\mu,\sigma^2,\infty)$ model in which the Rubin estimator is unbiased. Nevertheless, at least in the case of large m both the Rubin estimator and the optimal unbiased one have the same asymptotic MSE. Since the formulas for the optimal coefficients α_*, β_* and $\alpha_{*,u}$, $\beta_{*,u}$ (u stands for the unbiased estimator) are quite complicated we propose their approximate $\alpha_{*,m}$, $\beta_{*,m}$ or simplified α_{**}, β_{**} versions. Numerical experiments show that performances of the optimal, approximate optimal and simplified optimal estimators are comparable under the model with non-informative prior. Therefore, in practice, when the extraordinary precision is not neccessary, we recommend to use non-informative prior with either simplified (when both m and n are large) or approximate (when n is large only) Rubin-type estimators.

APPENDIX

1. PROOF OF THEOREM 1

Proof. The form of the imputation estimator as given in (2) follows immediately from the way, (1), we impute variables. Note that in the $\mathrm{GmG}(\mu,\sigma^2,\kappa)$ model $X_i = M + \sigma Z_i$, $i \in R$, where Z_i, $i \in R$, are iid standard normal random variables and $(Z_i, i \in R)$ and M are independent. Therefore, $S_R^2 = \frac{\sigma^2}{r-1}\sum_{i\in R}(Z_i - \bar{Z})^2$ and $\bar{X}_R = M + \sigma\bar{Z}$ are independent too. Consequently,

$$\mathbb{V}\mathrm{ar}\,\bar{X}_{Imp} = \left(\tfrac{(n\kappa+1)f}{r\kappa+1}\right)^2 \mathbb{V}\mathrm{ar}\,\bar{X}_R + (1-f)^2\,\mathbb{V}\mathrm{ar}\,S_R\bar{W}.$$

Moreover,

$$\mathbb{V}\mathrm{ar}\,\bar{X}_R = \mathbb{V}\mathrm{ar}\,M + \sigma^2\,\mathbb{V}\mathrm{ar}\,\bar{Z} = \tfrac{\sigma^2}{r}(r\kappa+1). \tag{39}$$

Note also that $\bar{W} \sim \mathrm{N}\left(0,\tilde{\tau}^2\right)$, where

$$\tilde{\tau}^2 = \tfrac{n\kappa+1}{(n-r)(r\kappa+1)} = \tfrac{\tau^2}{r(1-f)} \tag{40}$$

and thus

$$\mathbb{V}\mathrm{ar}\,S_R\bar{W} = \mathbb{E}\,S_R^2\,\mathbb{E}\,\bar{W}^2 = \sigma^2\tilde{\tau}^2.$$

Therefore,

$$\mathbb{V}\mathrm{ar}\,\bar{X}_{Imp} = \left(\tfrac{(n\kappa+1)f}{r\kappa+1}\right)^2 \tfrac{\sigma^2}{r}(r\kappa+1) + (1-f)^2\,\sigma^2\,\tfrac{n\kappa+1}{(n-r)(r\kappa+1)}.$$

After simplifications one gets (3).

Note that

$$\mathrm{MSE}\,\bar{X}_{Imp} = \mathbb{E}((\bar{X}_{Imp}-\mu)-(M-\mu))^2 = \mathbb{V}\mathrm{ar}\,\bar{X}_{Imp} + \mathbb{V}\mathrm{ar}\,M - 2\mathbb{C}ov(\bar{X}_{Imp},M).$$

Since

$$\mathbb{C}ov(\bar{X}_{Imp},M) = \tfrac{(n\kappa+1)f}{r\kappa+1}\mathbb{C}ov(\bar{X}_R,M) = \tfrac{(n\kappa+1)f}{r\kappa+1}\mathbb{V}\mathrm{ar}\,M$$

by (3) we get

$$\mathrm{MSE}\,\bar{X}_{Imp} = \tfrac{\sigma^2}{n}(n\kappa+1) + \sigma^2\kappa - 2\tfrac{(n\kappa+1)f}{r\kappa+1}\sigma^2\kappa.$$

After simplifications we arrive at (4). □

2. PROOF OF THEOREM 2

Proof. By (1) directly from the definition of S^2_{Imp} we have

$$(n-1)S^2_{Imp} = \sum_{i\in R}(X_i - \bar{X}_{Imp})^2 + \sum_{i\in R^c}\left(\tfrac{r\kappa\bar{X}_R-\mu}{r\kappa+1} + S_R W_i - \bar{X}_{Imp}\right)^2 =: I_1 + I_2.$$

Using (2) for $\bar{X}_{Imp}$, after computation, we get

$$I_1 = (r-1)S^2_R + r(1-f)^2\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1} - S_R\bar{W}\right)^2,$$
$$I_2 = (n-r-1)S^2_R S^2_Z + rf(1-f)\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1} - S_R\bar{W}\right)^2,$$

where for I_2 we use additonally the fact that $S^2_W = S^2_Z$. Thus (5) follows.

Note also that

$$\mathbb{E}S^2_R = \sigma^2, \quad \mathbb{E}S^2_Z = 1.$$

Moreover, $\bar{X}_R$, S_R and $\bar{W}$ are independent and thus

$$\mathbb{E}\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1} - S_R\bar{W}\right)^2 = \mathbb{E}\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 + \mathbb{E}S^2_r\bar{W}^2. \tag{41}$$

By (39) we get

$$\mathbb{E}\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 = \tfrac{1}{(r\kappa+1)^2}\mathbb{V}ar\bar{X}_R = \tfrac{\sigma^2}{r(r\kappa+1)}.$$

Since $\mathbf{X}_R$ and $\bar{W}$ are independent

$$\mathbb{E}S^2_R\bar{W}^2 = \tilde{\tau}^2\,\mathbb{E}S^2_R. \tag{42}$$

Consequently, due to (40), we see that (5) implies

$$\mathbb{E}S^2_{Imp} = \tfrac{\sigma^2}{n-1}\left(n-2+nf(1-f)\left(\tfrac{1}{r(r\kappa+1)} + \tilde{\tau}^2\right)\right) = \sigma^2,$$

and thus S^2_{Imp} is unbiased for σ^2.

To prove (6) we first note that S^2_{Imp} can be rewritten as

$$(n-1)S^2_{Imp} = S^2_R\left(r-1+(n-r-1)S^2_Z + nf(1-f)\bar{W}^2\right)$$

$$+r(1-f)\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2-2r(1-f)\tfrac{\bar{X}_R-\mu}{r\kappa+1}S_R\bar{W}=:A_1+A_2+A_3.$$

Since $\bar{X}_R$ and the random vector $(S_R^2, S_Z^2, \bar{W})$ are independent it follows that $\mathbb{C}ov(A_1,A_2)=0$. Moreover, since additionally $\mathbb{E}(\bar{X}_R-\mu)=0=\mathbb{E}(\bar{X}_R-\mu)^3$ it follows that $\mathbb{C}ov(A_1,A_3)=0$ $=\mathbb{C}ov(A_2,A_3)$. Therefore

$$(n-1)^2\mathbb{V}\mathrm{ar}S^2_{Imp}=V_1+\left(\tfrac{r(1-f)}{r\kappa+1}\right)^2\left(\tfrac{V_2}{(r\kappa+1)^2}+4V_3\right), \tag{43}$$

where $V_1=\mathbb{V}\mathrm{ar}S_R^2Y$ with

$$Y=r-1+(n-r-1)S_Z^2+r(1-f)\bar{W}^2, \tag{44}$$

$$V_2=\mathbb{V}\mathrm{ar}(\bar{X}_R-\mu)^2 \qquad \text{and} \qquad V_3=\mathbb{V}\mathrm{ar}(\bar{X}_R-\mu)S_R\bar{W}.$$

Note that for independent random variables A, B we have

$$\mathbb{V}\mathrm{ar}AB=\mathbb{V}\mathrm{ar}A\,\mathbb{V}\mathrm{ar}B+(\mathbb{E}A)^2\mathbb{V}\mathrm{ar}B+\mathbb{V}\mathrm{ar}A\,(\mathbb{E}B)^2. \tag{45}$$

Consequently,

$$V_1=\mathbb{V}\mathrm{ar}S_R^2\,\mathbb{V}\mathrm{ar}Y+(\mathbb{E}S_R^2)^2\,\mathbb{V}\mathrm{ar}Y+\mathbb{V}\mathrm{ar}S_R^2\,(\mathbb{E}Y)^2.$$

Since $\frac{r-1}{\sigma^2}S_R^2$ has the $\chi^2(r-1)$ distribution we have

$$\mathbb{V}\mathrm{ar}S_R^2=\tfrac{2}{r-1}\sigma^4. \tag{46}$$

Moreover, see (42) and note that $r(1-f)\tilde{\tau}^2=\tau^2$,

$$\mathbb{E}Y=r-1+(n-r-1)\mathbb{E}S_Z^2+r(1-f)\mathbb{E}\bar{W}^2=n-2+\tau^2. \tag{47}$$

Since S_Z^2 and $\bar{W}$ are independent

$$\mathbb{V}\mathrm{ar}Y=(n-r-1)^2\mathbb{V}\mathrm{ar}S_Z^2+r^2(1-f)^2\mathbb{V}\mathrm{ar}\bar{W}^2.$$

Note that $(n-r-1)S_Z^2$ has the $\chi^2(n-r-1)$ distribution. Consequently,

$$\mathbb{V}\mathrm{ar}S_Z^2=\tfrac{2}{n-r-1}.$$

Since $\bar{W}\sim\mathrm{N}(0,\tilde{\tau}^2)$ it follows that $\mathbb{V}\mathrm{ar}\bar{W}^2=2\tilde{\tau}^4$. Summing up we get

$$\mathbb{V}\mathrm{ar}Y=2\left(n-r-1+\tau^4\right). \tag{48}$$

Consequently,

$$\mathbb{V}\mathrm{ar}\, S_R^2\, \mathbb{V}\mathrm{ar}\, Y = \tfrac{4\sigma^4}{r-1}\left(n-r-1+\tau^4\right), \qquad (\mathbb{E}\, S_R^2)^2\, \mathbb{V}\mathrm{ar}\, Y = 2\sigma^4\left(n-r-1+\tau^4\right)$$

and

$$\mathbb{V}\mathrm{ar}\, S_R^2\, (\mathbb{E}\, Y)^2 = \tfrac{2\sigma^4}{r-1}\left(n-2+\tau^2\right)^2.$$

Therefore, combining the first two terms of V_1, we get

$$V_1 = \tfrac{2\sigma^4(r+1)}{r-1}\left(n-r-1+\tau^4\right) + \tfrac{2\sigma^4}{r-1}\left(n-2+\tau^2\right)^2.$$

Since $\bar{X}_R \sim \mathrm{N}(\mu, \sigma^2\, \frac{r\kappa+1}{r})$ we get

$$V_2 = 2\sigma^4\left(\tfrac{r\kappa+1}{r}\right)^2. \tag{49}$$

By independence of $\bar{X}_R$, S_R^2, $\bar{W}$ we conclude that

$$V_3 = \mathbb{V}\mathrm{ar}\, \bar{X}_R\, \mathbb{E}\, S_R^2\, \tilde{\tau}^2 = \tfrac{r\kappa+1}{r}\, \sigma^4\, \tilde{\tau}^2.$$

Finally, plugging the formulas for V_1, V_2 and V_3 into (43) we can write

$$\begin{aligned}&(n-1)^2 \mathbb{V}\mathrm{ar}\, S_{Imp}^2\\ &= 2\sigma^4\left\{\tfrac{(r+1)(n-r-1)+(n-2)^2+2(n-2)\tau^2+3\tau^4}{r-1} + (1-f)^2\left(\tfrac{1}{(r\kappa+1)^2} + 2\tfrac{r\tilde{\tau}^2}{r\kappa+1} + r^2\tilde{\tau}^4\right)\right\}.\end{aligned}$$

Plugging $\tilde{\tau}^2$ as given in (40) we get

$$(1-f)^2\left(\tfrac{1}{(r\kappa+1)^2} + 2\tfrac{r\tilde{\tau}^2}{r\kappa+1} + r^2\tilde{\tau}^4\right) = 1.$$

Hence, after some algebra, (6) follows. □

3. PROOF OF THEOREM 3

Proof. Each of imputed samples (10) gives rise to the imputation estimator

$$\bar{X}_{Imp}^{(\ell)} = f\, \tfrac{n\kappa+1}{r\kappa+1}\, \bar{X}_R + (1-f)\left(\tfrac{\mu}{r\kappa+1} + S_R \bar{W}^{(\ell)}\right), \quad \ell = 1,\ldots,m,$$

and thus (11) follows immediately from (7).

Plugging

$$\bar{X}^{(\ell)}_{Imp} - \bar{X}_{MImp} = (1-f)S_R\left(\bar{W}^{(\ell)} - \bar{\bar{W}}\right), \quad \ell = 1,\dots,m,$$

into (9) we get (12).

For $\ell = 1,\dots,m$, the imputation version of S^2 statistic, $\left(S^{(\ell)}_{Imp}\right)^2$ (see (5)), satisfies

$$\begin{aligned}(n-1)\left(S^{(\ell)}_{Imp}\right)^2 &= S^2_R\left(r-1+(n-r-1)S^2_{Z^{(l)}} + r(1-f)\left(\bar{W}^{(\ell)}\right)^2\right)\\ &\quad + r(1-f)\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 - 2r(1-f)\tfrac{\bar{X}_R-\mu}{r\kappa+1}S_R\bar{W}^{(\ell)}.\end{aligned}$$

Taking the mean according to (8) we get

$$\begin{aligned}n(n-1)\bar{U}_m &= S^2_R\left[r-1+(n-r-1)\bar{S}^2_Z + r(1-f)\left(\tfrac{m-1}{m}S^2_{\bar{W}} + \bar{\bar{W}}^2\right)\right]\\ &\quad + r(1-f)\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 - 2r(1-f)\tfrac{\bar{X}_R-\mu}{r\kappa+1}S_R\bar{\bar{W}}.\end{aligned}$$

Thus (13) follows.

□

4. PROOF OF THEOREM 4

Proof. The unbiasedness of $\bar{X}^{(\ell)}_{Imp}$, for any $\ell = 1,\dots,m$, implies immediately that $\bar{X}_{MImp}$ is unbiased.

Note that $\mathrm{MSE}\bar{X}_{MImp}$ can be computed as follows:

$$\tfrac{1}{m^2}\mathbb{E}\left(\sum_{\ell=1}^{m}(\bar{X}^{(\ell)}_{Imp} - M)\right)^2 = \tfrac{1}{m}\mathrm{MSE}\bar{X}_{Imp} + \tfrac{m-1}{m}\mathbb{C}ov\left(\bar{X}^{(1)}_{Imp} - M, \bar{X}^{(2)}_{Imp} - M\right).$$

From (2) we conclude that

$$\begin{aligned}&\mathbb{C}ov\left(\bar{X}^{(1)}_{Imp} - M, \bar{X}^{(2)}_{Imp} - M\right) = \mathbb{V}ar(\tau^2\bar{X}_R - M)\\ &\quad + (1-f)\,\mathbb{C}ov\left(\tau^2\bar{X}_R - M, S_R(\bar{W}^{(1)} + \bar{W}^{(2)})\right) + (1-f)^2\,\mathbb{C}ov\left(S_R\bar{W}^{(1)}, S_R\bar{W}^{(2)}\right).\end{aligned}$$

Since $(S_R, \bar{W})$ and $(\bar{X}, M)$ are independent the first covariance in the line above is zero. Moreover, independence of S_R, $\bar{W}^{(1)}$ and $\bar{W}^{(2)}$ together with the fact that $\mathbb{E}\,\bar{W}^{(\ell)} = 0$, $\ell = 1,2$, implies that the second covariance also vanishes.

Since $\bar{X}_R = M + \sigma \bar{Z}_R$, where Z_i, $i \in R$, are iid standard normal random variables and $(Z_i)_{i\in R}$ and M are independent we conclude that

$$\begin{aligned}\mathbb{C}ov\left(\bar{X}^{(1)}_{Imp} - M, \bar{X}^{(2)}_{Imp} - M\right) &= \mathbb{V}\mathrm{ar}\left(\tau^2 \sigma \bar{Z}_R + (\tau^2 - 1) M\right) \\ &= \left(\tfrac{(n\kappa+1)f}{r\kappa+1}\right)^2 \tfrac{\sigma^2}{r} + \left(\tfrac{1-f}{r\kappa+1}\right)^2 \kappa\sigma^2 = \sigma^2 \tfrac{n\kappa+f}{n(r\kappa+1)}.\end{aligned}$$

This formula together with (4), after computation, gives (14).

Since, for any $\ell = 1, \ldots, m$, $\left(S^{(\ell)}_{Imp}\right)^2$ is unbiased for σ^2, see Theorem 2, we get (15).

Similarly, (12) implies $\mathbb{E}\,B_m = (1-f)^2\,\mathbb{E}\,S^2_{\bar{R}}\,\mathbb{E}\,S^2_{\bar{W}} = (1-f)^2\sigma^2\tilde{\tau}^2$ and thus (16) follows from (40).

Using (14), (15), (16) and (40) we see that the Rubin estimate of the variance of $\bar{X}_{MImp}$ is biased with the bias

$$\begin{aligned}\mathbb{E}\,v^2_{Rub} - \mathrm{MSE}\,\bar{X}_{MImp} &= \tfrac{\sigma^2}{n} + \tfrac{m+1}{m}(1-f)^2\tilde{\tau}^2\sigma^2 - \left(\tfrac{n\kappa+f}{n(r\kappa+1)} + \tfrac{(1-f)^2\tilde{\tau}^2}{m}\right)\sigma^2 \\ &= \tfrac{\sigma^2}{n}\left(1 + \tfrac{(1-f)(n\kappa+1) - n\kappa - f}{r\kappa+1}\right).\end{aligned}$$

After calculation we get the formula (17). □

5. PROOF OF PROPOSITION 6

Proof. From the representation (13) we get

$$\mathbb{V}\mathrm{ar}\,\bar{U}_m = \tfrac{1}{n^2(n-1)^2}\left\{\mathbb{V}\mathrm{ar}\,S^2_{\bar{R}}\,\bar{Y} + r^2(1-f)^2\left(\mathbb{V}\mathrm{ar}\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 + 4\mathbb{V}\mathrm{ar}\,\tfrac{\bar{X}_R-\mu}{r\kappa+1}\,S_R\bar{\bar{W}}\right)\right\},$$

where

$$\bar{Y} = r - 1 + (n-r-1)\bar{S}^2_Z + r(1-f)\left(\bar{\bar{W}}^2 + \tfrac{m-1}{m}S^2_{\bar{W}}\right).$$

Since $\bar{X}_R, S_R^2, \bar{Y}, \bar{\bar{W}}$ are (jointly) independen, $\mathbb{E}\bar{\bar{W}} = 0$ and $\mathbb{E}\bar{X}_R = \mu$ it follows that all the covariances expected to be in $\mathbb{V}ar\bar{U}_m$ are equal zero.

Note that $\bar{Y} = \frac{1}{m}\sum_{\ell=1}^{m} Y^{(\ell)}$, where $Y^{(1)}, \ldots, Y^{(m)}$ are iid copies of Y defined in (44) (with Z and $\bar{W}$ changed into $Z^{(\ell)}$ and $\bar{W}^{(\ell)}$, respectively). Thus $Y^{(\ell)} \stackrel{d}{=} Y$, $\ell = 1, \ldots, m$. Hence $\mathbb{E}\bar{Y} = \mathbb{E}Y$ and $\mathbb{V}ar\bar{Y} = \frac{1}{m}\mathbb{V}arY$. Consequently, (45) yields

$$\mathbb{V}ar S_R^2\bar{Y} = \tfrac{1}{m}\mathbb{E}S_R^4\mathbb{V}arY + \mathbb{V}ar S_R^2(\mathbb{E}Y)^2.$$

Note that $\mathbb{E}S_R^4 = \mathbb{V}ar S_R^2 + (\mathbb{E}S_R^2)^2 = \frac{2\sigma^4}{r-1} + \sigma^4 = \frac{\sigma^4(r+1)}{r-1}$. Therefore (47) and (48) yield

$$\mathbb{V}ar S_R^2\bar{Y} = \tfrac{2\sigma^4}{r-1}\left[\left(n-2+\tau^2\right)^2 + \tfrac{r+1}{m}\left(n-r-1+\tau^4\right)\right].$$

Moreover, due to (49)

$$\mathbb{V}ar\left(\tfrac{\bar{X}_R-\mu}{r\kappa+1}\right)^2 = \tfrac{2\sigma^4}{(r(r\kappa+1))^2}.$$

For the third term in the expression for $\mathbb{V}ar\bar{U}_m$ we obtain, see (39) and (40),

$$\mathbb{V}ar\tfrac{\bar{X}_R-\mu}{r\kappa+1}S_R\bar{\bar{W}} = \mathbb{V}ar\tfrac{\bar{X}_R-\mu}{r\kappa+1}\,\mathbb{E}S_R^2\,\mathbb{E}\bar{\bar{W}}^2 = \tfrac{\sigma^4\tau^2}{mr^2(r\kappa+1)(1-f)}.$$

Combining the three terms, after some algebra we arrive at (20).

To compute the variance of B_m we use (12) and thus we get

$$\mathbb{V}ar B_m = (1-f)^4(\mathbb{V}ar S_R^2\mathbb{V}ar S_{\bar{W}}^2 + \mathbb{V}ar S_R^2(\mathbb{E}S_{\bar{W}}^2)^2 + (\mathbb{E}S_R^2)^2\mathbb{V}ar S_{\bar{W}}^2).$$

Since $\frac{(m-1)S_{\bar{W}}^2}{\tilde{\tau}^2}$ has the chi-square distribution with $m-1$ degrees of freedom it follows that, see (40), $\mathbb{E}S_{\bar{W}}^2 = \tilde{\tau}^2$ and

$$\mathbb{V}ar S_{\bar{W}}^2 = \tfrac{2}{m-1}\tilde{\tau}^4. \tag{50}$$

Consequently,

$$\mathbb{V}ar B_m = (1-f)^4\left(\tfrac{2\sigma^4}{r-1}\left(\tfrac{2}{m-1}\tilde{\tau}^4 + \tilde{\tau}^4\right) + \sigma^4\tfrac{2}{m-1}\tilde{\tau}^4\right) = 2(1-f)^4\sigma^4\tilde{\tau}^4\tfrac{r+m}{(r-1)(m-1)}.$$

To compute $\mathbb{C}ov(\bar{U}_m, B_m)$ we first note, that

$$\mathbb{C}ov(\bar{U}_m, B_m) = \tfrac{(1-f)^2}{n(n-1)}\mathbb{C}ov(S_R^2\bar{Y}, S_R^2S_{\bar{W}}^2).$$

since similarly as in the case of $\mathbb{V}ar\bar{U}_m$ remaining covariances are zero. Now, from the definition of $\bar{Y}$ we have

$$\begin{aligned}\mathbb{C}ov(S_R^2\bar{Y}, S_R^2 S_{\bar{W}}^2) &= (r-1)\mathbb{C}ov(S_R^2, S_R^2 S_{\bar{W}}^2) + (n-r-1)\mathbb{C}ov(S_R^2\bar{S}_Z^2, S_R^2 S_{\bar{W}}^2)\\ &\quad + r(1-f)\tfrac{m-1}{m}\mathbb{V}ar\, S_R^2 S_{\bar{W}}^2 + r(1-f)\mathbb{C}ov(S_R^2\bar{\bar{W}}^2, S_R^2 S_{\bar{W}}^2).\end{aligned}$$

Now we compute three covariances and the variance of the above formula. First we note that independence of S_R^2 and $S_{\bar{W}}^2$ implies

$$\mathbb{C}ov(S_R^2, S_R^2 S_{\bar{W}}^2) = (\mathbb{V}ar\, S_R^2)\,\mathbb{E}\, S_{\bar{W}}^2 = \tfrac{2\sigma^4}{r-1}\,\tilde{\tau}^2.$$

For the second covariance $\mathbb{C}ov(S_R^2\bar{S}_Z^2, S_R^2 S_{\bar{W}}^2)$ we first note that $S_{\bar{W}}^2 = S_{\bar{Z}}^2 + 2\sqrt{\frac{\kappa}{r\kappa+1}}\,S_{\bar{Z},U}^2 + \frac{\kappa}{r\kappa+1}S_U^2$, where $S_{\bar{Z}}^2 = \frac{1}{m-1}\sum_{l=1}^m(\bar{Z}^{(l)} - \bar{\bar{Z}})^2$, $S_{\bar{Z},U}^2 = \frac{1}{m-1}\sum_{l=1}^m(\bar{Z}^{(l)} - \bar{\bar{Z}})(U^{(l)} - \bar{U})$ and $S_U^2 = \frac{1}{m-1}\sum_{l=1}^{m-1}(U^{(l)} - \bar{U})^2$. That is, $S_{\bar{W}}^2$ is a function of $(U^{(l)}, l = 1,\ldots,m)$ and $(\bar{Z}^{(l)}, l = 1,\ldots,m)$, while $\bar{S}_Z^2$ is a function of $((S_Z^{(l)})^2, l = 1,\ldots,m)$. Consequently, S_R^2, $\bar{S}_Z^2$ and $S_{\bar{W}}^2$ are independent and thus

$$\mathbb{C}ov(S_R^2\bar{S}_Z^2, S_R^2 S_{\bar{W}}^2) = \mathbb{V}ar\, S_R^2\,\mathbb{E}\,\bar{S}_Z^2\,\mathbb{E}\, S_{\bar{W}}^2 = \tfrac{2\sigma^4}{r-1}\,\tilde{\tau}^2.$$

In view of (12) and (21) it follows immediately that

$$\tfrac{m-1}{m}\mathbb{V}ar\, S_R^2 S_{\bar{W}}^2 = \tfrac{2\sigma^4\tilde{\tau}^4}{r-1}\left(1 + \tfrac{r}{m}\right).$$

To complete the computation of $\mathbb{C}ov(\bar{U}_m, B_m)$ we note that independence of S_R^2, $\bar{\bar{W}}$ and $S_{\bar{W}}^2$ implies

$$\mathbb{C}ov(S_R^2\bar{\bar{W}}^2, S_R^2 S_{\bar{W}}^2) = \mathbb{V}ar\, S_R^2\,\mathbb{E}\,\bar{\bar{W}}^2\,\mathbb{E}\, S_{\bar{W}}^2 = \tfrac{2\sigma^4\tilde{\tau}^4}{m(r-1)}.$$

Therefore,

$$\mathbb{C}ov(\bar{U}_m, B_m) = \tfrac{(1-f)^2}{n(n-1)}\left(2\sigma^4\tilde{\tau}^2 + (n-r-1)\tfrac{2\sigma^4\tilde{\tau}^2}{r-1} + nf(1-f)\tfrac{2\sigma^4\tilde{\tau}^4}{r-1}\left(1 + \tfrac{r+1}{m}\right)\right).$$

And thus (22) follows. □

6. PROOF OF THEOREM 8

Proof. Since

$$\begin{aligned}\mathrm{MSE}(\nu^2(\alpha,\beta)) &= \mathbb{V}\mathrm{ar}(\nu^2(\alpha,\beta)) + \mathbb{B}^2(\nu^2(\alpha,\beta)) \qquad (51)\\ &= \alpha^2\mathbb{V}\mathrm{ar}\,\bar{U}_m + \beta^2\,\mathbb{V}\mathrm{ar}\,B_m + 2\alpha\beta\,\mathbb{C}ov(\bar{U}_m, B_m) + (\alpha\,\mathbb{E}\,\bar{U}_m + \beta\,\mathbb{E}\,B_m - \mathrm{MSE}\,\bar{X}_{MImp})^2\end{aligned}$$

we have to minimize the function

$$T(\alpha,\beta) = \alpha^2\mathbb{E}\,\bar{U}_m^2 + \beta^2\mathbb{E}\,B_m^2 + 2\alpha\beta\mathbb{E}\,\bar{U}_m B_m - 2\alpha\mathbb{E}\,\bar{U}_m\,\mathrm{MSE}\,\bar{X}_{MImp}$$

$$-2\beta\mathbb{E}\,B_m\,\mathrm{MSE}\,\bar{X}_{MImp} + (\mathrm{MSE}\,\bar{X}_{MImp})^2.$$

Therefore we differentiate T with respect to α and β to find the stationary point which gives the minimum. Differentation leads to the system of linear equations

$$\alpha\,\mathbb{E}\,\bar{U}_m^2 + \beta\,\mathbb{E}\,\bar{U}_m B_m = \mathbb{E}\,\bar{U}_m\,\mathrm{MSE}\,\bar{X}_{MImp}, \qquad (52)$$

$$\alpha\,\mathbb{E}\,\bar{U}_m B_m + \beta\mathbb{E}\,B_m^2 = \mathbb{E}\,B_m\,\mathrm{MSE}\,\bar{X}_{MImp}. \qquad (53)$$

Therefore,

$$\alpha = \frac{\mathrm{MSE}\,\bar{X}_{MImp}\left(\mathbb{E}\,\bar{U}_m\mathbb{E}\,B_m^2 - \mathbb{E}\,B_m\mathbb{E}\,\bar{U}_m B_m\right)}{\mathbb{E}\,\bar{U}_m^2\mathbb{E}\,B_m^2 - (\mathbb{E}\,\bar{U}_m B_m)^2}, \qquad (54)$$

$$\beta = \frac{\mathrm{MSE}\,\bar{X}_{MImp}\left(\mathbb{E}\,B_m\mathbb{E}\,\bar{U}_m^2 - \mathbb{E}\,\bar{U}_m\mathbb{E}\,\bar{U}_m B_m\right)}{\mathbb{E}\,\bar{U}_m^2\mathbb{E}\,B_m^2 - (\mathbb{E}\,\bar{U}_m B_m)^2}. \qquad (55)$$

Consequently,

$$\mathbb{E}\,\bar{U}_m^2\mathbb{E}\,B_m^2 - (\mathbb{E}\,\bar{U}_m B_m)^2 = \left(\frac{\sigma^4(r+1)\tau^2(1-f)}{nr(r-1)}\right)^2 \cdot A_1,$$

with A_1 defined in (25),

$$\mathbb{E}\,\bar{U}_m\mathbb{E}\,B_m^2 - \mathbb{E}\,B_m\,\mathbb{E}\,\bar{U}_m B_m = \frac{2\sigma^6\tau^4(1-f)^2(r+1)}{nr^2(r-1)} \cdot A_2,$$

with A_2 defined in (26),

$$\mathbb{E}\,\bar{U}_m^2\mathbb{E}\,B_m - \mathbb{E}\,\bar{U}_m\,\mathbb{E}\,\bar{U}_m B_m = 2\frac{\sigma^6\tau^2(1-f)(r+1)}{n^2(n-1)^2r(r-1)} \cdot A_3,$$

with A_3 defined in (27),

Note that (14) can be rewritten as

$$\mathrm{MSE}(\bar{X}_{MImp}) = \frac{\sigma^2}{r} \cdot A_4,$$

with A_4 defined in (28),

Therefore, plugging the above identities into (54), (55) we obtain the formulas for the optimal α and β as given in (24).

To obtain optimal MSE as given in (29) we compute $T(\alpha_*, \beta_*)$. Due to (52) and (53) we get

$$T(\alpha_*, \beta_*) = \mathrm{MSE}\bar{X}_{MImp}(\mathrm{MSE}\bar{X}_{MImp} - \alpha_* \mathbb{E}\bar{U}_m - \beta^* \mathbb{E}B_m)$$

which, after referring to formulas for $\mathrm{MSE}\bar{X}_{MImp}$, $\mathbb{E}\bar{U}_m$ and $\mathbb{E}B_m$ gives the final result. □

7. PROOF OF THEOREM 10

Proof. Note that

$$\lim_{\kappa\to\infty} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 + \frac{2n(1-f)}{(n-1)^2 m} \\ \frac{(1-f)^2}{f^2} \frac{m+1}{m-1} \\ \frac{(1-f)}{f}\left(1 + \frac{2}{m(n-1)}\right) \end{bmatrix}.$$

Moreover

$$\lim_{\kappa\to\infty} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix} = \begin{bmatrix} \left(1 + \frac{2n(1-f)}{(n-1)^2 m}\right) \frac{m+1}{m-1} - \left(1 + \frac{2}{m(n-1)}\right)^2 \\ \frac{1}{m-1} - \frac{1}{m(n-1)} \\ -\frac{r-1}{m} \\ 1 + \frac{1-f}{m} \end{bmatrix}.$$

Plugging the above values in the formulas (24) gives $\alpha_{*,\infty} = \lim_{\kappa\to\infty} \alpha_*$ and $\beta_{*,\infty} = \lim_{\kappa\to\infty} \beta_*$ as in (30) .

The MSE follows from taking the limit as $\kappa \to \infty$ in (29), which is the same as inserting in this formula the limiting values of a, b, c and A_1, A_2, A_3, A_4 as obtained above. Thus the result follows. □

8. PROOF OF THEOREM 12

Proof. The standard Lagrange approach to the minimization problem for $\mathbb{V}\mathrm{a}r\,v^2(\alpha,\beta)$ under the unbiasedness condition $\alpha\mathbb{E}\bar{U}_m+\beta\,\mathbb{E}B_m=\mathrm{MSE}\bar{X}_{MImp}$ leads to the solution

$$\alpha_{*,u}=\frac{J(\alpha)\,\mathrm{MSE}\bar{X}_{MImp}}{J(\alpha)\,\mathbb{E}\bar{U}_m+J(\beta)\,\mathbb{E}B_m}\quad\text{and}\quad\beta_{*,u}=\frac{J(\beta)\,\mathrm{MSE}\bar{X}_{MImp}}{J(\alpha)\,\mathbb{E}\bar{U}_m+J(\beta)\,\mathbb{E}B_m},$$

where

$$J(\alpha)=\mathbb{E}\bar{U}_m\,\mathbb{V}\mathrm{a}r\,B_m-\mathbb{E}B_m\mathbb{C}ov(\bar{U}_m,B_m)\quad\text{and}\quad J(\beta)=\mathbb{V}\mathrm{a}r\,\bar{U}_m\,\mathbb{E}B_m-\mathbb{E}\bar{U}_m\mathbb{C}ov(\bar{U}_m,B_m).$$

We note that for $\tau^2=1$ (i.e. when $\kappa\to\infty$)

$$J(\alpha)=\frac{2\sigma^6(1-f)^2(r+1)}{nr^2(r-1)}\left(\frac{1}{m-1}-\frac{1}{m(n-1)}\right),\quad J(\beta)=-\frac{2\sigma^6(1-f)(r+1)}{n^2(n-1)^2rm}$$

and

$$\mathrm{MSE}\bar{X}_{MImp}=\frac{\sigma^2}{r}\left(1+\frac{1-f}{m}\right).$$

Hence we obtain $\alpha_{*,u}$ and $\beta_{*,u}$ as given in (34) and (35). □

9. PROOF OF PROPOSITION 14

Proof. Note that

$$\mathbb{V}\mathrm{a}r\,v^2_{Rub}=\mathbb{V}\mathrm{a}r\,\bar{U}_m+(1+\tfrac{1}{m})^2\mathbb{V}\mathrm{a}r\,B_m+2(1+\tfrac{1}{m})\mathbb{C}ov(\bar{U}_m,B_m)$$

where $\mathbb{V}\mathrm{a}r\,\bar{U}_m$, $\mathbb{V}\mathrm{a}r\,B_m$ and $\mathbb{C}ov(\bar{U}_m,B_m)$ are given in (20), (21) and (22), respectively. Plugging $\tau^2=1$ in these formulas, after calculations we get (36). □

10. PROOF OF THEOREM 15

Proof. Note that (36) implies

$$\lim_{m\to\infty}\mathbb{V}\mathrm{a}r\,v^2_{Rub}=\frac{2\sigma^4}{r^2(r-1)}.$$

From (34) and (35) we get

$$\lim_{m\to\infty}\alpha_{*,u}=\frac{(n-1)(n-2)}{f[(n-1)(n-2)-(r-1)]}\qquad\text{and}\qquad\lim_{m\to\infty}\beta_{*,u}=-\frac{r-1}{(1-f)[(n-1)(n-2)-(r-1)]}.$$

Since $\lim_{\kappa\to\infty}\tau^2=1$, from (20), (21) and (22) we get

$$\lim_{\substack{\kappa\to\infty\\ m\to\infty}}\mathbb{V}ar\,\bar{U}_m=\frac{2\sigma^4}{(r-1)n^2},\quad\lim_{\substack{\kappa\to\infty\\ m\to\infty}}\mathbb{V}ar\,B_m=\frac{2\sigma^4(1-f)^2}{r^2(r-1)},\quad\lim_{\substack{\kappa\to\infty\\ m\to\infty}}\mathbb{C}ov(\bar{U}_m,B_m)=\frac{2\sigma^4(1-f)}{r(r-1)n}.$$

Thus

$$\lim_{m\to\infty}\mathbb{V}ar\,v^2(\alpha_{*,u},\beta_{*,u})=\frac{2\sigma^4}{r^2(r-1)}\lim_{m\to\infty}(f\alpha_{*,u}+(1-f)\beta_{*,u})^2$$

and the result follows since $\lim_{m\to\infty}(f\alpha_{*,u}+(1-f)\beta_{*,u})=1$.

We note that

$$\lim_{m\to\infty}\alpha_{*,\infty}=\frac{r-1}{r+1}\lim_{m\to\infty}\alpha_{*,u}\quad\text{and}\quad\lim_{m\to\infty}\beta_{*,\infty}=\frac{r-1}{r+1}\lim_{m\to\infty}\beta_{*,u}.$$

Therefore

$$\lim_{m\to\infty}\mathbb{V}ar\,v^2(\alpha_{*,\infty},\beta_{*,\infty})=\left(\frac{r-1}{r+1}\right)^2\lim_{m\to\infty}\mathbb{V}ar\,v^2(\alpha_{*,u},\beta_{*,u})=\frac{2\sigma^4(r-1)}{r^2(r+1)^2}.$$

Similarly,

$$\lim_{m\to\infty}\mathbb{B}\,v^2(\alpha_{*,\infty},\beta_{*,\infty})=\lim_{m\to\infty}\left[\frac{r-1}{r+1}\left(\alpha_{*,u}\,\mathbb{E}\,\bar{U}_m+\beta_{*,u}\,\mathbb{E}\,B_m\right)-\mathrm{MSE}\,\bar{X}_{MImp}\right].$$

Since $v^2(\alpha_{*,u},\beta_{*,u})$ is ubiased for $\mathrm{MSE}\,\bar{X}_{MImp}$ it follows that

$$\lim_{m\to\infty}\mathbb{B}\,v^2(\alpha_{*,\infty},\beta_{*,\infty})=\left(\frac{r-1}{r+1}-1\right)\lim_{m\to\infty}\mathrm{MSE}\,\bar{X}_{MImp}=-\frac{2\sigma^2}{r(r+1)}.$$

Finally, we get

$$\lim_{m\to\infty}\mathrm{MSE}\,v^2(\alpha_{*,\infty},\beta_{*,\infty})=\frac{2\sigma^4(r-1)}{r^2(r+1)^2}+\left(\frac{2\sigma^2}{r(r+1)}\right)^2=\frac{2\sigma^4}{r^2(r+1)}.$$

□

References

[1] AKANDE, O., LI, F., REITER, J. An empirical comparison of multiple imputation methods for categorical data. *Amer. Statist.* (2017) DOI 10.1080/00031305 .2016.1277158.

[2] BJØRNSTAD, J.F. Non-Bayesian multiple imputation. *J. Offic. Statist.* **23(4)** (2007), 433-452.

[3] BODNER, T.E. What improves with increasing number of missing data imputations. *Struct. Equat. Model.* **15** (2008), 651-675.

[4] DI ZIO, M., GUARNERA, U. On multiple imputation through finite mixture Gaussian models. In: *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thiemme, eds., Springer, Berlin 2008, 111-118.

[5] FAY, R.E. When are inferences from multiple imputation valid? In: *Proc. Surv. Res. Meth. Sec.*, Amer. Statist. Assoc., Alexandria 1992, 227-232.

[6] GRAHAM, J.W., OLCHOWSKI, A.E., GILREATH, T.D. How many imputatuions are really needed? Some practical clarfications of multiple impuatation theory. *Prev. Sci.* **8(3)** (2007), 206-213.

[7] HUGHES, R.A. STERNE, J.A.C., TILLING, K. Comparison of imputation variance estimators. *Statist. Meth. Med. Res.* **25(6)** (2016), 2541-2557.

[8] HAYATI REZVAN, P., LEE, K.J, SIMPSON, J.A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med. Res. Meth.* **15:30** (2015), 1-14.

[9] KIM, J.K. Finite sample properties of multiple imputation estimators. *Ann. Statist.* **32(2)** (2004), 766-782.

[10] KIM, J.K., BRICK, J.M., FULLER, W.A., KALTON, G. On the bias of the multiple-imputation variance estimator in survey sampling. *JRSS B* **68(3)** (2006), 509-521.

[11] KOTT, P.S. A paradox of multiple imputation. In: *Proc. Surv. Res. Meth. Sec.*, Amer. Statist. Assoc. 1995, 380-383.

[12] NIELSEN, S.F. Proper and improper multiple imputation. *Int. Statist. Rev.* **71** (2003), 592-627.

[13] LAAKSONEN, S. Multiple imputation for a continuous variable. *J. Math. Statist. Sci.* **2(10)** (2016a), 624-643.

[14] LAAKSONEN, S. A new framework for multiple imputation and applications to a binary variable. *Model. Ass. Statist. Appl.* **11(3)** (2016b), 191-201.

[15] LALL, R. How multiple imputation makes a difference. *Polit. Anal.* **24(4)** (2016), 414-433.

[16] ROBINS, J.M., WANG, M. Inference for imputation estimators. *Biometrika* **87** (2000) 113-124.

[17] RUBIN, D. *Multiple Imputation for Nonresponse in Surveys* Wiley, New York 1987.

[18] SKINNER, C.J. Discussion of J.F. Bjørnstad "Non-Bayesian multiple imputation". *J. Offic. Statist.* **23(4)** (2007), 463-465.

[19] VAN BUUREN, S. *Flexible Imputation of Missing Data* CRC Press, Boca Raton 2018.

[20] VON HIPPEL, P.T. How many imputations are needed? A comment on Hershbereger and Fischer (2003). *Struct. Equat. Model.* **12(2)** (2005), 334-335.

[21] VON HIPPEL, P.T. Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Soc. Meth.* **37(1)** (2007), 83-117.

[22] VON HIPPEL, P.T. Should a normal imputation model be modified to impute skewed variables? *Soc. Method. Res.* **42(4)** (2013), 105-138.

[23] VON HIPPEL, P.T. The bias and efficiency of incomplete-data estimators in small univariate normal samples. *Soc. Method. Res.* **42(1)** (2013), 531-558.

[24] WANG, M., ROBINS, J.M. Large sample theory for parametric multiple imputation procedures. *Biometrika* **85** (1998), 935-948.

[25] WESOŁOWSKI, J. Non-admissibility of the Rubin estimator of the variance in multiple imputation in Gaussian models. (2017), 1-28 - unpublished manuscript.

[26] WESOŁOWSKI, J., TARCZYŃSKI, J. Mathematical basics of imputation techniques. *Wiad. Statyst.* **9(664)** (2016), 7-54. (in Polish)

Bartłomiej Bosek[1], **Sebastian Czerwiński**[2], **Michał Dębski**[3], **Jarosław Grytczuk**[3], **Zbigniew Lonc**[3], **Paweł Rzążewski**[3]

[1] Theoretical Computer Science Department, Faculty of Mathematics and Computer Science, Jagiellonian University in Kraków, Poland
[2] Institute of Mathematics, University of Zielona Góra, Zielona Góra, Poland
[3] Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

COLORING CHAIN HYPERGRAPHS

Manuscript received: 4 July 2020
Manuscript accepted: 11 August 2020

Abstract: In this article we present a collection of problems and results concerning a special type of hypergraphs which emerged recently in discrete geometry, in a context of multiple coverings. These are uniform hypergraphs on the set of positive integers whose edges can be linearly ordered by a relation inherited from the natural order of the integers. We call them *chain hypergraphs*. One transparent example is based on the family of *homogeneous* arithmetic progressions of fixed length k, which are sets of the form $\{a, 2a, \ldots, ka\}$, with $a \in \mathbb{N}$. Is there a k-coloring of $\mathbb{N}$ such that every member of this family is *rainbow*? This innocently looking question has some unexpected connections to deep number-theoretic problems. Another challenge concerns the *weak* version of a hypergraph coloring, in which it is sufficient that no edge of a hypergraph is monochromatic. It is conjectured that every chain hypergraph (with sufficiently large edges) is weakly 2-colorable. We discuss possible ways of attacking these and other related problems. We also pose some new questions involving other kinds of coloring for chain hypergraphs.
Keywords: hypergraph coloring, chain hypergraph, shift hypergraph, arithmetic progression
Mathematics Subject Classification (2020): 05C65 (primary), 05C15, 11B25

1. INTRODUCTION

A *hypergraph* is a pair $H = (V, E)$, where V is a set whose elements are called the *vertices* of H and E is any collection of non-empty subsets of V, called the *edges* of H. A hypergraph is *k-uniform* if each edge is of cardinality k. A *coloring* of a hypergraph is any mapping from its vertex set V to a certain set of colors. We will discuss several coloring problems for

[0]Supported by the Polish National Science Center, Grant Number: NCN 2017/26/D/ST6/00264

a special family of hypergraphs that emerged recently in discrete geometry (see Pálvölgyi [18] and Pach and Pálvölgyi [17]). It is defined as follows.

Let $V = \{1,2,\ldots,n\}$. Let $A = \{a_1,\ldots,a_k\}$ and $B = \{b_1,\ldots,b_k\}$ be two k-element subsets of V, numbered increasingly, that is, $a_i < a_j$ and $b_i < b_j$ for every pair of indices $1 \leqslant i < j \leqslant k$. We write $A \leqslant B$ if $a_i \leqslant b_i$ for all $i \in \{1,2,\ldots,k\}$.

A k-uniform hypergraph H on the vertex set V is called a *chain hypergraph*, or shortly a *chain*, if for every pair of edges A,B of H we have either $A \leqslant B$ or $B \leqslant A$. Equivalently, that the whole collection of edges of H can be put into a linear order accordingly to this relation.

A coloring of a hypergraph H is *proper* if in this coloring no edge of H is monochromatic. The *chromatic number* $\chi(H)$ of a hypergraph H is the least number of colors in a proper coloring of H. The following intriguing problem concerning the chromatic number of chain hypergraphs was posed in [18].

Conjecture 1 (Pálvölgyi [18])**.** *Every k-uniform chain hypergraph H satisfies $\chi(H) = 2$, for a sufficiently large k.*

It is easy to see that 2-uniform hypergraphs (which are simple graphs) can demand more than two colors for a proper coloring. In [18] it was demonstrated that also 3-uniform chain hypergraphs may have the chromatic number greater than 2. However, no example of a 4-uniform chain with the chromatic number greater than 2 is known. On the other hand, one can easily prove that every chain hypergraph H satisfies $\chi(H) \leqslant 3$ (see Theorem 3).

It is perhaps worth noticing that to prove Conjecture 1 it is sufficient to confirm it for just one specific value of k. Indeed, for every $m \geqslant k$, any m-uniform chain H can be restricted to a k-uniform chain H' by taking the first k elements of every edge. So, every proper coloring of H' is a proper coloring of H and therefore we have $\chi(H) \leqslant \chi(H')$.

We will also consider other types of colorings for chain hypergraphs. For instance, in a *rainbow coloring* no edge may contain two vertices with the same color. Let us denote by $\chi_r(H)$ the *rainbow chromatic number* of a hypergraph H, that is, the least number of colors in a rainbow coloring of H. Clearly, for every k-uniform hypergraph we have $\chi_r(H) \geqslant k$.

Our favorite rainbow coloring problem for chains concerns a special sub-family of hypergraphs, whose edges are *homogeneous* arithmetic progressions, that is, sets of the form $\{a,2a,\ldots,ka\}$ for any $a \in \mathbb{N}$. We call them *homogeneous arithmetic chains*. The following conjecture was posed independently by Bosek (see [6]) and Pach and Pálvölgyi (see [17]).

Conjecture 2. *Every k-uniform homogeneous arithmetic chain H satisfies*

$$\chi_r(H) = k.$$

One may easily prove that the conjecture is true for $k = p-1$, where p is any prime number (Proposition 23). However, in general the conjecture might be hard to prove. Indeed, it is stronger than the statement of the famous Graham's *gcd-problem*, which was eventually proved by deep methods of analytic number theory (see [3]).

We will further discuss the above problems in subsequent sections including some results here and there.

2. SKELETONS OF CHAIN HYPERGRAPHS

We start with proving an upper bound for the chromatic number of general chain hypergraphs. The proof shows that every 2-uniform chain is 3-colorable, which is actually a known fact. Indeed, 2-uniform chain hypergraphs are the same as well-known and intensively studied 1-*queue* graphs, which are known to be 3-colorable (see [11]). We decided to include our algorithmic proof, which is slightly different from the one presented in [11].

Theorem 3. *Every chain hypergraph H satisfies $\chi(H) \leqslant 3$.*

Proof. As mentioned above, it suffices to prove the assertion of the theorem for graphs (2-uniform chains). Let $G = (V,E)$ be a chain graph. Consider the following algorithm:

Algorithm 1: Lazy-Greedy Coloring
Output: proper coloring f of G

```
1 c ← 1
2 for each v ∈ V (in increasing order) do
3     if v has a neighbor u such that f(u) = c then
4         c ← minimum color not appearing in {f(u): u ∈ N(v) and u < v}
5     set f(v) = c
```

If a vertex v has a predecessor which is its neighbor, then we call it *important*. Let w_v be the leftmost neighbor of an important vertex v. For two vertices $x < y$, let $[x,y) = \{u : x \leqslant u < y\}$.

We claim that the Algorithm 1 colors any chain graph using at most 3 colors. To prove that we shall show that the following invariant holds:

Invariant: Before an iteration of the algorithm, for an important vertex v, the vertices of $[w_v,v)$ are colored with at most 2 colors such that for some vertex $x \in [w_v,v)$, the vertices of $[w_v,x)$ are colored with one color and the vertices of $[x,v)$ are colored with the other color.

Clearly, the invariant is true before the pass of the **for** loop for the leftmost (the first) important vertex.

Now suppose the invariant holds for all important vertices preceding a vertex v. Let u be the most right important predecessor of v and let x be a vertex such that the vertices of $[w_u,x)$ are colored with blue and the vertices of $[x,u)$ are colored with red. If u has a neighbor y in $[x,u)$, then u is colored with a color, say c, different from red. All vertices of $[u,v) \setminus \{u\}$ are colored with c too because they are not important. Moreover, $x \leqslant w_v$ because G is a chain (consider the edges uy and vw_v). One can easily observe that the invariant holds for v.

Thus, assume that u does not have a neighbor in $[x,u)$. Then, we are done again because all vertices of $[u,v)$ are colored with red and $w_u \leqslant w_v$ (since G is a chain).

From the invariant it is easy to observe that the algorithm uses at most 3 colors. Indeed, since all colored neighbors of v received one of the two colors, we can always use the third color to color v.

□

To obtain a general upper bound for the rainbow chromatic number of chain hypergraphs we need to consider an auxiliary structure – the skeleton graph of a hypergraph. Let H be a hypergraph on the set of vertices V. The *skeleton graph* $G(H)$ of a hypergraph H is a graph with the same vertex set V and the edges joining all pairs of vertices which appear in a common edge of H. It is obvious that any rainbow coloring of H is at the same time a proper coloring of $G(H)$, and vice versa. So, for every hypergraph H we have

$$\chi_r(H) = \chi(G(H)). \tag{1}$$

Recall that the *clique number* of a graph G, denoted by $\omega(G)$, is the largest integer t such that G contains the complete graph K_t as a subgraph. It is clear that every graph G satisfies $\omega(G) \leqslant \chi(G)$. Hence, every hypergraph H satisfies

$$\omega(G(H)) \leqslant \chi_r(H). \tag{2}$$

If H is k-uniform, then $\omega(G(H)) \geqslant k$.

The following simple proposition will be used in bounding the rainbow chromatic number and the clique number of skeletons of chain hypergraphs:

Proposition 4. *Let H be a k-uniform chain hypergraph H on the vertex set V and let $v \in V$. There exists a k-uniform chain hypergraph H' such that $G(H) - v \subseteq G(H')$.*

Proof. Let H be a fixed k-uniform chain. To simplify the notation we denote by $G(H)$ the skeleton of G by G' the graph $G - v$.

Let H'' be the subhypergraph of H induced by $V \setminus \{v\}$, that is, we remove from H the vertex v and all edges containing v (note that some vertices may become isolated). Let G'' denote $G(H'')$. Clearly $V(G'') = V(G')$ and $E(G'') \subseteq E(G')$.

We shall construct H' iteratively. If $E(G'') = E(G')$, then we are done. Suppose then that there are two vertices u, w such that $uw \in E(G')$ and $uw \notin E(G'')$. Then there is an edge $B = \{b_1, b_2, \ldots, b_k\}$ in H, such that $u, w \in B$ (clearly B is not an edge in H''). Moreover, $v \in B$.

Consider the following two cases:

Case 1. There exists an edge in H'', which precedes B in the order $\leqslant$ defining chain hypergraphs. Let $A = \{a_1, a_2, \ldots, a_k\}$ be the edge of H'' directly preceding B in this order. Define $A' = \{a'_1, a'_2, \ldots, a'_k\}$ as follows:

$$a'_i = \begin{cases} a_i \text{ if } b_i \neq u, w \\ b_i \text{ if } b_i = u, w. \end{cases}$$

Let H' be the hypergraph H'' with an additional edge A'. Since $A \leqslant A' \leqslant B$, H' is a k-uniform chain hypergraph. Moreover, since $u, w \in A'$, uw is an edge in $G(H')$.

Case 2. There is no edge in H'' preceding B in the order $\leqslant$. Let $C = \{c_1, c_2, \ldots, c_k\}$ be the first edge in H''. Define $C' = \{c'_1, c'_2, \ldots, c'_k\}$ as follows:

$$c'_i = \begin{cases} c_i \text{ if } b_i \neq u, w \\ b_i \text{ if } b_i = u, w. \end{cases}$$

Let H' be the hypergraph H'' with an additional edge C'. Since $B \leqslant C' \leqslant C$, H' is a k-uniform chain hypergraph. Moreover, since $u, w \in C'$, uw is an edge in $G(H')$.

The statement of the proposition follows from repeated application of the above procedure. □

By applying Proposition 4 a number of times, we obtain the following result:

Corollary 5. *For every k-uniform chain hypergraph H and every induced subgraph G' of $G(H)$ there exists a k-uniform chain hypergraph H' such that G' is a spanning subgraph of $G(H')$.*

A k-uniform chain hypergraph H is *maximal* if for every two of consecutive edges $A \leqslant B$, $|A \cap B| = k-1$. Observe that every chain hypergraph is a subhypergraph of some maximal chain hypergraph on the same vertex set. The following proposition gives an upper bound on the number of edges in a k-uniform chain on n vertices:

Proposition 6 (Pálvölgyi [18])**.** *A k-uniform chain hypergraph H on n vertices has at most $k(n-k)+1$ edges.*

Clearly the chain hypergraph with the maximum number of edges is maximal.

Corollary 7. *A skeleton graph of a k-uniform chain hypergraph on n vertices has at most*

$$(k(n-k)+1)\binom{k}{2} < \frac{k^2(k-1)}{2}n \tag{3}$$

edges.

By this corollary we immediately get the aforementioned general bound on $\chi_r(H)$ for chain hypergraphs.

Theorem 8. *Every k-uniform chain hypergraph H satisfies*

$$\chi_r(H) \leqslant k^2(k-1). \tag{4}$$

Proof. Let H be a k-uniform chain hypergraph with n vertices. We will demonstrate that $\chi(G(H))$ satisfies the asserted inequality.

If $n \leqslant k$ then the theorem is obviously true. Suppose the claim holds for all chain hypergraphs with less than n vertices. By Corollary 7, $G(H)$ has a vertex v with degree smaller than $k^2(k-1)$.

Consider the graph $G' = G(H) - v$. By Corollary 5, there is a k-uniform chain hypergraph H', such that G' is the spanning subgraph of $G(H')$. By the induction hypothesis, $G(H')$ (and thus G') can be colored with $c = k^2(k-1)$ colors. Since the degree of v is strictly smaller than $k^2(k-1)$, we can always extend this coloring to the coloring of $G(H)$ without using any additional colors. □

We are able to get much better upper bound than (4) for the cardinality of the largest clique in the skeleton graph of a chain hypergraph.

Let H be a k-uniform chain hypergraph and let C be a clique in $G(H)$; suppose it has m vertices. By Corollary 5, there exists a k-uniform hypergraph H' such that C is a spanning subgraph of $G(H')$. Since C is a clique, $C = G(H')$. Let H'' denote a maximal k-uniform chain hypergraph on vertex set $V(C)$ with the largest number of edges, such that H' is a subhypergraph of H''. Clearly $G(H'') = C$. Let $V(H'') = \{1,2,\ldots,m\}$ and $E(H'') = \{A_1, A_2, \ldots, A_s\}$ (both sets are ordered).

On one hand, there are $\binom{m}{2}$ edges in C and each of them is covered by at least one edge of H''. The first edge of H'' covers $\binom{k}{2}$ edges of C. Since H'' is maximal, there are $k-1$ edges covered by A_i, which were not covered by A_{i-1}. Thus, we obtain the following inequality:

$$\binom{m}{2} \leqslant \binom{k}{2} + k(m-k)(k-1). \tag{5}$$

From this it follows that $m \leqslant 2k^2 - 3k + 1$, so we have proved the following statement.

Theorem 9. *Every k-uniform chain hypergraph H satisfies*

$$\omega(G(H)) \leqslant 2k^2 - 3k + 1.$$

It would be nice to know how large the dissonance between $\chi_r(H)$ and $\omega(G(H))$ may be for chain hypergraphs. Currently, we do not know of any example where these two numbers differ. Hence, we dare to state the most provocative conjecture.

Conjecture 10. *Every chain hypergraph H satisfies $\chi_r(H) = \omega(G(H))$.*

In the next section we present a class of chain hypergraphs supporting this conjecture.

3. SPECIAL CHAIN HYPERGRAPHS

A chain hypergraph H is called *special* if for every pair of edges A, B such that $A \leqslant B$ the last element in $A \setminus B$ is smaller than the first element in $B \setminus A$. This type of chain hypergraphs was introduced in [17] in connection to the decomposable coverings problem.

Recall that a graph G is *perfect* if every induced subgraph F of G satisfies $\chi(F) = \omega(F)$. We will prove that skeletons of special chain hypergraphs are perfect.

Let H be a k-uniform special chain hypergraph on the vertex set $V = \{1, 2, \ldots, n\}$, with no isolated vertices. Define a relation $\prec$ on the set V in the following way:

$$a \prec b \iff a < b \text{ and no edge in } H \text{ contains both vertices } a \text{ and } b.$$

Observe that $\prec$ is a strict partial order. Irreflexivity and asymmetry are obvious. For the transitivity, consider $a, b, c \in V$ such that $a \prec b$ and $b \prec c$. It is clear that $a < c$. Suppose there is an edge A in H, such that $a, c \in A$. Since $a \prec b$, we know that $b \notin A$. Let B be an edge containing b. Clearly it does not contain a nor c. Assume that $A \leqslant B$ (the other case is symmetric). Thus, we obtain that:

$$\max(A \setminus B) \geqslant c > b \geqslant \min(B \setminus A)$$

which contradicts the speciality of H.

By $\preccurlyeq$ we denote the union of the relations $\prec$ and $=$. Recall that the *incomparability graph* of a partially ordered set P is a graph on P in which every pair of incomparable elements is joined by an edge (and no other edges are present). It is not hard to verify the following proposition:

Proposition 11. *For every special chain hypergraph H on the vertex set V, the skeleton graph $G(H)$ is the incomparability graph of the partially ordered set $(V, \preccurlyeq)$.*

It is well-known that incomparability graphs are perfect (see [10] or [4]). Hence, we get the aforementioned statement.

Corollary 12. *The skeleton graph of every special chain hypergraph is perfect. In particular, every special chain hypergraph H satisfies $\chi_r(H) = \omega(G(H))$.*

We now give an upper bound on the rainbow chromatic number of special chain hypergraphs.

Theorem 13. *Every k-uniform special chain hypergraph H satisfies*

$$\chi_r(H) \leqslant 2k - 1.$$

Moreover, this bound is in general best possible.

Proof. For a hypergraph H with the vertex set V and $V' \subseteq V$, by $H[V']$ we denote the hypergraph with vertex set V' and edge set $\{A \cap V' \colon A \in E(H)\}$. Clearly $\omega(G(H')) \leqslant \omega(G(H))$ for any H and $H' = H[V']$.

We will prove by induction that the maximum clique in a skeleton of special chain hypergraph whose edges have at most k elements has at most $2k - 1$ vertices.

For $k = 1$ the claim is trivial. Now assume $k > 1$ and the claim holds for all hypergraphs with edges of cardinality smaller than k. Let H be a k-uniform special chain hypergraph with

$\omega(G(H))$ of the largest possible value, say $\omega(G(H)) = t$, and let V' be the vertex set of the maximum clique in $G(H)$. By H' we denote $H[V']$. Let x and y denote the the first and the last vertex in V', respectively.

Since H' is special, every edge of H' should contain x or y (this follows from properties of special chain hypergraphs). Let $V'' = V' \setminus \{x,y\}$ and $H'' = H'[V'']$. Every edge in H'' has at most $k-1$ elements, so by inductive hypothesis we know that $\omega(G(H'')) \leqslant 2(k-1)-1$. Thus $t = \omega(G(H')) \leqslant 2(k-1)-1+2 = 2k-1$.

For the lower bound, consider the following hypergraph H. The vertex set of H is $V = \{v_1, v_2, \ldots, v_{2k-1}\}$. The edge set is:

$$\bigcup_{i=k}^{2k-1} \{\{v_1, v_2, \ldots, v_{k-1}, v_i\}\} \cup \{\{v_{k-1}, v_k, \ldots, v_{2k-1}\}\}.$$

It is not hard to verify that it is special and its skeleton is a clique. □

Concerning the weak chromatic number of special chain hypergraphs, it is known that they satisfy Conjecture 1.

Theorem 14 (Pach and Pálvölgyi [17])**.** *Every special chain hypergraph H, with edges of size at least 3, satisfies $\chi(H) = 2$.*

This result has been generalized in [16] as follows:

Theorem 15 (Keszegh and Pálvölgyi [16])**.** *Every $(2t-1)$-uniform special chain hypergraph is t-colorable so that every edge contains at least one point of each color.*

This result is related to the classic theorem of Erdős and Lovász that will be discussed in the next section.

4. SHIFT HYPERGRAPHS

Another type of chain hypergraphs was introduced in a seminal paper by Erdős and Lovász [13], where the famous *local lemma* was invented. A hypergraph H is called a *shift hypergraph* if every edge is a translated copy of a fixed finite subset A of the integers. More formally, H is a shift hypergraph if there exists a set of integers $A = \{a_1, a_2, \ldots, a_k\}$ such that all edges od H are of the form $t + A$, where

$$t + A = \{t + a_1, t + a_2, \ldots, t + a_k\}.$$

We will prove that there exist k-uniform shift hypergraphs whose rainbow chromatic number is of order at least $\Omega(k^2)$. Recall that a set C of vertices in a k-uniform hypergraph H

is a *weak clique* if for each two vertices in C there is an edge in H containing both of these vertices. Clearly, the set C is a clique in the skeleton graph $G(H)$.

A *t-complete sparse ruler* $R(m,k,t)$ of length m with k marks is a sequence of integers $a_1,a_2,...,a_k$ called *marks*, where $0=a_1<a_2<...<a_k=m$, such that for every ℓ, $0\leqslant\ell\leqslant t$, there are marks a_i and a_j such that $a_j-a_i=\ell$. If $t=m$, then a t-complete sparse ruler is called *complete sparse ruler* $R(m,k)$.

There is a close relationship between weak cliques in shift hypergraphs and complete sparse rulers that is captured by the following statement:

Proposition 16. *There exists a k-uniform shift hypergraph with a $(t+1)$-element weak clique whose vertices are consecutive integers if and only if there exists a t-complete sparse ruler $R(m,k,t)$ for some m.*

Proof. ($\Leftarrow$) Let A be the ruler $R(m,k,t)$ and let H be the k-uniform shift chain $\{s+A:\ s=0,1,\ldots,m+t\}$. We claim that the set $C=\{m,m+1,\ldots,m+t\}$ is a weak clique in H. Consider $a,b\in C$, $a<b$, and let $d=b-a$. As $d\leqslant t$, there are two elements a_i,a_j in A such that $a_j-a_i=d$. Clearly, the set $(a-a_i)+A$ contains both a and b.

($\Rightarrow$) Let $H=\{s+A:\ s=0,1,\ldots,r\}$ be a k-uniform shift chain with a $(t+1)$-element weak clique C whose vertices are consecutive integers. We claim that A is a t-complete sparse ruler $R(m,k,t)$, where m is the largest element in A. Let $0\leqslant\ell\leqslant t$. There are two elements a,b in C such that $b-a=\ell$. As some edge $s+A$ contains both a and b, there are $a_i,a_j\in A$ such that $a_j-a_i=\ell$. □

By this proposition we get the aforementioned lower bound for the rainbow chromatic number of shift hypergraphs.

Theorem 17. *For every k there is a k-uniform shift hypergraph H satisfying*

$$\chi_r(H)\geqslant\frac{k^2}{3}-2k+4.$$

Proof. Wichmann [20] constructed for every m a complete sparse ruler $R(m,k)$ with $k\leqslant\lfloor\sqrt{3m}\rfloor+3$. Thus, Proposition 16 implies the existence of k-uniform shifts hypergraphs with weak cliques of cardinality at least $\frac{k^2}{3}-2k+4$ for every k. □

The first coloring problem stated for shift hypergraphs was the following, innocently looking question asked by Strauss (see [13]): For a given t, does there exist a finite k such that for any set S of k integers, there is a t-coloring of the integers such that every integer translate of S (i.e. every set of the form $\ell+S$, where ℓ is an integer) meets every color class?

Let $f(t)$ denote the least such k. In [13] it was proved that $f(t)$ is actually finite. It was the first use of the celebrated local lemma. More specifically, it was shown there that

$$f(t)\leqslant(3+o(1))t\ln t.$$

Then in [1] the following lower bound was derived:

$$f(t) \geqslant (1-o(1))t\ln t.$$

Currently best upper bound obtained in [15] meets asymptotically this lower bound:

$$f(t) \leqslant (1+o(1))t\ln t.$$

A similarly defined function may be studied for more general chain hypergraphs. For instance, Theorem 15 asserts that $f(t) \leqslant 2t-1$ in the case of the special chain hypergraphs. It would be nice to know what happens in the general case.

5. ARITHMETIC CHAIN HYPERGRAPHS

A chain hypergraph H is *arithmetic* if each edge of H is an arithmetic progression. A special case is obtained by allowing only *homogeneous* arithmetic progressions, that is, sets of the form $\{a,2a,\ldots,ka\}$ with $a \in \mathbb{N}$. We will call them *homogeneous arithmetic* chains.

The following innocently looking question was asked by Graham [14]: Is it true that among any n distinct positive integers $a_1,a_2,\ldots,a_n$ there is always a pair a_i,a_j satisfying

$$\frac{a_i}{\gcd(a_i,a_j)} \geqslant n?$$

The question was answered in the affirmative for sufficiently large n by Szegedy [19] and independently by Zaharescu [21]. Then Balasubramanian and Soundararajan [3] gave a complete solution by using deep methods of analytic number theory.

It is not hard to see that Graham's question is equivalent to the following: Is it true that the clique number of the skeleton of any k-uniform homogeneous arithmetic chain is equal to k? The result of [3] gives a positive answer to this question.

Theorem 18 (Balasubramanian and Soundararajan [3])**.** *Every k-uniform homogeneous arithmetic chain H satisfies $\omega(G(H)) = k$.*

Additionally, it was proved in [3] that the maximum biclique in the skeleton graph $G(H)$ is the complete bipartite graph $K_{k,k}$. Perhaps to solve Conjecture 2 one will need to recognize the structure of skeletons $G(H)$ more deeply. Unfortunately, for $k = 3$ these graphs are not perfect. A number of results in that direction was proved in [6]. For instance, the following result showing that Conjecture 2 is asymptotically true:

Theorem 19. *Every k-uniform homogeneous arithmetic chain H satisfies*

$$\chi_r(H) = (1+o(1))k.$$

We conclude this section with the following generalization of Conjecture 2:

Conjecture 20. *For every $t = 2,3,\ldots,k$, every k-uniform homogeneous arithmetic chain hypergraph is t-colorable so that each edge meets every color.*

Actually, the statement is equivalent to Conjecture 2, as the case $t = k$ implies all smaller cases (by any partition of the set of k colors into t non-empty parts). Thus, proving it for each particular value of t gives a step towards the full conjecture.

It is not hard to prove the conjecture for $t = 2$. Actually, in this case one may prove a much stronger statement corresponding to *equitable* partition of the set of k colors. Let us call a 2-coloring of a hypergraph H *perfectly balanced* if every edge has the same number of vertices in each color (or almost the same when k is odd).

Theorem 21 (Bosek and Grytczuk [7])**.** *Every k-uniform homogeneous arithmetic chain has a perfectly balanced 2-coloring.*

Proof (sketch). A basic idea of the proof is simple. Let k be a fixed positive integer. Let f be a *completely multiplicative* function assigning to every positive integer one of the two values $\{-1,+1\}$. This means that $f(ab) = f(a)f(b)$ for any pair $a,b \in \mathbb{N}$. Notice that such a function is completely determined by specifying values $f(p)$ for all prime numbers p. Notice also that it must be $f(1) = +1$.

Suppose now, that we have a completely multiplicative function f that satisfies

$$\sum_{i=1}^{k} f(i) \in \{0,-1,+1\}. \tag{6}$$

Then, for every $a \in \mathbb{N}$ we have

$$\sum_{i=1}^{k} f(ai) = \sum_{i=1}^{k} f(a)f(i) = f(a)\sum_{i=1}^{k} f(i) \in \{0,-1,+1\}. \tag{7}$$

This means that the coloring f is perfectly balanced on every edge $\{a,2a,\ldots,ka\}$. Hence, to prove the assertion of the theorem it suffices to construct, for every k, a completely multiplicative function f satisfying condition (6). This can be done as follows:

We start with a completely multiplicative function g specified by taking $g(p) = \pm 1$ in accordance to whether a prime p is congruent to $+1$ or -1 modulo 3, with $g(3) = +1$. It can be proved that $\sum_{i=1}^{k} g(i)$ is exactly equal to the number of 1's in the ternary expansion of k (see [5]). In particular, this sum is never negative and bounded from above by $\log_3 k + 1$. So, to get a function f with a desired property it suffices to change the sign $+1$ into -1 of at most $\log_3 k + 1$ primes of the form $3t + 1$ lying in the interval $[k/2,k]$. This operation will not affect multiplicativity of g. The fact that there exists a sufficient number of primes of that form in this interval follows from the celebrated Dirichlet's Theorem on primes in arithmetic progressions (see [2]). This gives the result for a sufficiently large k. Complete proof demands more delicate tricks together with some computational experiments (see [7]). □

A perfectly balanced coloring may be defined for more than two colors in the same way. In view of the above result, we state the following strengthening of the last conjecture (still equivalent to Conjecture 2):

Conjecture 22. *For every $t = 2,3,\ldots,k$, every k-uniform homogeneous arithmetic chain has a perfectly balanced t-coloring.*

It is not hard to prove the statement for $k = p-1$, where p is a prime number (see [6], [9]).

Proposition 23. *Let p be a prime number. Then every $(p-1)$-uniform homogeneous arithmetic chain has a perfectly balanced t-coloring, for every $t \in \{2,3,\ldots,p-1\}$.*

Proof. As noticed above, it is enough to prove the statement for $t = p-1$. Let H be a fixed chain hypergraph whose edges are homogeneous arithmetic progressions of length $p-1$.

We define a desired coloring of H as follows: write a natural number n as $n = p^s m$, where m is not divisible by p. Let $r(m)$ be the residue of m modulo p. We assign $r(m)$ as a color of the number n. Since residue zero is excluded, there are $p-1$ different colors.

It is not hard to see that no two elements of the same edge in H may have the same color. Indeed, let an and bn be any two distinct elements of the progression $\{n,2n,\ldots,(p-1)n\}$. Since a and b are not divisible by p, we have $an = p^s ma$ and $bn = p^s mb$. Consequently, the color of an is $r(ma)$ and the color of bn is $r(mb)$. If these two colors are equal, then, by multiplication properties of residues modulo p, also $r(a) = r(b)$, which means that $a = b$. Hence, $an = bn$ and the proof is complete. □

6. THE LAST CHALLENGE

We conclude the paper with a challenging problem inspired by the recent breakthrough result concerning the queue number of planar graphs.

Let G be a graph whose vertices are linearly ordered. The *queue number* of G is the least number of colors needed to color the edges of G so that each color class forms a (2-uniform) chain hypergraph (with respect to the common linear order of vertices). A long-standing open question was whether the queue number of planar graphs was finite. It was recently solved in the affirmative in [12] by proving a deep structural result for more general minor-closed classes of graphs.

A natural generalization of this statement could be formulated by using chain hypergraphs and some natural hypergraph extension of planar graphs. Let us call a hypergraph H *planar* if there is a realization of H by a *pseudo-disk arrangement* in the sense that the vertices are embedded as points and the edges as pseudo-disks such that a point is contained in a pseudo-disk if and only if the respective vertex is in the respective edge (see [8]).

A definition of the analogue of the queue number, which could be called the *chain number* of a hypergraph, is analogous: it is the least number of colors needed to color the edges of an ordered k-uniform hypergraph so that each color class forms a chain hypergraph.

Conjecture 24. *For every $k \geqslant 2$, the chain number of planar k-uniform hypergraphs is bounded.*

References

[1] N. Alon, I. Kříž, and J. Nešetřil, How to color shift hypergraphs, Studia Sci. Math. Hungar. 30 (1995) 1–11.
[2] T. M. Apostol, Introduction to Analytic Number Theory, Springer-Verlag, New York, 1998.
[3] R. Balasubramanian and K. Soundararajan, On a conjecture of R. L. Graham, Acta Arithmetica LXXV.1 (1996) 1–38.
[4] C. Berge and V. Chvátal, Topics on Perfect Graphs, Annals of Discrete Mathematics 21, Elsevier 1984.
[5] P. Borwein, S. K. K. Choi, M. Coons, Completely multiplicative functions taking values in $\{-1, 1\}$, Trans. Amer. Math. Soc. 362 (2010) 6279–6291.
[6] B. Bosek, M. Dębski, J. Grytczuk, J. Sokół, M. Śleszyńska-Nowak, W. Żelazny, Graph coloring and Graham's greatest common divisor problem, Discrete Math. 341 (2018) 781–785
[7] B. Bosek and J. Grytczuk, Reflection on the Erdős Discrepancy Problem, arXiv:2005.14283, 2020.
[8] S. Buzaglo, R. Pinchasi, G. Rote, Topological hypergraphs. In: Pach, J. (ed.) Thirty Essays on Geometric Graph Theory, pp. 71–81. Springer, New York (2013).
[9] A. E. Caicedo, T. A. C. Chartier, P. P. Pach, Coloring the n-smooth numbers with n colors, arXiv:1902.00446, 2019.
[10] R. P. Dilworth, A Decomposition Theorem for Partially Ordered Sets, Annals of Mathematics, 51 (1) (1950) 161–166.
[11] V. Dujmovič and D. R. Wood, On linear layouts of graphs, Discrete Mathematics and Theoretical Computer Science 6 (2004) 339–358.
[12] V. Dujmović, G. Joret, P Micek, P. Morin, T. Ueckerdt, and D. R. Wood, Planar graphs have bounded queue-number, arXiv:1904.04791, 2019.
[13] P. Erdős and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions. In Infinite and Finite Sets, volume 11 of Colloq. Math. Soc. J. Bolyai, pages 609–627. North-Holland, 1975.
[14] R. L. Graham, Unsolved problem 5749, American Math. Monthly 77 (1970) 775.
[15] D. Harris and A. Srinivasan, A note on near-optimal coloring of shift hypergraphs, Random Structures Algorithms 48 (2016) 53–56.
[16] B. Keszegh and D. Pálvölgyi, An abstract approach to polychromatic coloring: shallow hitting sets in ABA-free hypergraphs and pseudohalfplanes, arxiv:1410.0258, 2014.
[17] J. Pach and D. Pálvölgyi, Unsplitable coverings in the plane, Proceedings of 41st International Workshop WG 2015, Lecture Notes in Computer Science Volume 9224, Springer, 2016, 281–298.
[18] D. Pálvölgyi, Decomposition of geometric set systems and graphs, PhD thesis, EPFL, Lausanne, 2010, arXiv:1009.4641.
[19] M. Szegedy, The solution of Graham's greatest common divisor problem, Combinatorica 6 (1986) 67–71.
[20] B. Wichmann. A note on restricted difference bases, J. London Math. Soc. 38 (1962) 465–466.
[21] A. Zaharescu, On a conjecture of Graham, J. Number Theory 27 (1987) 33–40.

Krzysztof Chełmiński

Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

MATERIAL STABILITY IN QUASISTATIC MELAN–PRAGER MODEL

Manuscript received: 30 May 2020
Manuscript accepted: 30 July 2020

Abstract: Quasistatic models in the inelastic deformation theory are very often used in the engineering practice for the numerical analysis of observed real deformation processes. In these models there appear material constants, which are determined experimentally only. For this reason it is important to study the stability of solutions with respect to the material constants. In this article we study the material stability in the quasistatic Melan-Prager model.
Keywords: inelastic deformations, Melan-Prager model, material stability
Mathematics Subject Classification (2020): 35B30, 74C10

1. THEORY OF INELASTIC DEFORMATIONS - SHORT INTRODUCTION

The theory of inelastic deformations is a part of continuum mechanics. In this theory systems of equations are considered, which model viscoplastic deformations of solids at small strains in the quasistatic or in the dynamic setting of the problem. In this article we will study the quasistatic case only. Such systems consist of linear partial differential equations coupled with nonlinear differential inclusions (or ordinary differential equations) for the vector of internal variables. The partial differential equations result from general mechanical laws. The differential inclusions are experimental, and depend on the kind of considered material. Therefore, in engineering sciences there are many inelastic constitutive equations, always specially adapted to the material under consideration. The first part of the system in the quasistatic setting of the problem in all models represents the balance of forces acting on the material

$$\operatorname{div}_x T(x,t) = -F(x,t) \qquad (x,t) \in \Omega \times (0,T_e). \tag{1}$$

Here $T_e > 0$ is a fixed length of a time interval, $\Omega \subset \mathbb{R}^3$ is a bounded domain with a smooth boundary $\partial\Omega$, $u : \Omega \times (0,T_e) \to \mathbb{R}^3$ denotes the displacement field, $T : \Omega \times (0,T_e) \to \mathcal{S}^3 = \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ is the Cauchy stress tensor, and $F : \Omega \times (0,T_e) \to \mathbb{R}^3$ describes the external forces acting on the body. The other parts of the system consist of constitutive relations

$$T(x,t) = \mathcal{F}_{0\le s\le t}\Big(\nabla u(x,s)\Big), \tag{2}$$

where the right-hand side denotes a functional depending on the history of the displacement gradient. The theory, which we are going to present, assumes that the functional $\mathcal{F}_{0\le s\le t}$ consists of the elastic constitutive equation

$$T(x,t) = \mathcal{D}\Big(\varepsilon(x,t) - \varepsilon^p(x,t)\Big), \tag{3}$$

where $\varepsilon = \varepsilon(u) = \frac{1}{2}(\nabla u + \nabla^T u)$ is the symmetrized displacement gradient, $\varepsilon^p : \Omega \times (0,T_e) \to \mathcal{S}^3$ describes the plastic part of the deformation and $\mathcal{D} : \mathcal{S}^3 \to \mathcal{S}^3$ is the elasticity tensor, which is assumed to be constant with respect to $x \in \Omega$ and $t \in (0,T_e)$, symmetric and positive definite. This equation is coupled with the inelastic constitutive relation formulated in general as a differential inclusion

$$z_t(x,t) \in f\Big(\varepsilon(x,t), z(x,t)\Big), \tag{4}$$

where $z : \Omega \times (0,T_e) \to \mathbb{R}^N$ is the vector of internal variables. z consists of ε^p and other components $\tilde{z}$ which are introduced to describe the deformation process more appropriately. $f : D(f) \subset \mathcal{S}^3 \times \mathbb{R}^N \to \mathcal{P}(\mathbb{R}^N)$ is the constitutive multifunction and causes the system of equations (1)+(3)+(4) to become nonlinear. Thermodynamical considerations yield that there exists a free energy function $\psi : D(f) \subset \mathcal{S}^3 \times \mathbb{R}^N \to \mathbb{R}_+$ such that for all $(\varepsilon, z) \in D(f)$

$$T = \frac{\partial \rho\psi(\varepsilon,z)}{\partial \varepsilon} \qquad \text{(hyperelasticity)}, \tag{5}$$

$$\frac{\partial \rho\psi(\varepsilon,z)}{\partial z} \cdot w^* \;\le\; 0 \quad \text{for all} \quad w^* \in f(\varepsilon,z). \tag{6}$$

The existence of the free energy function ψ implies that the considered problem with $F = 0$ and with homogeneous boundary conditions possesses a natural semi-invariant, namely the total energy does not increase in time

$$\mathcal{E}(u,z)(t) \overset{\mathrm{df}}{=} \int_\Omega \rho\psi(\varepsilon,z)\,dx \le \mathcal{E}(u,z)(0). \tag{7}$$

Using the properties of the elasticity tensor $\mathcal{D}$ (5) implies that the free energy function has to be of the form

$$\rho\psi(\varepsilon,z) = \tfrac{1}{2}\mathcal{D}(\varepsilon - \varepsilon^p)\cdot(\varepsilon - \varepsilon^p) + \psi_1(z), \tag{8}$$

where the function ψ_1 depends on the vector z only. There is no precise relationship between free energy functions and constitutive multifunctions such that the reduced dissipation inequality (6) would hold. We restrict our considerations to a subclass of problems, for which

(6) will be satisfied automatically. We say that the considered problem is of pre-monotone type if the constitutive multifunction is of the form

$$f(\varepsilon,z) = g\Big(-\rho\nabla_z\psi(\varepsilon,z)\Big) \tag{9}$$

with a multifunction $g : D(g) \subset \mathbb{R}^N \to \mathcal{P}(\mathbb{R}^N)$ satisfying

$$\forall\, z \in D(g)\,, \quad \forall\, z^* \in g(z) \qquad z^* \cdot z \geq 0 \tag{10}$$

and with the free energy function ψ given as a positive semi-definite quadratic form

$$\rho\,\psi(\varepsilon,z) = \tfrac{1}{2}\mathcal{D}(\varepsilon - Bz)\cdot(\varepsilon - Bz) + \tfrac{1}{2}Lz\cdot z. \tag{11}$$

Here $L \in \mathbb{R}^{N\times N}_{\mathrm{sym}}$, $L \geq 0$ and $Bz = B(\varepsilon^p,\tilde{z}) \stackrel{\mathrm{df}}{=} \varepsilon^p$ is the orthogonal projection of the vector z on the direction ε^p. Moreover, we assume that the symmetric operator $L + B^T\mathcal{D}B$ is positive definite. This class of models was introduced by H.-D. Alber in [1, Definition 3.1.1]. Assuming that $0 \in g(0)$ we can say that (10) gives the monotonicity of g at the point 0. If additionally the inelastic constitutive multifunction is monotone then we say that the considered model is of monotone type. For such flow rules we can find in the literature many articles, (see for example [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]) which prove existence and uniqueness results in dynamical and quasistatic setting of the problem.
Some special subclass of models is very important in the theory of inelastic deformations. If the operator L is positive definite then the energy controls the whole vector z and therefore the strain tensor is also controlled by the energy in this case. Such models are called coercive in the literature and in this article we are going to work in this subclass of models only.

2. MELAN–PRAGER MODEL AND MAIN THEOREM

In this section we are going to present the Melan–Prager flow rule. The vector z of internal variables consists of two components from $\mathcal{S}^3$. The first one is the inelastic strain tensor ε^p and the second one is the so called backstress b. The flow rule is given by the following system of a differential inclusion coupled with a differential equation

$$\varepsilon^p_t \;\in\; \partial I_{\mathcal{K}}(T-b)\,, \tag{12}$$

$$b_t \;=\; \alpha\varepsilon^p_t\,. \tag{13}$$

Here, $\mathcal{K}$ denotes the set of admissible stresses, which is assumed to be of the following form $\mathcal{K} = \{T \in \mathcal{S}^3 : |PT| \leq C_K\}$, where C_K is a positive constant depending on the material under consideration. $P : \mathcal{S}^3 \to P\mathcal{S}^3$ is the projector on the deviatoric part of symmetric matrices: $PS = S - \frac{1}{3}\mathrm{tr}S\cdot I$. The function $I_{\mathcal{K}} : \mathcal{S}^3 \to \mathbb{R}_+$ is the indicator function of the set $\mathcal{K}$, which means that

$$I_{\mathcal{K}}(T) = \begin{cases} 0 & \text{for } T \in \mathcal{K}\,, \\ \infty & \text{for } T \notin \mathcal{K}\,. \end{cases} \tag{14}$$

Moreover, the positive constant α depends on the considered material. This means that in this flow rule we have two positive material constants C_K and α. The first constant describes how large is the elastic domain of the considered material and the second constant describes the hardening of the material. If α goes to zero then the considered material becomes pure elasto-plastic (this result in the dynamical case is proved in [6]).
The free energy associated with the Melan–Prager flow rule has the form

$$\rho\psi(\varepsilon,\varepsilon^p,b) = \tfrac{1}{2}\mathcal{D}(\varepsilon-\varepsilon^p)\cdot(\varepsilon-\varepsilon^p) + \tfrac{1}{2\alpha}|b|^2\,. \tag{15}$$

Then it is easy to see that the operator L is given by $Lz = L(\varepsilon^p,b) = (0,\alpha^{-1}b)$ and is semi-positive definite. Moreover, the operator $(L+B^T\mathcal{D}B)z = (\mathcal{D}\varepsilon^p,\alpha^{-1}b)$ and is positive definite. It is also very easy to verify that the flow rule of Melan–Prager is thermodynamically admissible (satisfies the dissipation inequality (6)) and is of monotone type.
The whole system of equations describing a quasistatic deformation process with the flow of Melan–Prager is in the form:

$$\begin{aligned}
\operatorname{div}_x T(x,t) &= -F(x,t)\,,\\
T(x,t) &= \mathcal{D}(\varepsilon(u(x,t)) - \varepsilon^p(x,t))\,,\\
\varepsilon(u(x,t)) &= \frac{1}{2}(\nabla_x u(x,t) + \nabla_x^T u(x,t))\,,\\
\varepsilon_t^p(x,t) &\in \partial I_K\Big(T(x,t) - b(x,t)\Big)\,,\\
b_t(x,t) &= \alpha\varepsilon_t^p(x,t)\,,
\end{aligned} \tag{16}$$

where the displacement vector u, the inelastic strain tensor ε^p and the backstress b are the unknowns. We consider system (16) with boundary conditions of mixed type. The Dirichlet boundary condition on $\Gamma_1 \subset \partial\Omega$

$$u(x,t) = g_D(x,t) \;\text{ for } x\in\Gamma_1 \;\text{ and } t\geq 0 \tag{17}$$

and the Neumann boundary condition on $\Gamma_2 \subset \partial\Omega$

$$T(x,t)\cdot n(x) = g_N(x,t) \;\text{ for } x\in\Gamma_2 \;\text{ and } t\geq 0, \tag{18}$$

where $n(x)$ is the exterior unit normal vector to the boundary $\partial\Omega$ at the point x, Γ_1 and Γ_2 are open in $\partial\Omega$, disjoint, smooth sets satisfying $\partial\Omega = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$ and $\mathcal{H}_2(\Gamma_1) > 0$, where $\mathcal{H}_2$ denotes the 2-dimensional Hausdorff measure. The functions g_D, g_N are given and describe boundary data. Finally, the initial conditions for the inelastic strain tensor and the backstress are given by

$$\varepsilon^p(x,0) = \varepsilon^{p,0}(x)\,, \tag{19}$$

$$b(x,0) = b^0(x) \tag{20}$$

with given initial data $\varepsilon^{p,0} : \Omega \to PS^3$ and $b^0 : \Omega \to PS^3$.

The operator L in the Melan–Prager model does not control all directions of the vector z and

the considered model is not coercive. However, using the very simple form of the equation (13) from the flow rule, we can integrate this equation in time and obtain

$$b(x,t) = \alpha\varepsilon^p(x,t) + b^0(x) - \alpha\varepsilon^{p,0}(x)\,. \tag{21}$$

For simplicity we assume that $b^0 = \alpha\varepsilon^{p,0}$. This assumption is not important and it is easy to see that all results obtained in this article are also true without this technical simplification. Inserting (21) into (12) we obtain the following new form of the Melan–Prager flow rule

$$\varepsilon_t^p(x,t) \in \partial I_{\mathcal{K}}(T(x,t) - \alpha\varepsilon^p(x,t))\,. \tag{22}$$

The free energy function associated with this new flow rule is now in the form

$$\rho\,\psi(\varepsilon,\varepsilon^p) = \tfrac{1}{2}\mathcal{D}(\varepsilon-\varepsilon^p)\cdot(\varepsilon-\varepsilon^p) + \tfrac{\alpha}{2}|\varepsilon^p|^2\,. \tag{23}$$

We see that the vector of internal variables is now reduced to the inelastic strain tensor only. The free energy controls the direction ε^p. Moreover, the operator L in this new setting is defined by $L\varepsilon^p = \alpha\varepsilon^p$ and is positive definite. Consequently, the considered model of Melan–Prager in this new setting is coercive. Hence, the system of equations, which we are going to study is in the form

$$\begin{aligned} \mathrm{div}_x T(x,t) &= -F(x,t)\,, \\ T(x,t) &= \mathcal{D}(\varepsilon(u(x,t)) - \varepsilon^p(x,t))\,, \\ \varepsilon_t^p(x,t) &\in \partial I_{\mathcal{K}}\Big(T(x,t) - \alpha\varepsilon^p(x,t)\Big)\,. \end{aligned} \tag{24}$$

This system will be considered with boundary conditions (17,18) and with initial condition (19) reduced to the inelastic strain only. Let us define the notion of a strong solution to system (24).

Definition 1. *Functions (u,ε^p) are called a strong solution to system (24) with boundary conditions (17,18) and with initial condition (19) if*

$$\begin{aligned} u \in \mathbb{C}([0,T_e],\mathbb{H}^1(\Omega;\mathbb{R}^3))\,, \qquad u_t \in \mathbb{L}^\infty((0,T_e),\mathbb{H}^1(\Omega;\mathbb{R}^3))\,, \\ \varepsilon^p \in \mathbb{C}([0,T_e],\mathbb{L}^2(\Omega;\mathcal{S}^3))\,, \qquad \varepsilon_t^p \in \mathbb{L}^\infty((0,T_e),\mathbb{L}^2(\Omega;\mathcal{S}^3))\,, \end{aligned}$$

the system (24) is satisfied pointwise for almost all $(x,t) \in \Omega\times[0,T]$, boundary conditions (17,18) are satisfied in the sense of traces and initial condition (19) is satisfied in classical sense.

The initial function $\varepsilon^{p,0}$ generates initial values for the stress and the displacement. Let us denote by T^0 and by u^0 the unique solution of the linear problem

$$\begin{aligned} \mathrm{div}_x T^0(x) &= -F(x,0)\,, \\ T^0(x) &= \mathcal{D}\Big(\varepsilon(u^0(x)) - \varepsilon^{p,0}(x))\Big)\,, \\ u^0(x)_{|\Gamma_1} &= g_D(x,0)\,, \quad T^0(x)\cdot n(x)_{|\Gamma_2} = g_N(x,0)\,. \end{aligned} \tag{25}$$

We will say that the initial function $\varepsilon^{p,0}$ is **admissible** if $T^0(x) - \alpha\varepsilon^{p,0}(x) \in \mathcal{K}$ for almost all $x \in \Omega$. This means that the initial value for the argument of the nonlinear operator $\partial I_{\mathcal{K}}$ belongs to the domain of this operator.

Lemma 2. *If the boundary data and the external forces have the following regularity: for all $T_e > 0$*

$$F \in \mathbb{W}^{2,\infty}((0,T_e);\mathbb{L}^2(\Omega;\mathbb{R}^3)),$$
$$g_D \in \mathbb{W}^{3,\infty}((0,T_e);\mathbb{H}^{\frac{1}{2}}(\Gamma_1;\mathbb{R}^3)), \quad g_N \in \mathbb{W}^{2,\infty}((0,T_e);\mathbb{H}^{-\frac{1}{2}}(\Gamma_2;\mathbb{R}^3)), \tag{26}$$

and the initial inelastic strain tensor $\varepsilon^{p,0} \in \mathbb{L}^2(\Omega;P\mathcal{S}^3)$ is admissible then system (24) with boundary conditions (17,18) and with initial condition (19) possesses a global in time, unique strong solution.

Proof. This lemma follows directly from [2], where a similar result is proved for all coercive models. □

Now we are ready to present the problem, which we are going to study in this article. In the engineering practice quasistatic models from the theory of inelastic deformations are used in the numerical analysis of observed real deformation processes. In the Malan–Prager model two material constants appear, which are determined experimentally only. For this reason it is important to study continuous dependence of solutions with respect to material constants. This kind of continuity is called material stability in the literature. Let us assume that we have two convergent positive sequences $\alpha^l \to \alpha$ and $C_K^l \to C_K$ and the limits are also positive. Let us consider strong solutions $(u^l, \varepsilon^{p,l})$ to the system

$$\begin{aligned} \operatorname{div}_x T^l(x,t) &= -F(x,t), \\ T^l(x,t) &= \mathcal{D}(\varepsilon(u^l(x,t)) - \varepsilon^{p,l}(x,t)), \\ \varepsilon_t^{p,l}(x,t) &\in \partial I_{\mathcal{K}^l}\Big(T^l(x,t) - \alpha^l\varepsilon^{p,l}(x,t)\Big). \end{aligned} \tag{27}$$

where $\mathcal{K}^l = \{T \in \mathcal{S}^3 : |PT| \le C_K^l\}$. This system is considered with boundary conditions (17,18) and with initial condition (19). We are going to prove that the sequence of strong solutions converges to the strong solution of (24).

Theorem 3. *Let us assume that the boundary data, the external forces and the initial data have the regularity from Lemma 2. Moreover, assume that the initial data is admissible for every l. Then the sequence of strong solutions $(u^l, \varepsilon^{p,l})$ to problem (27) with boundary conditions (17,18) and with initial condition (19) converges in the topology $\mathbb{C}([0,T_e];\mathbb{H}^1(\Omega;\mathbb{R}^3) \times \mathbb{L}^2(\Omega;\mathcal{S}^3))$ to the strong solution of (24).*

3. UNIFORM ENERGY ESTIMATES

In this section we will prove uniform estimates for the sequence of strong solutions to (27). We begin with an easy observation that the initial displacement vector and the initial stress defined by (26) do not depend on l. This follows from the fact that the boundary data, the initial data and the external forces in initial-boundary problem (27) do not depend on l.

Theorem 4. *Let us assume that the boundary data, the external forces and the initial data satisfy all requirements from Theorem 3. Moreover, assume that $(u^l, \varepsilon^{p,l})$ is the strong solution to problem (27) with boundary conditions (17,18) and with initial condition (19). Then there exists a positive constant $C = C(T_e)$, which does not depend on l such that for all $t \in (0, T_e)$ the following inequality holds*

$$\mathcal{E}^l(u^l, \varepsilon^{p,l})(t) \stackrel{\text{df}}{=} \frac{1}{2}\int_\Omega \mathcal{D}(\varepsilon(u^l) - \varepsilon^{p,l}) \cdot (\varepsilon(u^l) - \varepsilon^{p,l})\, dx + \frac{\alpha^l}{2}\int_\Omega |\varepsilon^{p,l}|^2\, dx \le C(T_e). \quad (28)$$

Proof. This theorem can be concluded from the general energy estimate in [9]. However, this result is very crucial in the proof of the main theorem and therefore we present the whole proof in this article. Let us calculate the time derivative of the energy $\mathcal{E}^l(u^l, \varepsilon^{p,l})(t)$.

$$\begin{aligned}\frac{d}{dt}\mathcal{E}^l(u^l, \varepsilon^{p,l})(t) &= \int_\Omega \mathcal{D}(\varepsilon(u^l_t) - \varepsilon^{p,l}_t) \cdot (\varepsilon(u^l) - \varepsilon^{p,l})\, dx + \alpha^l \int_\Omega \varepsilon^{p,l}_t \cdot \varepsilon^{p,l}\, dx \\ &= \int_\Omega \varepsilon(u^l_t) \cdot T^l\, dx - \int_\Omega \varepsilon^{p,l}_t \cdot (T^l - \alpha^l \varepsilon^{p,l})\, dx \le \int_\Omega \nabla u^l_t \cdot T^l\, dx \qquad (29) \\ &= \int_\Omega u^l_t F\, dx + \int_{\Gamma_1} g_D T^l n\, dS + \int_{\Gamma_2} u^l_t g_N\, dS.\end{aligned}$$

Integrating inequality (29) on $(0,t)$ we have

$$\begin{aligned}\mathcal{E}^l(u^l, \varepsilon^{p,l})(t) &\le \mathcal{E}^l(u^l, \varepsilon^{p,l})(0) \\ &\quad + \int_0^t \int_\Omega u^l_t F\, dx\, d\tau + \int_0^t \int_{\Gamma_1} g_D T^l n\, dS\, d\tau + \int_0^t \int_{\Gamma_2} u^l_t g_N\, dS\, d\tau. \qquad (30)\end{aligned}$$

Let us denote the three integrals on the right hand side of (30) by K_i for $i = 1,2,3$. From the observation that the initial displacement does not depend on l we see that the initial energy $\mathcal{E}^l(u^l, \varepsilon^{p,l})(0)$ is constant. Next we will estimate the integrals K_i for $i = 1,2,3$. The energy function does not depend on the time derivative of the displacement vector. Therefore we write K_1 in another form integrating by parts with respect to t.

$$\int_0^t \int_\Omega u^l_t F\, dx\, d\tau = -\int_0^t \int_\Omega u^l F_t\, dx\, d\tau + \int_\Omega u^l F\, dx - \int_\Omega u^0 F(0)\, dx. \quad (31)$$

We see that the last integral on the right hand side of (31) is equal to a constant and we have to estimate the remaining integrals.

$$\left|\int_0^t \int_\Omega u^l F_t\, dx\, d\tau\right| + \left|\int_\Omega u^l F\, dx\right| \le (1 + T_e) \sup_{(0,t)} [\|u^l\|_{\mathbb{L}^2(\Omega)} (\|F_t\|_{\mathbb{L}^2(\Omega)} + \|F\|_{\mathbb{L}^2(\Omega)})]. \quad (32)$$

Using the Poincare inequality we have

$$\|u^l\|_{\mathbb{L}^2(\Omega)} \leq D(\|\varepsilon(u^l)\|_{\mathbb{L}^2(\Omega)} + \|g_D\|_{\mathbb{H}^{1/2}(\Gamma_1)}), \tag{33}$$

where the constant D depends on the set Ω only. From the elastic constitutive relation we also have that

$$\|\varepsilon(u^l)\|_{\mathbb{L}^2(\Omega)} \leq \|\mathcal{D}^{-1}T^l\|_{\mathbb{L}^2(\Omega)} + \|\varepsilon^{p,l}\|_{\mathbb{L}^2(\Omega)} \leq L\sqrt{\mathcal{E}^l(u^l,\varepsilon^{p,l})}, \tag{34}$$

where the positive constant L does not depend on l. In the last estimation we have used the fact that the sequence α^l is bounded from below and from above by positive constants. Inserting (34) into (33) and the resulting inequality into (32) we conclude that for $\eta > 0$ there exists $C(\eta, T_e) > 0$ such that

$$\begin{aligned}\left|\int_0^t \int_\Omega u^l F_t \, dx\, d\tau\right| &\leq \eta \mathcal{E}^l(u^l,\varepsilon^{p,l}) \\ &\quad + C(\eta,T_e)\sup_{(0,t)}(\|F_t\|^2_{\mathbb{L}^2(\Omega)} + \|F\|^2_{\mathbb{L}^2(\Omega)} + \|g_D\|^2_{\mathbb{H}^{1/2}(\Gamma_1)} + 1).\end{aligned} \tag{35}$$

Next we are going to will estimate K_2.

$$\left|\int_0^t \int_{\Gamma_1} g_D T^l n \, dS\, d\tau\right| \leq \sup_{(0,t)} \|g_{D,t}\|_{\mathbb{H}^{1/2}(\Gamma_1)} \sup_{(0,t)} \|T^l n\|_{\mathbb{H}^{-1/2}(\Gamma_1)}. \tag{36}$$

According to the trace theorem in the space $\mathbb{L}^2(\mathrm{div})$ (this space contains all vector fields from $\mathbb{L}^2(\Omega)$ which weak divergence belongs also to $\mathbb{L}^2(\Omega)$) we have

$$\begin{aligned}\|T^l n\|_{\mathbb{H}^{-1/2}(\Gamma_1)} &\leq \|T^l n\|_{\mathbb{H}^{-1/2}(\partial\Omega)} \\ &\leq M(\|\mathrm{div} T^l\|_{\mathbb{L}^2(\Omega)} + \|T^l\|_{\mathbb{L}^2(\Omega)}) \leq \tilde{M}(\|F\|_{\mathbb{L}^2(\Omega)} + \sqrt{\mathcal{E}^l(u^l,\varepsilon^{p,l})}),\end{aligned} \tag{37}$$

where the constants $M, \tilde{M}$ do not depend on l. Inserting (37) into (36) we obtain that for every positive κ there exists a positive constant $C(\kappa, T_e)$ such that

$$\left|\int_0^t \int_{\Gamma_1} g_D T^l n \, dS\, d\tau\right| \leq \kappa a \mathcal{E}^l(u^l,\varepsilon^{p,l}) + C(\kappa,T_e)\sup_{(0,t)}(\|F\|^2_{\mathbb{L}^2(\Omega)} + \|g_{D,t}\|^2_{\mathbb{H}^{1/2}(\Gamma_1)}). \tag{38}$$

Finally, we have to estimate integral K_3. First we shift the time derivative from the displacement vector onto the data.

$$\int_0^t \int_{\Gamma_2} u^l_t g_N \, dS\, d\tau = -\int_0^t \int_{\Gamma_2} u^l g_{N,t} \, dS\, d\tau + \int_{\Gamma_2} u^l g_N \, dS - \int_{\Gamma_2} u^0 g_N(0) \, dS. \tag{39}$$

The last integral on the right hand side of (39) is constant and we only need to estimate the other two.

$$\left|\int_0^t \int_{\Gamma_2} u^l g_{N,t} \, dS\, d\tau\right| + \left|\int_{\Gamma_2} u^l g_N \, dS\right| \leq T_e \sup_{(0,t)}[\|u^l\|_{\mathbb{H}^{1/2}(\partial\Omega)}(\|g_{N,t}\|_{\mathbb{H}^{-1/2}(\Gamma_2)} + \|g_N\|_{\mathbb{H}^{-1/2}(\Gamma_2)})]. \tag{40}$$

Using the trace theorem in the space $\mathbb{H}^1(\Omega)$ and the Poincare inequality we get

$$\|u^l\|_{\mathbb{H}^{1/2}(\partial\Omega)} \leq E\|u^l\|_{\mathbb{H}^1(\Omega)} \leq \tilde{E}(\|\varepsilon(u^l)\|_{\mathbb{L}^2(\Omega)} + \|g_D\|_{\mathbb{H}^{1/2}(\Gamma_1)}), \tag{41}$$

where the constants $E, \tilde{E}$ do not depend on l. Using (34) we can conclude that for positive constant μ there exists a positive constant $C(\mu, T_E)$ such that

$$\left|\int_0^t \int_{\Gamma_2} u_t^l g_N \, dS \, d\tau\right| \leq \mu a \mathcal{E}^l(u^l, \varepsilon^{p,l}) + C(\mu, T_e) \sup_{(0,t)} (\|g_{N,t}\|^2_{\mathbb{H}^{-1/2}(\Gamma_2)} + \|g_N\|^2_{\mathbb{H}^{-1/2}(\Gamma_2)} + \|g_D\|_{\mathbb{H}^{1/2}(\Gamma_1)} + 1). \tag{42}$$

Adding estimates (35),(38) and (42) side to side and choosing constants η, κ, μ so that $\eta + \kappa + \mu < 1$ we end the proof. □

4. UNIFORM ENERGY ESTIMATES FOR TIME DERIVATIVES

In the previous section we proved that the sequence $\{u^l\}$ is bounded in the space $\mathbb{C}([0,T_e], \mathbb{H}^1(\Omega;\mathbb{R}^3))$ and the sequence $\{\varepsilon^{p,l}\}$ is bounded in the space $\mathbb{C}([0,T_e], \mathbb{L}^2(\Omega;\mathcal{S}^3))$. The next step is to obtain uniform estimates for the time derivatives of the strong solutions to system (27).

Theorem 5. *Let us assume that the boundary data, the external forces and the initial data satisfy all requirements from Theorem 3. Moreover, assume that $(u^l, \varepsilon^{p,l})$ is the strong solution to problem (27) with boundary conditions (17,18) and with initial condition (19). Then there exists a positive constant $\tilde{C} = \tilde{C}(T_e)$ which does not depend on l and such that for all $t \in (0, T_e)$ the energy evaluated on the time derivatives can be estimated as follows*

$$\mathcal{E}^l(u_t^l, \varepsilon_t^{p,l})(t) \stackrel{\text{df}}{=} \frac{1}{2}\int_\Omega \mathcal{D}(\varepsilon(u_t^l) - \varepsilon_t^{p,l}) \cdot (\varepsilon(u_t^l) - \varepsilon_t^{p,l}) \, dx + \frac{\alpha^l}{2}\int_\Omega |\varepsilon_t^{p,l}|^2 \, dx \leq \tilde{C}(T_e). \tag{43}$$

Proof. Let us denote by $(\varepsilon(u^l)_h, \varepsilon_h^{p,l})$ the shifted functions $(\varepsilon(u^l)(x,t+h), \varepsilon^{p,l}(x,t+h))$ for $h \in (0,T_e)$. Let us define by v^l the velocity vector u_t^l. For simplicity we will write $\varepsilon^l = \varepsilon(u^l)$.

Calculating the time derivative of the function $\mathcal{E}^l(\varepsilon_h^l-\varepsilon^l,\varepsilon_h^{p,l}-\varepsilon^{p,l})$ we obtain

$$\begin{aligned}\frac{d}{dt}\mathcal{E}^l(\varepsilon_h^l-\varepsilon^l,\varepsilon_h^{p,l}-\varepsilon^{p,l}) &= \int_\Omega \mathcal{D}(\varepsilon(v_h^l)-\varepsilon(v^l)-\varepsilon_{h,t}^{p,l}+\varepsilon_t^{p,l})\cdot(\varepsilon_h^l-\varepsilon^l-\varepsilon_h^{p,l}+\varepsilon^{p,l})\,dx\\ &\quad+\alpha^l\int_\Omega(\varepsilon_{h,t}^{p,l}+\varepsilon_t^{p,l})\cdot(\varepsilon_h^{p,l}-\varepsilon^{p,l})\,dx\\ &= \int_\Omega(\varepsilon(v_h^l)-\varepsilon(v^l))\cdot(T_h^l-T^l)\,dx\\ &\quad-\int_\Omega(\varepsilon_{h,t}^{p,l}+\varepsilon_t^{p,l})\cdot(T_h^l-\alpha^l\varepsilon_h^{p,l}+T^l-\alpha^l\varepsilon^{p,l})\,dx \qquad (44)\\ &\le \int_\Omega(\nabla v_h^l-\nabla v^l)(T_h^l-T^l)\,dx\\ &= \int_\Omega(v_h^l-v^l)(F_h-F)\,dx+\int_{\Gamma_1}(g_{D,h}^t-g_D^t)(T_h^l-T^l)n\,dS\\ &\quad+\int_{\Gamma_2}(v_h^l-v^l)(g_{N,h}-g_N)\,dS,\end{aligned}$$

where g_D^t denotes the time derivative $\partial_t g_D$ and $F_h, v_h, g_{D,h}^t, g_{N,h}$ denote the shifted functions F, v^l, g_D^t, g_N respectively. Next we integrate (44) on $(0,t)$, shift all difference operators onto the given data, divide by h^2 and go to the limit as $h\to 0^+$. Thus, we obtain the inequality

$$\begin{aligned}\mathcal{E}^l(v^l,\varepsilon_t^{p,l})(t) &\le \mathcal{E}^l(u_t^l,\varepsilon_t^{p,l})(0)+\int_0^t\|F_{tt}\|_{\mathbb{L}^2(\Omega)}\|v^l\|_{\mathbb{L}^2(\Omega)}\,d\tau\\ &\quad+B(T_e)(\sup_{(0,t)}\|g_{D,ttt}\|_{\mathbb{H}^{1/2}(\Gamma_1)}+\sup_{(0,t)}\|g_{D,tt}\|_{\mathbb{H}^{1/2}(\Gamma_1)}+1)\sup_{(0,t)}\|T^l n\|_{\mathbb{H}^{-1/2}(\partial\Omega)} \qquad (45)\\ &\quad+B(T_e)(\sup_{(0,t)}\|g_{N,tt}\|_{\mathbb{H}^{-1/2}(\Gamma_2)}+\sup_{(0,t)}\|g_{N,t}\|_{\mathbb{H}^{-1/2}(\Gamma_2)}+1)\sup_{(0,t)}\|v^l\|_{\mathbb{H}^{1/2}(\partial\Omega)}\\ &\quad+\sup_{(0,t)}\|F_t\|_{\mathbb{L}^2(\Omega)}\|v^l\|_{\mathbb{L}^2(\Omega)}+B(T_e),\end{aligned}$$

where the positive constant $B(T_e)$ does not depend on l. From the admissibility of the initial data we conclude that the sequence of norms $\{\|\varepsilon_t^{p,l}(0)\|_{\mathbb{L}^2(\Omega)}\}$ is bounded (the sequence $\{\|T^)-\alpha\varepsilon^{p,0}\|_{\mathbb{L}^2(\Omega)}\}$ is bounded and the operator $\partial I_{\mathcal{K}}$ is maximal monotone). Hence, the sequence $\{\mathcal{E}^l(u_t^l,\varepsilon_t^{p,l})(0)\}$ is finite and constant. We estimate the boundary norm $\|T^l n\|_{\mathbb{H}^{-1/2}(\partial\Omega)}$ using (37). From the Poincare inequality and from the coerciveness of the flow rule we estimate the sequence $\{v^l\}$ in the space $\mathbb{L}^2(\Omega)$ in the way similar to the method used in (33) and in (34). Finally, we estimate the boundary norm $\|v^l\|_{\mathbb{H}^{1/2}(\partial\Omega)}$ as in (41). Hence, inserting all these estimates into (45) and using the method from the proof of Theorem 4 we end the proof. □

5. STRONG CONVERGENCE OF STRESSES AND STRAINS

The boundedness of the sequence of strong solutions to system (27) obtained in Theorem 4 implies the existence of a weakly convergent subsequence. In this section we are going to prove that the whole sequence converges strongly.

Theorem 6. *Assume that the boundary data, the external forces and the initial data satisfy all requirements from Theorem 3. Moreover, assume that $(u^l, \varepsilon^{p,l})$ is the strong solution to problem (27) with boundary conditions (17,18) and with initial condition (19). Then $T^l \to T$, and $\varepsilon^{p,l} \to \varepsilon^p$ in the space $\mathbb{C}([0,T_e],\mathbb{L}^2(\Omega;\mathcal{S}^3))$.*

Proof. This result is crucial in the proof of the main theorem. Let us calculate the time derivative of the limit energy function

$$\mathcal{E}^\infty(u,\varepsilon^p) = \frac{1}{2}\int_\Omega \mathcal{D}(\varepsilon(u)-\varepsilon^p)\cdot(\varepsilon(u)-\varepsilon^p)\,dx + \frac{\alpha}{2}\int_\Omega \|\varepsilon^p\|^2\,dx$$

evaluated on the differences $(u^l-u^k, \varepsilon^{p,l}-\varepsilon^{p,k})$.

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}^\infty(u^l-u^k,\varepsilon^{p,l}-\varepsilon^{p,k}) &= \int_\Omega \mathcal{D}(\varepsilon_t^l-\varepsilon_t^k-\varepsilon_t^{p,l}+\varepsilon_t^{p,k})\cdot(\varepsilon^l-\varepsilon^k-\varepsilon^{p,l}+\varepsilon^{p,k})\,dx \\
&\quad + \alpha\int_\Omega(\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot(\varepsilon^{p,l}-\varepsilon^{p,k})\,dx \qquad (46)\\
&= \int_\Omega(\varepsilon(u_t^l)-\varepsilon(u_t^k))\cdot(T^l-T^k)\,dx - \int_\Omega(\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot(T^l-T^k-\alpha\varepsilon^{p,l}+\alpha\varepsilon^{p,k})\,dx \\
&= \int_\Omega(\nabla u_t^l-\nabla u_t^k)\cdot(T^l-T^k)\,dx - \int_\Omega(\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot(T^l-T^k-\alpha^l\varepsilon^{p,l}+\alpha^k\varepsilon^{p,k})\,dx \\
&\quad + (\alpha-\alpha^l)\int_\Omega(\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot\varepsilon^{p,l}\,dx + (\alpha^k-\alpha)\int_\Omega(\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot\varepsilon^{p,k}\,dx \\
&\stackrel{\text{df}}{=} I_1+I_2+I_3+I_4\,.
\end{aligned}$$

It is easy to see that $I_1 = 0$ because system (27) is considered under the assumption that boundary data and external forces do not depend on l. Moreover, we also see that

$$\begin{aligned}
|I_3|+|I_4| &\le |\alpha^l-\alpha|\big(\sup_{(0,T_e)}[(\|\varepsilon_t^{p,l}\|_{\mathbb{L}^2(\Omega)}+\|\varepsilon_t^{p,k}\|_{\mathbb{L}^2(\Omega)})\|\varepsilon^{p,l}\|_{\mathbb{L}^2(\Omega)}] \\
&\quad + |\alpha^k-\alpha|\big(\sup_{(0,T_e)}[(\|\varepsilon_t^{p,l}\|_{\mathbb{L}^2(\Omega)}+\|\varepsilon_t^{p,k}\|_{\mathbb{L}^2(\Omega)})\|\varepsilon^{p,k}\|_{\mathbb{L}^2(\Omega)}] \qquad (47)\\
&\le (|\alpha^l-\alpha|+|\alpha^k-\alpha|)2(\tilde{C}(T_e)+C(T_e))\,,
\end{aligned}$$

where the positive constants $C(T_e)\,,\tilde{C}(T_e)$ are from Theorem 4 and Theorem 5 respectively. Let us denote by Π_r the orthogonal projector of $\mathcal{S}^3$ onto the cylinder $K_r = \{T\in\mathcal{S}^3 : |PT|\le r\}$. We will use this projector in the estimate of I_2. Let us assume that $C_K^l > C_K^k$.

Then

$$
\begin{aligned}
&-\int_\Omega (\varepsilon_t^{p,l}-\varepsilon_t^{p,k})\cdot(T^l-T^k-\alpha^l\varepsilon^{p,l}+\alpha^k\varepsilon^{p,k})\,dx\\
&\quad=-\int_\Omega \varepsilon_t^{p,l}\cdot(T^l-T^k-\alpha^l\varepsilon^{p,l}+\alpha^k\varepsilon^{p,k})\,dx\\
&\qquad+\int_\Omega \varepsilon_t^{p,k}\cdot(\Pi_{C_K^k}(T^l-\alpha^l\varepsilon^{p,l})-T^k+\alpha^k\varepsilon^{p,k})\,dx\\
&\qquad+\int_\Omega \varepsilon_t^{p,k}\cdot(T^l+\alpha^l\varepsilon^{p,l}-\Pi_{C_K^k}(T^l-\alpha^l\varepsilon^{p,l}))\,dx\\
&\quad\le |C_K^l-C_K^k|\|\varepsilon_t^{p,k}\|_{\mathbb{L}^2(\Omega)}\|T^l-\alpha^l\varepsilon^{p,l}\|_{\mathbb{L}^2(\Omega)}\le D|C_K^l-C_K^k|(\tilde{C}(T_e)+C(T_e))
\end{aligned}
\tag{48}
$$

where the positive constant D does not depend on l and k. The case $C_K^k > C_K^l$ can be considered on the same way and if $C_K^l = C_K^k$ then $I_2 \le 0$. Inserting (47) and (48) into (46) and using the fact that system (27) is considered with initial data independent of l and k we end the proof. □

6. PROOF OF THE MAIN THEOREM

From Theorem 6 we have that $u^l \to u$ in the space $\mathbb{C}([0,T_e],\mathbb{H}^1(\Omega;\mathbb{R}^3))$. To end the proof of Theorem 3 it remains to prove that (u,ε^p) is the strong solution to system (24).

Proof. The boundedness of the time derivatives of the sequence $(u^l,\varepsilon^{p,l})$ implies that $u_t^l \overset{*}{\rightharpoonup} u_t$ in the space $\mathbb{L}^\infty((0,T_e),\mathbb{H}^1(\Omega;\mathbb{R}^3))$ and that $\varepsilon_t^{p,l} \overset{*}{\rightharpoonup} \varepsilon_t^p$ in the space $\mathbb{L}^\infty((0,T_e),\mathbb{L}^2(\Omega;\mathcal{S}^3))$. Consequently, the functions (u,ε^p) have the regularity required by Definition 1. It is easy to see that the limit stress T satisfies the balance of forces $\mathrm{div}_x T = -F$ and the linear elastic constitutive relation is also satisfied by $T, \varepsilon(u)$ and ε^p. Moreover, since the boundary data and the initial data do not depend on l we immediately obtain that the limit functions satisfy (17,18) and (19). To end the proof we have to show that the limit functions satisfy the Melan–Prager flow rule. For all l we have that $|P(T^l-\alpha^l\varepsilon^{p,l})| \le C_K^l$. From Theorem 6 sequences of stresses and inelastic strains converge strongly in the space $\mathbb{C}([0,T_e],\mathbb{L}^2(\Omega;\mathcal{S}^3))$. Then possibly switching to a subsequence we have also the pointwise convergence for almost all $(x,t)\in\Omega\times[0,T_e]$. Hence, we conclude that $|P(T-\alpha\varepsilon^p)| \le C_K$. Let us assume that a test stress $S\in\mathcal{K}$. It remains to prove that

$$\varepsilon_t^p\cdot(S-T+\alpha\varepsilon^p)\le 0$$

for almost $(x,t)\in\Omega\times(0,T_e)$. Let $\varphi:\Omega\times(0,T_e)\to\mathbb{R}$ be a nonnegative smooth function with compact support. Since $(u^l,\varepsilon^{p,l})$ is the strong solution to system (27) we have that

$$\int_0^{T_e}\int_\Omega \varepsilon_t^{p,l}\cdot(S-T^l+\alpha^l\varepsilon^{p,l})\varphi\,dx\,d\tau\le 0\,. \tag{49}$$

Going to the limit as $l\to\infty$ we obtain (49) for the limit functions and the proof is complete. □

Remark

Using the same methods as presented in this article we can prove that the solution operator to system (24) is also continuous with respect to the external forces, the boundary data and the initial data.

References

[1] Alber H.-D., *Materials with memory*, Lecture Notes in Math., **1682**, Springer, Berlin Heidelberg New York, 1998.

[2] Alber H.-D., Chełmiński K., *Quasistatic problems in viscoplasticity theory I: Models with linear hardening*. in I. Gohberg et al. Operator theoretical methods and applications to mathematical physics. The Erhard Meister memorial volume. 105-129, Birkhäuser, Basel, 2004.

[3] Alber H.-D., Chełmiński K., *Quasistatic problems in viscoplasticity theory II: Models with nonlinear hardening*. Math. Meth. Mod. in App. Sci. **17** (2), 189-213, 2007.

[4] Chełmiński K., Coercive limits for a subclass of monotone constitutive equations*i*n the theory of inelastic material behaviour of metals, Mat. Stos. **40**, 41–81, 1997.

[5] Chełmiński K., *On monotone plastic constitutive equations with polynomial growth condition*, Math. Meth. App. Sci. **22**, 547–562, 1999.

[6] Chełmiński K. *Coercive approximation of viscoplasticity and plasticity*, Asymptot. Anal. **26**, 115-135, 2001.

[7] Chełmiński K. *Global existence of weak-type solutions for models of monotone type in the theory of inelastic deformations* Math. Meth. in App. Sci. **25**, 1195-1230, 2002.

[8] Chełmiński K. *Coercive and self-controlling quasistatic models of the gradient type with convex composite inelastic constitutive equations*, Cent. Eur. J. Math., **1**, 670-689, 2003.

[9] Chełmiński K., Gwiazda P. *Convergence of coercive approximations for strictly monotone quasistatic models in inelastic deformation theory* Math. Meth. App. Sci., **30** (12), 1357-1374, 2007.

[10] Chełmiński K., Naniewicz Z., *Coercive limits for constitutive equations of monotone-gradient type*, Nonlinear Anal. TMA **48**, 1197-1214, 2002.

[11] Kisiel K., Chełmiński K., *Prandtl–Reuss dynamical elasto-perfect plasticity without safe-load conditions*, Nonlinear Anal. TMA **192**, 1-28, 2020.

[12] Kisiel K., Chełmiński K., *On strong solutions of viscoplasticity without safe-load conditions*, J. Diff. Eqn. **269**(3), 2264-2327, 2020.

[13] R. Temam, *A generalized Norton-Hoff model and the Prandtl-Reuss law of plasticity*, Arch. Rational Mech. Anal. **95**, 137–183, 1986.

Wojciech Domitrz, Stanisław Janeczko

Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

HAMILTONIAN VECTOR FIELDS ON SINGULAR VARIETIES

Manuscript received: 15 August 2020
Manuscript accepted: 31 August 2020

Abstract: We define Hamiltonian vector fields on singular subvarieties of the symplectic space. We describe Hamiltonian vector fields on smooth submanifolds, singular planar curves with ADE singularities and regular union singularities.
Keywords: symplectic geometry, Hamiltonian vector fields, singularities
Mathematics Subject Classification (2020): 58K50 (primary), 37J39

1. INTRODUCTION

Let M be a smooth $2n$-dimensional manifold, endowed with a nondegenerate, closed 2-form ω. The 2-form ω is called symplectic and the pair (M,ω) is a symplectic manifold. We introduce the canonical symplectic structure $\dot{\omega}$ on TM using the vector bundle morphism $\beta : TM \ni u \mapsto \omega(u,\cdot) \in T^*M$, namely the pullback of the Liouville symplectic form $d\theta$ defined on the cotangent bundle T^*M, $\dot{\omega} = \beta^* d\theta$. A smooth vector field $X : M \to TM$ is said to be Hamiltonian if the form $\omega(X,\cdot)$ is exact. A function $H : M \to \mathbb{R}$ is called Hamiltonian for X if $\omega(X,\cdot) = -dH(\cdot)$. If X is Hamiltonian, then its image $X(M) \subset TM$ is a Lagrangian submanifold of $(TM,\dot{\omega})$ generated by H. In local Darboux coordinates, $M \cong \mathbb{R}^{2n}$, $\omega = \sum_{i=1}^n dy_i \wedge dx_i$, and $\dot{\omega} = \beta^* d\theta = \sum_{i=1}^n (d\dot{y}_i \wedge dx_i - d\dot{x}_i \wedge dy_i)$, where $(q,\dot{q}) = ((x,y),(\dot{x},\dot{y}))$ are coordinates on $T\mathbb{R}^{2n} \equiv \mathbb{R}^{2n} \times \mathbb{R}^{2n}$.

To generalize this notion, we introduce a concept of a Hamiltonian system as a general Lagrangian submanifold N of the symplectic tangent bundle $(TM,\dot{\omega})$. If $\tau|_N : N \to M$ is singular, where τ is tangent bundle projection, we also call N an implicit Hamiltonian system (cf. [12], [7]). Important property of such systems around singularities is their solvability, i.e. existence of smooth local curve $\gamma : (-\varepsilon,\varepsilon) \to M$ such that its tangent lifting $\dot{\gamma}(t)$ belongs to N around each point of N. An immediate necessary condition for solvability

is tangential solvability condition, which is satisfied if $\dot{q} \in d(\tau\mid_N)_v(T_vN)$ for each point $v=(q,\dot{q}) \in N$. It is proved (cf. [7]) that, for certain naturally generic implicit Hamiltonian systems, they are solvable if they fulfill this tangential solvability condition. Another generalization following P.A.M. Dirac (cf. [3]) is provided by constrained Lagrangian submanifolds (cf. [11]) as Hamiltonian systems. The generalized Hamiltonian function for such system is a generating family (Morse family) for the corresponding Lagrangian submanifold L_h; $F(x,y,\lambda)=\sum_{i=1}^{k} a_i(x,y)\lambda_i + h(x,y)$ over the constraint K defined by smooth functions $a_i(x,y)=0$. The condition of solvability $\{\frac{\partial F}{\partial \lambda_i}, F\}=0$ for $(x,y,\lambda)\in S\times\mathbb{R}^{2n}$ defines the section of L_h which is tangent to K. The general sections of L_h give the vector fields which are Hamiltonian on the constrained submanifold.

In this work we concentrate on the vector fields of symplectic space (M,ω), which are Hamiltonian on a subvariety of M. As we do not exclude singularities, our approach is local and we consider mainly germs of subvarieties and germs of vector fields. We find the spaces of vector fields, which are Hamiltonian on symplectic, isotropic and coisotropic submanifolds of (M,ω) and we provide the classification of Hamiltonian vector fields on singular varieties: planar curves of type A_k, D_k, E_6, E_7, E_8, regular union of three 1-dimensional submanifolds, regular union of two 2-dimensional isotropic submanifolds, and regular union of two 2-dimensional symplectic submanifolds. We use the Mathematica package Exterior Differential Calculus for calculations.

2. HAMILTONIAN SYSTEMS ON SUBMANIFOLDS

Let K be a submanifold of $\mathbb{R}^{2n}$ and $h: K\to\mathbb{R}$ be a smooth function on K. The notion of generalized Hamiltonian system (generalized Hamiltonian dynamics) was introduced by P.A.M. Dirac in [3]. A generalized Hamiltonian system is the following sub-bundle L_h of $T\mathbb{R}^{2n}$ over K (cf. [13]):

$$L_h=\{v\in T\mathbb{R}^{2n} : \omega(v,u)=-dh(u) \quad \forall_{u\in TK}\}. \tag{1}$$

It is easy to see that L_h is a Lagrangian submanifold of $(T\mathbb{R}^{2n},\dot{\omega})$.

In local coordinates, the generalized Hamiltonian system (1) can be written, using generating family $F:\mathbb{R}^{2n}\times\mathbb{R}^k\to\mathbb{R}$, in the following way:

$$F(x,y,\lambda)=\sum_{\ell=1}^{k} a_\ell(x,y)\lambda_\ell + H(x,y), \tag{2}$$

where K is defined as a zero-level set of the mapping $a:(x,y)\mapsto(a_1(x,y),\ldots,a_k(x,y))$, $H(x,y)$ is an arbitrary smooth extension of the function $h:K\to\mathbb{R}$ and a is a maximal rank map-germ.

The generalized Hamiltonian system L is given by an immersion $\phi : C_F \to L \subset (T\mathbb{R}^{2n}, \dot{\omega})$ defined by

$$\phi(x,y,\lambda) = (x,y,\frac{\partial F}{\partial y}(x,y,\lambda), -\frac{\partial F}{\partial x}(x,y,\lambda)), \quad (x,y,\lambda) \in C_F.$$

Since $\frac{\partial F}{\partial \lambda_\ell}(x,y,\lambda) = a_\ell(x,y)$, we have $C_F = K \times \mathbb{R}^k$. Then L can be described as

$$L = \phi(C_F) = \{(x,y,\frac{\partial F}{\partial y}(x,y,\lambda), -\frac{\partial F}{\partial x}(x,y,\lambda)) \in T\mathbb{R}^{2n} : (x,y,\lambda) \in K \times \mathbb{R}^k\}.$$

L is a skew-conormal bundle to K and its smooth sections are called Hamiltonian systems on K with Hamiltonian H. This may be extended to Hamiltonian system on M taking Hamiltonian function

$$F(x,y) = \sum_{l=1}^{k} \lambda_l(x,y) a_l(x,y) + H(x,y)$$

for some smooth functions $\lambda_l(x,y)$.

Vector fields, which are Hamiltonian on K are given in the form:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \lambda_j(x,y)(\frac{\partial a_j}{\partial y_i}(x,y)\frac{\partial}{\partial x_i} - \frac{\partial a_j}{\partial x_i}(x,y)\frac{\partial}{\partial y_i}) + \sum_{i=1}^{n}(\frac{\partial H}{\partial y_i}(x,y)\frac{\partial}{\partial x_i} - \frac{\partial H}{\partial x_i}(x,y)\frac{\partial}{\partial y_i}). \tag{3}$$

If we consider the functions $\lambda_j(x,y)$ which are smooth solutions of the system of linear equations (cf. [8]),

$$\sum_{j=1}^{k} \{a_i, a_j\}(x,y)\lambda_j = \{H, a_i\}(x,y), \quad i = 1,\ldots,k, \tag{4}$$

then the vector fields (3) are the logarithmic Hamiltonian vector fields over K.

3. HAMILTONIAN VECTOR FIELDS ON SINGULAR VARIETIES

Let (M,ω) be a symplectic manifold. Let N be a subset of M.

Definition 1. *A smooth vector field X on M is called* **Hamiltonian on** *N if there exists a smooth function H on M such that*

$$(X \rfloor \omega)|_x = -dH|_x, \text{ for every } x \in N. \tag{5}$$

Example 2. *Let $(\mathbb{R}^{2n},\omega_0)$ be the standard symplectic space. Let $N\subset(\mathbb{R}^{2n},\omega_0)$ be the germ of a hypersurface with isolated singularity at 0. Assume that the ideal of smooth function-germs vanishing on N is generated by a smooth function-germ g on $\mathbb{R}^{2n}$. Let H be a smooth function-germ on $\mathbb{R}^{2n}$ and let X_H be a Hamiltonian vector field-germ on $(\mathbb{R}^{2n},\omega_0)$ with a Hamiltonian H i.e. $X_H\rfloor\omega=-dH$. Let Y be a smooth vector field-germ on $\mathbb{R}^{2n}$. Then the vector field-germ $gY+X_H$ is Hamiltonian on N.*

A smooth k-form β on M vanishes on N if $\beta|_x=0$ for every $x\in N$.

Definition 3. *A smooth k-form α on M has zero algebraic restriction to N if there exist a smooth k-form β on M vanishing on N and a smooth $(k-1)$-form γ on M vanishing on N such that*

$$\alpha=\beta+d\gamma. \tag{6}$$

Let $\mathcal{A}_0^k(N,M)$ denote the space of smooth k-forms with zero algebraic restriction to N. Since $d(\mathcal{A}_0^k(N,M))\subset\mathcal{A}_0^{k+1}(N,M)$, the complex $(\mathcal{A}_0^*(N,M),d)$ is a subcomplex of the de Rham complex on M. We denote by $H^*(N,M)$ the cohomology groups of the complex $(\mathcal{A}_0^*(N,M),d)$.

Proposition 4. *A smooth vector field X on M is Hamiltonian on N if and only if there exists a smooth function H on M such that $X\rfloor\omega+dH$ has zero algebraic restriction to N.*

Proof. Definition 1 is equivalent to the following condition:

$$X\rfloor\omega+dH=\sum_{i=1}^{k}f_i\alpha_i, \tag{7}$$

where $\alpha_1,\cdots,\alpha_k$ are smooth 1-forms on M, $H,f_1,\cdots,f_k$ are smooth functions on M such that $f_1|_N=\cdots=f_k|_N=0$. But this implies that $X\rfloor\omega+dH$ has zero algebraic restriction to N.

On the other hand, if there exists a smooth function H on M such that $X\rfloor\omega+dH$ has zero algebraic restriction to N, then

$$X\rfloor\omega+dH=\sum_{i=1}^{k}f_i\alpha_i+dg, \tag{8}$$

where $\alpha_1,\cdots,\alpha_k$ are smooth 1-forms on M, $H,f_1,\cdots,f_k,g$ are smooth functions on M such that $f_1|_N=\cdots=f_k|_N=g|_N=0$. But this can be written in the following way:

$$X\rfloor\omega+d(H-g)=\sum_{i=1}^{k}f_i\alpha_i, \tag{9}$$

which implies that X is Hamiltonian on N. □

The above definition and proposition are the motivation for the following definition of the symplectic vector field on N:

Definition 5. *A smooth vector field X on M is called* **symplectic on** *N if $\mathcal{L}_X\omega$ has zero algebraic restriction to N.*

It is obvious that a vector field, which is Hamiltonian on N, is symplectic on N. The inverse implication is not always true. The necessary and sufficient conditions are given in the following proposition:

Proposition 6. *The vector field-germ X is Hamiltonian on N if and only if X is symplectic on N and $\mathcal{L}_X\omega$ define the zero cohomology class in $H^2(N,M)$.*

Corollary 7. *If $H^2(N,M)=\{0\}$, then any symplectic vector field-germ on N is Hamiltonian on N.*

Definition 8. *The germ at 0 of a set $N\subset\mathbb{R}^m$ is called quasi-homogeneous if there exist a local coordinate system $x_1,\ldots,x_m$ and positive numbers $\lambda_1,\ldots,\lambda_m$ such that the following holds: if a point with coordinates $x_i=a_i$ belongs to N, then for any $t\in[0,1]$ the point with coordinates $x_i=t^{\lambda_i}a_i$ also belongs to N.*

It was proved that if N is quasi-homogeneous, then $H^k(N,M)=\{0\}$ for $k>0$. (e.g. see [4]). It implies the following proposition:

Proposition 9. *If N is quasi-homogeneous, then any symplectic vector field-germ on N is Hamiltonian on N.*

4. GERMS OF HAMILTONIAN VECTOR FIELDS ON SMOOTH SUBMANIFOLDS

If S is a smooth submanifold of M, then a smooth k-form α on M has zero algebraic restriction to M if and only if the pullback of α to M vanishes. Thus, we obtain the following result:

Corollary 10. *Let S be a smooth submanifold of M. Let $\imath: S\hookrightarrow M$ be an embedding of S. A smooth vector field X on M is Hamiltonian on S if and only if there exists a smooth function H on M such that*

$$\imath^*(X\rfloor\omega)=d(H\circ\imath). \tag{10}$$

Thus, by the above corollary we obtain the following:

$$\omega(X(x),v)=-dH(v),\ \text{for every } x\in S,\ \text{and for every } v\in T_xS. \tag{11}$$

It means that if the vector field X is Hamiltonian on a smooth submanifold S of M, then X is a section of the bundle L.

By Poincare Lemma and Corollary 10 we have

Proposition 11. *Let S be a smooth submanifold of M. Let $\imath: S\hookrightarrow M$ be an embedding of S. A smooth vector field X on M is Hamiltonian on S if and only if $d(\imath^*(X\rfloor\omega))=0$.*

4.1. SYMPLECTIC SUBMANIFOLDS

Let S be the germ of a symplectic submanifold of dimension $2k$ of the symplectic manifold $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{n} dx_i \wedge dy_i)$. Then, by the Darboux-Givental Theorem (see [1]), S is symplectomorphic to

$$S_0 = \{(x,y) \in \mathbb{R}^{2n} | x_i = y_i = 0 \text{ for } i = k+1, \cdots, n\}.$$

If $(\tilde{x}, \tilde{y}) = (x_1, \cdots, x_k, y_1, \cdots, y_k)$ and $\imath : S \ni (\tilde{x}, \tilde{y}) \mapsto (\tilde{x}, 0, \tilde{y}, 0) \in \mathbb{R}^{2n}$, then a smooth vector field-germ

$$X = \sum_{i=1}^{n} f_i(x,y) \frac{\partial}{\partial x_i} + g_i(x,y) \frac{\partial}{\partial y_i}$$

at 0 on $\mathbb{R}^{2n}$ is Hamiltonian on S_0 if $d(\imath^*(X \rfloor \omega) = 0$.

It implies that $d(\sum_{i=1}^{k} f_i(\tilde{x}, 0, \tilde{y}, 0) dy_i - g_i(\tilde{x}, 0, \tilde{y}, 0) dx_i) = 0$. Thus, the vector field-germ

$$\tilde{X} = \sum_{i=1}^{k} f_i(\tilde{x}, 0, \tilde{y}, 0)) \frac{\partial}{\partial x_i} + g_i(\tilde{x}, 0, \tilde{y}, 0)) \frac{\partial}{\partial y_i}$$

on S_0 is Hamiltonian on a symplectic manifold $(S_0, \imath^* \omega = \sum_{i=1}^{k} dx_i \wedge dy_i)$. Let us notice that $\tilde{X}|_{\pi(x,y)} = \pi_*(X|_{(x,y)})$, where $\pi : \mathbb{R}^{2n} \ni (x,y) \mapsto (\tilde{x}, \tilde{y}) \in S_0$. Since $\pi \circ \imath = Id_{S_0}$, we obtain the following proposition:

Proposition 12. *A smooth vector field-germ X on $(\mathbb{R}^{2n}, \omega)$ is Hamiltonian on the symplectic submanifold-germ S_0, if the vector field-germ $\pi_*(X \circ \imath)$ on S_0 is Hamiltonian on the symplectic manifold $(S_0, \imath^* \omega)$.*

4.2. COISOTROPIC SUBMANIFOLDS

Let C be the germ of a coisotropic submanifold of codimension k of the symplectic manifold $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{n} dx_i \wedge dy_i)$. Then, by the Darboux-Givental Theorem (see [1]), C is symplectomorphic to

$$C_0 = \{(x,y) \in \mathbb{R}^{2n} | x_i = 0 \text{ for } i = 1, \cdots, k\}.$$

If $\tilde{x} = (x_{k+1}, \cdots, x_n)$ and $\imath : C_0 \ni (\tilde{x}, y) \mapsto (0, \tilde{x}, y) \in \mathbb{R}^{2n}$, then a smooth vector field-germ

$$X = \sum_{i=1}^{n} f_i(x,y) \frac{\partial}{\partial x_i} + g_i(x,y) \frac{\partial}{\partial y_i}$$

at 0 on $\mathbb{R}^{2n}$ is Hamiltonian on C_0 if $d(\imath^*(X \rfloor \omega) = 0$. It implies that the 1-form-germ

$$\imath^*(X \rfloor \omega) = \sum_{i=1}^{n} f_i(0, \tilde{x}, y) dy_i - \sum_{i=k+n}^{k} g_i(0, \tilde{x}, y) dx_i$$

on C_0 is exact. Hence, there exists a smooth function-germ on C_0 such that $g_i(0,\tilde{x},y) = -\frac{\partial h}{\partial x_i}(\tilde{x},y)$ for $i = k+1,\cdots,n$ and $f_i(0,\tilde{x},y) = \frac{\partial h}{\partial y_i}(\tilde{x},y)$ for $i = 1,\cdots,n$. Thus, we obtain the following proposition:

Proposition 13. *A smooth vector field-germ*

$$X = \sum_{i=1}^{n} f_i(x,y)\frac{\partial}{\partial x_i} + g_i(x,y)\frac{\partial}{\partial y_i}$$

on $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{n} dx_i \wedge dy_i)$ *is Hamiltonian on the coisotropic submanifold-germ*

$$C_0 = \{(x,y) \in \mathbb{R}^{2n} |\ x_i = 0 \text{ for } i = 1,\cdots,k\},$$

if there exists a smooth function germ h *on* C_0 *such that* $g_i(0,\tilde{x},y) = -\frac{\partial h}{\partial x_i}(\tilde{x},y)$ *for* $i = k+1,\cdots,n$ *and* $f_i(0,\tilde{x},y) = \frac{\partial h}{\partial y_i}(\tilde{x},y)$ *for* $i = 1,\cdots,n$.

4.3. ISOTROPIC SUBMANIFOLDS

Let I be the germ of an isotropic submanifold of dimension k of the symplectic manifold $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{n} dx_i \wedge dy_i)$. Then, by the Darboux-Givental Theorem (see [1]), I is symplectomorphic to

$$I_0 = \{(x,y) \in \mathbb{R}^{2n} | y = 0,\ x_i = 0 \text{ for } i = k+1,\cdots,n\}.$$

If $\tilde{x} = (x_1,\cdots,x_k)$ and $\iota : I_0 \ni \tilde{x} \mapsto (\tilde{x},0) \in \mathbb{R}^{2n}$, then a smooth vector field-germ

$$X = \sum_{i=1}^{n} f_i(x,y)\frac{\partial}{\partial x_i} + g_i(x,y)\frac{\partial}{\partial y_i}$$

at 0 on $\mathbb{R}^{2n}$ is Hamiltonian on I_0 if $d(\iota^*(X \rfloor \omega) = 0$. It implies that the 1-form-germ

$$\iota^*(X \rfloor \omega) = -\sum_{i=1}^{k} g_i(\tilde{x},0)dx_i$$

on I_0 is exact. Hence, there exists a smooth function-germ on I_0 such that $g_i(\tilde{x},0) = \frac{\partial h}{\partial x_i}(\tilde{x})$ for $i = 1,\cdots,k$. Thus, we obtain the following proposition:

Proposition 14. *A smooth vector field-germ*

$$X = \sum_{i=1}^{n} f_i(x,y)\frac{\partial}{\partial x_i} + g_i(x,y)\frac{\partial}{\partial y_i}$$

on $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^{n} dx_i \wedge dy_i)$ *is Hamiltonian on the isotropic submanifold*

$$I_0 = \{(x,y) \in \mathbb{R}^{2n} | y = 0,\ x_i = 0 \text{ for } i = k+1,\cdots,n\},$$

if there exists a smooth function-germ h *on* I_0 *such that* $g_i(\tilde{x},0) = \frac{\partial h}{\partial x_i}(\tilde{x})$ *for* $i = 1,\cdots,k$.

In particular by Proposition 11 we obtain

Corollary 15. *If C is a regular curve (1-dimensional smooth submanifold), then any smooth vector field-germ on $\mathbb{R}^{2n}$ is Hamiltonian.*

Proof. Any smooth 1-form on C is closed. □

5. GERMS OF HAMILTONIAN VECTOR FIELDS ON SINGULAR CURVES

In this section we describe germs of Hamiltonian vector fields on singular curves at a singular point. By Corollary 15 any smooth vector field-germ is Hamiltonian on a regular curve.

PLANAR CURVES OF TYPES A_K, D_K, E_6, E_7, E_8

A planar curve in the symplectic space $(\mathbb{R}^{2n}, \omega)$ is a curve which is embedded in a smooth 2-dimensional submanifold S of $(\mathbb{R}^{2n}, \omega)$. Let $\imath : S \hookrightarrow \mathbb{R}^{2n}$ be an embedding of S.

We assume that the germ of the curve is locally diffeomorphic to $N = \{x \in \mathbb{R}^{2n} | G(x_1, x_2) = x_{\geq 3} = 0\}$, where G has the following properties:

1. $G(0,0) = 0, dG(0,0) = 0$,
2. the ideal of smooth function-germs on $\mathbb{R}^2$ vanishing on $\{(x_1, x_2) \in \mathbb{R}^2 | G(x_1, x_2) = 0\}$ is generated by G.
3. G is quasi-homogeneous polynomial.

Then, we can take locally $S = \{x \in \mathbb{R}^{2n} | x_{\geq 3} = 0\}$ and $\imath(x_3, \cdots, x_{2n}) = (0, 0, x_3, \cdots, x_{2n})$. A smooth vector field-germ on $(\mathbb{R}^{2n}, \omega)$ is Hamiltonian on N if and only if $d(\imath^*(X \rfloor \omega)) = d(G(x_1, x_2)\alpha)$ for some smooth 1-form-germ α on $\mathbb{R}^2$.

By Theorem 4.11 in [5] any curve-germ in the symplectic space $(\mathbb{R}^{2n}, \omega_0 = \sum_{i=0}^{n} dp_i \wedge dq_i)$, $n \geq 2$, which is diffeomorphic to the curve-germ at 0 $\{x \in \mathbb{R}^{2n} | G(x_1, x_2) = x_{\geq 3} = 0\}$ for smooth function-germs G in Tab. 1 is symplectomorphic to one and only one of the following curve-germs:

$$N^i = \{(p,q) \in \mathbb{R}^{2n} | G(p_1, p_2) = q_1 - \int_0^{p_2} F_i(p_1, t)dt = q_{\geq 2} = p_{\geq 3} = 0\} \subset (\mathbb{R}^{2n}, \omega_0), \quad (12)$$

for $i = 0, \cdots, \mu$, where smooth function-germs F_i are presented in Tab. 1.

Table 1

Classification of the algebraic restrictions to A_k, D_k, E_6, E_7, E_8

$G(x_1,x_2)$	$F_i(x_1,x_2),\ i=0,1,\ldots,\mu$
$A_k: x_1^{k+1}-x_2^2$ $k\geq 1$	$F_0=1$ $F_i=x_1^i,\ i=1,\ldots,k-1$ $F_k=0$
$D_k: x_1^2x_2-x_2^{k-1}$ $k\geq 4$	$F_0=1$ $F_i=bx_1+x_2^i,\ \ i=1,\ldots,k-4$ $F_{k-3}=(\pm 1)^k x_1+bx_2^{k-3},$ $F_{k-2}=x_2^{k-3},\ F_{k-1}=x_2^{k-2},\ F_k=0$
$E_6: x_1^3-x_2^4$	$F_0=1,\ F_1=\pm x_2+bx_1,\ F_2=x_1+bx_2^2,$ $F_3=x_2^2+bx_1x_2,\ F_4=\pm x_1x_2,\ F_5=x_1x_2^2,\ F_6=0$
$E_7: x_1^3-x_1x_2^3$	$F_0=1,\ F_1=x_2+bx_1,\ F_2=\pm x_1+bx_2^2,$ $F_3=x_2^2+bx_1x_2,\ F_4=\pm x_1x_2+bx_2^3,$ $F_5=x_2^3,\ F_6=x_2^4,\ F_7=0$
$E_8: x_1^3-x_2^5$	$F_0=\pm 1,\ F_1=x_2+bx_1,\ F_2=x_1+b_1x_2^2+b_2x_2^3$ $F_3=\pm x_2^2+bx_1x_2,\ F_4=\pm x_1x_2+bx_2^3,$ $F_5=x_2^3+bx_1x_2^2,\ F_6=x_1x_2^2,\ F_7=\pm x_1x_2^3,\ F_8=0$

Let $\iota:\mathbb{R}^2\to\mathbb{R}^{2n}$ be the following map-germ: $\iota(p_1,p_2)=(p_1,\int_0^{p_2}F_i(p_1,t)dt,p_2,0)$. A smooth vector field-germ

$$X=\sum_{i=1}^{n} f_i(p_1,q_1,\cdots,p_n,q_n)\frac{\partial}{\partial p_i}+g_i(p_1,q_1,\cdots,p_n,q_n)\frac{\partial}{\partial q_i}$$

on $(\mathbb{R}^{2n},\omega_0=\sum_{i=1}^n dp_i\wedge dq_i)$ is Hamiltonian on N^i if a smooth 2-form-germ at 0 on $\mathbb{R}^2$

$$\sigma=r(p_1,p_2)dp_1\wedge dp_2=d\left((f_1\circ\iota)d(\int_0^{p_2}F_i(p_1,t)dt)-(g_1\circ\iota)dp_1-(g_2\circ\iota)dp_2\right)$$

has zero algebraic restriction to $\{(p_1,p_2)\in\mathbb{R}^2|G(p_1,p_2)=0\}$.

By the direct calculation we obtain that

$$\begin{aligned} r(p_1,p_2)&=\left(\tfrac{\partial g_1}{\partial p_2}-\tfrac{\partial g_2}{\partial p_1}\right)(p_1,\textstyle\int_0^{p_2}F_i(p_1,t)dt,p_2,0)\\ &+F_i(p_1,p_2)\left(\tfrac{\partial f_1}{\partial p_1}+\tfrac{\partial g_1}{\partial q_1}\right)(p_1,\textstyle\int_0^{p_2}F_i(p_1,t)dt,p_2,0)\\ &-\textstyle\int_0^{p_2}\tfrac{\partial F_i}{\partial p_1}(p_1,t)dt\left(\tfrac{\partial f_1}{\partial p_2}+\tfrac{\partial g_2}{\partial q_1}\right)(p_1,\textstyle\int_0^{p_2}F_i(p_1,t)dt,p_2,0). \end{aligned}\tag{13}$$

If G is quasi-homogeneous, then a smooth 2-form $r(p_1,p_2)dp_1\wedge dp_2$ has zero algebraic restriction to $\{(p_1,p_2)\in\mathbb{R}^2|G(p_1,p_2)=0\}$ if and only if r belongs to the ideal $<\nabla G>$ generated by $\frac{\partial G}{\partial p_1}(p_1,p_2)$, $\frac{\partial G}{\partial p_2}(p_1,p_2)$ (see [5]).

Thus, we obtain the following proposition:

Proposition 16. *A smooth vector field-germ*

$$X=\sum_{j=1}^{n} f_j(p,q)\frac{\partial}{\partial p_j}+g_j(p,q)\frac{\partial}{\partial q_j}$$

is Hamiltonian on

$$N^i=\{(p,q)\in\mathbb{R}^{2n}|G(p_1,p_2)=q_1-\int_0^{p_2}F_i(p_1,t)dt=q_{\geq 2}=p_{\geq 3}=0\}\subset(\mathbb{R}^{2n},\omega_0),$$

where G and F_i are presented in Tab. 1, if and only if the function-germ r given by (13) belongs to the ideal $<\nabla G>$.

5.1. PLANAR CURVES OF TYPES A_K^I

By Proposition 16 we obtain the following:

Proposition 17. *Let us fix $k\in\mathbb{N}$ and $i=0,1,\cdots,k$. A smooth vector field-germ*

$$X=\sum_{j=1}^{n} f_j(p,q)\frac{\partial}{\partial p_j}+g_j(p,q)\frac{\partial}{\partial q_j}$$

on $(\mathbb{R}^{2n},\omega_0=\sum_{j=1}^n dp_j\wedge dq_j)$ is Hamiltonian on

$$A_k^i=\{(p,q)\in\mathbb{R}^{2n}|p_1^{k+1}-p_2^2=q_1-p_1^ip_2=q_{\geq 2}=p_{\geq 3}=0\}\ (i=0,1,\cdots,k-1)$$

or on

$$A_k^k=\{(p,q)\in\mathbb{R}^{2n}|p_1^{k+1}-p_2^2=q_{\geq 1}=p_{\geq 3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)=\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0)\ \textit{for}\ j=0,\cdots,i-1,$$

$$\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)=\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0)-\frac{j!}{(j-i)!}\left(\frac{\partial^{j-i+1}f_1}{\partial p_1^{j-i+1}}(0)+\frac{\partial^{j-i+1}g_2}{\partial p_1^{j-i}\partial q_1}(0)\right)\ \textit{for}\ j=i,\cdots,k-1.$$

Proof. For a planar curve of type A_k^i $(k\geq 1, i=0,\cdots,k)$, we have that $G(p_1,p_2)=p_1^{k+1}-p_2^2$ and $F_i(p_1,p_2)=p_1^i$ for $i=0,\cdots,k-1$ and $F_k(p_1,p_2)=0$ (see Tab. 1).

For A_k^0 singularity we have

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}+\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,p_2,p_2,0).$$

For A_k^i singularity $i=1,\cdots,k-1$ the function-germ r has the following form:

$$\begin{aligned} r(p_1,p_2) &= \left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,p_1^i p_2,p_2,0) \\ &+p_1^i\left(\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,p_1^i p_2,p_2,0) \\ &-ip_1^{i-1}p_2\left(\frac{\partial f_1}{\partial p_2}+\frac{\partial g_2}{\partial q_1}\right)(p_1,p_1^i p_2,p_2,0). \end{aligned} \tag{14}$$

For A_k^k singularity we get $F_k(p_1,p_2)=0$ and

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,0,p_2,0).$$

Since $<\nabla G>=<p_1^k,p_2>$, it is easy to see that $\mathcal{O}_2/<\nabla G>\cong\mathbb{R}\left\{1,p_1,\cdots,p_1^{k-1}\right\}$. The function-germ r belongs to $<\nabla G>$ if and only if $\frac{\partial^j r}{\partial p_1^j}(0,0)=0$ for $j=0,1,\cdots,k-1$. By a direct calculation we get that for $j=0,\cdots,i-1$

$$\frac{\partial^j r}{\partial p_1^j}(0,0)=\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)-\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0),$$

and for $j=i,\cdots,k-1$

$$\frac{\partial^j r}{\partial p_1^j}(0,0)=\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)-\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0)+\frac{j!}{(j-i)!}\left(\frac{\partial^{j-i+1}f_1}{\partial p_1^{j-i+1}}(0)+\frac{\partial^{j-i+1}g_2}{\partial p_1^{j-i}\partial q_1}(0)\right).$$

□

5.2. PLANAR CURVES OF TYPES D_K^I

For a planar curve of type D_k^i ($k\geq 4$, $i=0,\cdots,k$) we have that $G(p_1,p_2)=p_1^2p_2-p_2^{k-1}$. Then it is easy to see that $<\nabla G>=<p_1p_2,p_1^2-(k-1)p_2^{k-2}>$ and

$$\mathcal{O}_2/<\nabla G>\cong\mathbb{R}\left\{1,p_2,\cdots,p_2^{k-2},p_1\right\}.$$

For D_k^0 singularity we get $F_0(p_1,p_2)=1$ and

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}+\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,p_2,p_2,0).$$

For D_k^i singularity $i=1,\cdots,k-4$ we get $F_i(p_1,p_2)=bp_1+p_2^i$ and

$$\begin{aligned} r(p_1,p_2) &= \left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,bp_1p_2+\tfrac{p_2^{i+1}}{i+1},p_2,0) \\ &+(bp_1+p_2^i)\left(\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,bp_1p_2+\tfrac{p_2^{i+1}}{i+1},p_2,0) \\ &-bp_2\left(\frac{\partial f_1}{\partial p_2}+\frac{\partial g_2}{\partial q_1}\right)(p_1,bp_1p_2+\tfrac{p_2^{i+1}}{i+1},p_2,0). \end{aligned} \tag{15}$$

For D_k^{k-3} singularity we have $F_{k-3}(p_1,p_2)=(\pm1)^k p_1+bp_2^{k-3}$ and

$$\begin{aligned} r(p_1,p_2)=\left(\tfrac{\partial g_1}{\partial p_2}-\tfrac{\partial g_2}{\partial p_1}\right)(p_1,(\pm1)^k p_1p_2+b\tfrac{p_2^{k-2}}{k-2},p_2,0) \\ +((\pm1)^k p_1+bp_2^{k-3})\left(\tfrac{\partial f_1}{\partial p_1}+\tfrac{\partial g_1}{\partial q_1}\right)(p_1,(\pm1)^k p_1p_2+b\tfrac{p_2^{k-2}}{k-2},p_2,0) \\ -(\pm1)^k p_2\left(\tfrac{\partial f_1}{\partial p_2}+\tfrac{\partial g_2}{\partial q_1}\right)(p_1,(\pm1)^k p_1p_2+b\tfrac{p_2^{k-2}}{k-2},p_2,0). \end{aligned} \tag{16}$$

For D_k^{k-2} singularity we get $F_{k-2}(p_1,p_2)=p_2^{k-3}$ and

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,\frac{p_2^{k-2}}{k-2},p_2,0)+p_2^{k-3}\left(\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,\frac{p_2^{k-2}}{k-2},p_2,0).$$

For D_k^{k-1} singularity we get $F_{k-2}(p_1,p_2)=p_2^{k-2}$ and

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,\frac{p_2^{k-1}}{k-1},p_2,0)+p_2^{k-2}\left(\frac{\partial f_1}{\partial p_1}+\frac{\partial g_1}{\partial q_1}\right)(p_1,\frac{p_2^{k-1}}{k-1},p_2,0).$$

For D_k^k singularity we get $F_k(p_1,p_2)=0$ and

$$r(p_1,p_2)=\left(\frac{\partial g_1}{\partial p_2}-\frac{\partial g_2}{\partial p_1}\right)(p_1,0,p_2,0).$$

The function-germ r belongs to $<\nabla G>=<p_1p_2,p_1^2-(k-1)p_2^{k-2}>$ if and only if

$$\frac{\partial^j r}{\partial p_2^j}(0,0)=\frac{\partial r}{\partial p_1}(0,0)=0 \text{ for } j=0,1,\cdots,k-3 \tag{17}$$

and

$$\frac{\partial^2 r}{\partial p_1^2}(0,0)=\frac{2}{(k-1)!}\frac{\partial^{k-2} r}{\partial p_2^{k-2}}(0,0). \tag{18}$$

For general k conditions (17)-(18) are rather complicated in terms of partial derivatives of coefficient functions f_1, g_1, g_2 at 0. Therefore, we will present them only for D_4^i singularities for $i=0,1,\cdots,4$.

Let $X=\sum_{j=1}^n f_j(p,q)\frac{\partial}{\partial p_j}+g_j(p,q)\frac{\partial}{\partial q_j}$ be the smooth vector field-germ on $(\mathbb{R}^{2n},\omega_0=\sum_{j=1}^n dp_j\wedge dq_j)$. By a direct calculation Proposition 16 implies the following:

The vector field-germ X is Hamiltonian on

$$D_4^0=\{(p,q)\in\mathbb{R}^{2n}|p_1^2p_2-p_2^3=q_1-p_2=q_{\geq2}=p_{\geq3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0)=\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)+\frac{\partial^{j+1}g_1}{\partial p_1^j\partial q_1}(0)+\frac{\partial^{j+1}f_1}{\partial p_1^{j+1}}(0)\ \text{ for } j=0,1,$$

$$\frac{\partial^2 g_1}{\partial p_2^2}(0)+2\frac{\partial^2 g_1}{\partial q_1\partial p_2}(0)+\frac{\partial^2 g_1}{\partial q_1^2}(0)+\frac{\partial^2 f_1}{\partial p_1\partial p_2}(0)$$
$$-\frac{\partial^2 g_2}{\partial p_1\partial p_2}(0)+\frac{\partial^2 f_1}{\partial p_1\partial q_1}(0)-\frac{\partial^2 g_2}{\partial p_1\partial q_1}(0)=0,$$

$$3\left(\frac{\partial^3 g_1}{\partial p_1^2\partial p_2}(0)+\frac{\partial^3 g_1}{\partial p_1^2\partial q_1}(0)+\frac{\partial^3 f_1}{\partial p_1^3}(0)-\frac{\partial^3 g_2}{\partial p_1^3}(0)\right)$$
$$=\frac{\partial^3 g_1}{\partial p_2^3}(0)+3\frac{\partial^3 g_1}{\partial q_1\partial p_2^2}(0)+3\frac{\partial^3 g_1}{\partial q_1^2\partial p_2}(0)+\frac{\partial^3 g_1}{\partial q_1^3}(0)+\frac{\partial^3 f_1}{\partial p_1\partial p_2^2}(0)-\frac{\partial^3 g_2}{\partial p_1\partial p_2^2}(0)$$
$$+2\frac{\partial^3 f_1}{\partial p_1\partial q_1\partial p_2}(0)-2\frac{\partial^3 g_2}{\partial p_1\partial q_1\partial p_2}(0)+\frac{\partial^3 f_1}{\partial p_1\partial q_1^2}(0)-\frac{\partial^3 g_2}{\partial p_1\partial q_1^2}(0).$$

The vector field-germ X is Hamiltonian on

$$D_4^1=\{(p,q)\in\mathbb{R}^{2n}|p_1^2p_2-p_2^3=q_1-p_1p_2-b\frac{p_2^2}{2}=q_{\geq 2}=p_{\geq 3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial g_1}{\partial p_2}(0)=\frac{\partial g_2}{\partial p_1}(0),$$

$$\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)+\frac{\partial^2 g_1}{\partial p_1\partial p_2}(0)-\frac{\partial^2 g_2}{\partial p_1^2}(0)=0,$$

$$-\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial^2 g_1}{\partial p_2^2}(0)-\frac{\partial g_2}{\partial q_1}(0)+b\left(\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)\right)-\frac{\partial^2 g_2}{\partial p_1\partial p_2}(0)=0,$$

$$6\left(\frac{\partial^2 g_1}{\partial p_1\partial q_1}(0)+\frac{\partial^2 f_1}{\partial p_1^2}(0)\right)+3\left(\frac{\partial^3 g_1}{\partial p_1^2\partial p_2}(0)-\frac{\partial^3 g_2}{\partial p_1^3}(0)\right)$$
$$=\frac{\partial^3 g_1}{\partial p_2^3}(0)-\frac{\partial^3 g_2}{\partial p_1\partial p_2^2}(0)-2\left(\frac{\partial^2 f_1}{\partial p_2^2}(0)+\frac{\partial^2 g_2}{\partial q_1\partial p_2}(0)\right)$$
$$+b\left(3\frac{\partial^2 g_1}{\partial p_1\partial p_2}(0)+2\frac{\partial^2 f_1}{\partial p_1\partial p_2}(0)-\frac{\partial g_2}{\partial p_1}\partial q_1(0)\right).$$

The vector field-germ X is Hamiltonian on

$$D_4^2=\{(p,q)\in\mathbb{R}^{2n}|p_1^2p_2-p_2^3=q_1-\frac{p_2^2}{2}=q_{\geq 2}=p_{\geq 3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)=\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0) \text{ for } j=0,1,$$

$$\frac{\partial^2 g_1}{\partial p_2^2}(0)+\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)-\frac{\partial^2 g_2}{\partial p_1\partial p_2}(0)=0,$$

$$3\left(\frac{\partial^3 g_1}{\partial p_1^2\partial p_2}(0)-\frac{\partial^3 g_2}{\partial p_1^3}(0)\right)=$$

$$\frac{\partial^3 g_1}{\partial p_2^3}(0)+\frac{\partial^2 g_1}{\partial q_1\partial p_2}(0)+2\left(\frac{\partial^2 g_1}{\partial q_1\partial p_2}(0)+\frac{\partial^2 f_1}{\partial p_1\partial p_2}(0)\right)-\frac{\partial^3 g_2}{\partial p_1\partial p_2^2}(0)-\frac{\partial^2 g_2}{\partial p_1\partial q_1}(0).$$

The vector field-germ X is Hamiltonian on

$$D_4^3=\{(p,q)\in\mathbb{R}^{2n}|p_1^2p_2-p_2^3=q_1-\frac{p_2^3}{3}=q_{\geq 2}=p_{\geq 3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)=\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0) \text{ for } j=0,1,$$

$$\frac{\partial^2 g_1}{\partial p_2^2}(0)=\frac{\partial^2 g_2}{\partial p_1\partial p_2}(0),$$

$$3\left(\frac{\partial^3 g_1}{\partial p_1^2\partial p_2}(0)-\frac{\partial^3 g_2}{\partial p_1^3}(0)\right)=\frac{\partial^3 g_1}{\partial p_2^3}(0)+2\left(\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)\right)-\frac{\partial^3 g_2}{\partial p_1\partial p_2^2}(0).$$

The vector field-germ X is Hamiltonian on

$$D_4^4=\{(p,q)\in\mathbb{R}^{2n}|p_1^2p_2-p_2^3=q_{\geq 1}=p_{\geq 3}=0\}$$

if and only if the following conditions are satisfied:

$$\frac{\partial^{j+1}g_1}{\partial p_1^j\partial p_2}(0)=\frac{\partial^{j+1}g_2}{\partial p_1^{j+1}}(0) \text{ for } j=0,1,$$

$$\frac{\partial^2 g_1}{\partial p_2^2}(0) = \frac{\partial^2 g_2}{\partial p_1 \partial p_2}(0),$$

$$3\left(\frac{\partial^3 g_1}{\partial p_1^2 \partial p_2}(0) - \frac{\partial^3 g_2}{\partial p_1^3}(0)\right) = \frac{\partial^3 g_1}{\partial p_2^3}(0) - \frac{\partial^3 g_2}{\partial p_1 \partial p_2^2}(0).$$

In the same way by Proposition 16 one can obtain the necessary and sufficient conditions for the vector field-germ X to be Hamiltonian on planar curves with E_k^i singularities for $k = 6,7,8$ and $i = 0,1,\cdots,k$ (see Tab. 1). Please notice that for E_k^i singularity there are k independent conditions, therefore we do not present them.

6. GERMS OF HAMILTONIAN VECTOR FIELDS ON REGULAR UNION SINGULARITIES.

A regular union singularity N at 0 in $\mathbb{R}^{2n}$ is the union

$$N = N_1 \cup \cdots \cup N_s,\ s \geq 2 \tag{19}$$

of germs at 0 of smooth submanifolds $N_1,\cdots,N_s$ of $\mathbb{R}^{2n}$ (in what follows - strata) such that the dimension of the space

$$W = T_0N_1 + \cdots + T_0N_s \tag{20}$$

is equal to the sum of the dimensions of the strata, i.e. the sum (20) is direct. If the number of strata and their dimensions are fixed, then all such N are diffeomorphic. By Theorem 7.1 in [5] the germ of a closed 2-form σ has zero algebraic restriction to N if and only if its pullback to each of the strata N_i $(i = 1,\cdots,s)$ vanishes and the restriction of the germ σ to the space W vanishes. It implies the following:

Proposition 18. *A smooth vector field-germ X in the symplectic space $(\mathbb{R}^{2n},\omega)$ is Hamiltonian on a regular union singularity N if and only if the pullback of the germ $d(X\rfloor\omega)$ to each of the strata N_i $(i = 1,\cdots,s)$ vanishes and the restriction of the germ $d(X\rfloor\omega)$ to the space W vanishes.*

6.1. REGULAR UNION OF THREE 1-DIMENSIONAL SUBMANIFOLDS

Let us consider a regular union singularity of three germs at 0 of 1-dimensional submanifolds $N = N_1 \cup N_2 \cup N_3$ of the symplectic space $(\mathbb{R}^{2n}, \omega = \sum_{i=1}^n dp_i \wedge dq_i)$. These symplectic singularities are classified in [5].

Proposition 19 (Theorem 7.4 in [5])**.** *Any regular union singularity N with three 1-dimensional strata in the symplectic space $(\mathbb{R}^{2n},\omega)$, $n\geq 3$ (resp. $n=2$) is symplectomorphic to one and only one of the varieties N^0,N^1,N^2,N^3 (resp. N^0,N^1,N^2) given in Tab. 2. It holds if and only if the pair (ω,N) satisfies the condition in the last column of the table.*

Table 2

Classification of symplectic regular union singularities with three 1-dimensional strata. W denotes the 3-space spanned by the tangent lines at 0 to the strata

	Symplectic normal forms	Geometric condition
N^0	$q_2=p_1+p_2$, $p_1q_1=q_1p_2=p_2q_2=0$, $p_{\geq 3}=q_{\geq 3}=0$	$\omega\|_W\neq 0$, $\ker\omega\|_W\not\subset T_0N_i+T_0N_j$, for any $i,j\in\{1,2,3\}$;
N^1	$q_2=p_1$, $p_1q_1=q_1p_2=p_2p_1=0$, $p_{\geq 3}=q_{\geq 3}=0$	$\omega\|_W\neq 0$, $\ker\omega\|_W\subset T_0N_i+T_0N_j$, $\ker\omega\|_W\neq T_0N_i,T_0N_j$ for some $i,j\in\{1,2,3\}$;
N^2	$p_1q_1=q_1p_2=p_2p_1=0$, $p_{\geq 3}=$ $q_{\geq 2}=0$	$\omega\|_W\neq 0$, $\ker\omega\|_W=T_0N_i$ for some $i\in\{1,2,3\}$
N^3	$p_1p_2=p_2p_3=p_3p_1=0$, $p_{\geq 4}=$ $q_{\geq 1}=0$	$\omega\|_W=0$.

Since the strata are 1-dimensional, by Proposition 18, a smooth vector-field germ X is Hamiltonian on N if and only if $d(X\rfloor\omega)|_W=0$. Hence for singularities N^i for $i=0,1,\cdots,3$ we obtain the following conditions:

Let $X=\sum_{i=1}^n f_i(p,q)\frac{\partial}{\partial p_i}+g_i(p,q)\frac{\partial}{\partial q_i}$ be a smooth vector field-germ on $\mathbb{R}^{2n}$.
The vector field-germ X is Hamiltonian on N^0 if and only if

$$\frac{\partial f_1}{\partial q_2}(0)+\frac{\partial f_1}{\partial p_2}(0)-\frac{\partial f_2}{\partial q_1}(0)+\frac{\partial g_2}{\partial q_1}(0)=0,$$

$$\frac{\partial f_2}{\partial q_1}(0)+\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)+\frac{\partial f_2}{\partial q_1}(0)=0,$$

$$\frac{\partial g_1}{\partial q_2}(0)-\frac{\partial g_2}{\partial q_2}(0)-\frac{\partial f_2}{\partial p_2}(0)+\frac{\partial g_1}{\partial p_2}(0)+\frac{\partial f_2}{\partial p_1}(0)-\frac{\partial g_2}{\partial p_1}(0)=0.$$

The vector field-germ X is Hamiltonian on N^1 if and only if

$$\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial g_2}{\partial q_1}(0)=\frac{\partial f_1}{\partial q_2}(0)+\frac{\partial f_2}{\partial q_1}(0)+\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)=0,$$

$$\frac{\partial g_1}{\partial p_2}(0)-\frac{\partial g_2}{\partial q_2}(0)-\frac{\partial f_2}{\partial p_2}(0)-\frac{\partial g_2}{\partial p_1}(0)=0.$$

The vector field-germ X is Hamiltonian on N^2 if and only if

$$\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial g_2}{\partial q_1}(0)=\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)=\frac{\partial g_1}{\partial p_2}(0)-\frac{\partial g_2}{\partial p_1}(0)=0.$$

The vector field-germ X is Hamiltonian on N^3 if and only if

$$\frac{\partial g_2}{\partial p_3}(0)-\frac{\partial g_3}{\partial p_2}(0)=\frac{\partial g_1}{\partial p_2}(0)-\frac{\partial g_2}{\partial p_1}(0)=\frac{\partial g_1}{\partial p_3}(0)-\frac{\partial g_3}{\partial p_1}(0)=0.$$

6.2. REGULAR UNION OF TWO 2-DIMENSIONAL ISOTROPIC SUBMANIFOLDS

Now we consider the regular union singularity of two 2-dimensional isotropic submanifold-germs of the symplectic space. The following classification proposition was proved in [5]:

Proposition 20. *Any regular union singularity N of two 2-dimensional isotropic submanifold-germs in a symplectic space $(\mathbb{R}^{2n},\omega=\sum_{i=1}^{n}dp_i\wedge dq_i)$ is symplectomorphic to one and only one of the varieties N^0,N^1,N^4 in Tab. 3. The orbit of N^i has codimension i in the class of all regular union singularities with two 2-dimensional isotropic strata. The normal form N^i holds if and only if the pair (ω,N) satisfies the condition given in the last column of Tab. 3.*

Table 3

Classification of symplectic regular union singularities of two 2-dimensional isotropic submanifold-germs. W denotes the 4-space spanned by the tangent planes at 0 to the strata

	Symplectic normal forms	Geometric condition	codim
N^0	$\{p_{\geq 3}=q_{\geq 1}=0\}\cup$ $\{p_{\geq 1}=q_{\geq 3}=0\}$	rank $\omega\|_W=4$	0
N^1 ($n\geq 3$)	$\{p_{\geq 3}=q_{\geq 1}=0\}\cup$ $\{p_{\geq 1}=q_2=q_{\geq 4}=0\}$	rank $\omega\|_W=2$	1
N^4 ($n\geq 4$)	$\{p_{\geq 3}=q_{\geq 1}=0\}\cup$ $\{p_1=p_2=p_{\geq 5}=q_{\geq 1}=0\}$	$\omega\|_W=0$	4

By Proposition 18 a smooth vector field-germ X is Hamiltonian on N if and only if X is Hamiltonian on both of isotropic submanifold-germs N_1, N_2 and $d(X\rfloor\omega)|_W=0$.

Let $X=\sum_{i=1}^{n}f_i(p,q)\frac{\partial}{\partial p_i}+g_i(p,q)\frac{\partial}{\partial q_i}$ be a smooth vector field-germ on $\mathbb{R}^{2n}$. By Propositions 18 and 14 we obtain the following conditions:

The vector field-germ X is Hamiltonian on $N^0=N^0_1\cup N^0_2$ if and only if there exist a smooth function-germs h on $N^0_1=\{p_{\geq 3}=q_{\geq 1}=0\}$ and k on $N^0_2=\{q_{\geq 3}=p_{\geq 1}=0\}$ such that $g_i(p_1,p_2,0)=\frac{\partial h}{\partial p_i}(p_1,p_2)$ and $f_i(0,q_1,q_2,0)=\frac{\partial k}{\partial q_i}(q_1,q_2)$ for $i=1,2$, and

$$\frac{\partial g_2}{\partial q_2}(0)+\frac{\partial f_2}{\partial p_2}(0)=\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial g_2}{\partial q_1}(0)=\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)=\frac{\partial g_1}{\partial q_2}(0)+\frac{\partial f_2}{\partial p_1}(0)=0.$$

The vector field-germ X is Hamiltonian on $N^1=N^1_1\cup N^1_2$ if and only if there exist a smooth function-germs h on $N^1_1=\{p_{\geq 3}=q_{\geq 1}=0\}$ and k on $N^1_2=\{p_{\geq 1}=q_2=q_{\geq 4}=0\}$ such that $g_i(p_1,p_2,0)=\frac{\partial h}{\partial p_i}(p_1,p_2)$ for $i=1,2$ and $f_j(0,q_1,0,q_3,0)=\frac{\partial k}{\partial q_j}(q_1,q_3)$ for $j=1,3$, and

$$\frac{\partial g_2}{\partial q_3}(0)+\frac{\partial f_3}{\partial p_2}(0)=\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial g_2}{\partial q_1}(0)=\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)=\frac{\partial g_1}{\partial q_3}(0)+\frac{\partial f_3}{\partial p_1}(0)=0.$$

The vector field-germ X is Hamiltonian on $N^4=N^4_1\cup N^4_2$ if and only if there exist a smooth function-germs h on $N^4_1=\{p_{\geq 3}=q_{\geq 1}=0\}$ and k on $N^4_2=\{p_1=p_2=p_{\geq 5}=q_{\geq 1}=0\}$ such that $g_i(p_1,p_2,0)=\frac{\partial h}{\partial p_i}(p_1,p_2)$ for $i=1,2$ and $g_j(0,p_3,p_4,0)=\frac{\partial k}{\partial p_j}(p_3,p_4)$ for $j=3,4$, and

$$\frac{\partial g_2}{\partial p_3}(0)-\frac{\partial g_3}{\partial p_2}(0)=\frac{\partial g_2}{\partial p_4}(0)+\frac{\partial g_4}{\partial p_2}(0)=\frac{\partial g_1}{\partial p_3}(0)-\frac{\partial g_3}{\partial p_1}(0)=\frac{\partial g_1}{\partial p_4}(0)-\frac{\partial g_4}{\partial p_1}(0)=0.$$

6.3. REGULAR UNION OF TWO 2-DIMENSIONAL SYMPLECTIC SUBMANIFOLDS

In this subsection we consider Hamiltonian vector field-germs on regular union singularities with two 2-dimensional *symplectic* strata in a symplectic space $(\mathbb{R}^{2n},\omega)$. Recall that two germs of submanifolds N_1,N_2 of a symplectic space $(\mathbb{R}^{2n},\omega)$ are called ω-orthogonal if $\omega(v,u)=0$ for any vectors $v\in T_0N_1, u\in T_0N_2$. The symplectic classification of such N involves the following invariant:

Definition 21 (see Definition 7.6 in [5])**.** *The index of non-orthogonality between* 2*-dimensional symplectic submanifolds N_1 and N_2 of a symplectic space $(\mathbb{R}^{2n},\omega)$ is the number*

$$\alpha=\alpha(N_1,N_2)=1-\frac{(\omega\wedge\omega)(v_1,v_2,u_1,u_2)}{2\cdot\omega(v_1,v_2)\cdot\omega(u_1,u_2)}$$

where v_1,v_2 is a basis of T_0N_1 and u_1,u_2 is a basis of T_0N_2.

It is easy to see that the index of non-orthogonality $\alpha(N_1,N_2)$ is well-defined, i.e. it does not depend on the choice of the bases of T_0N_1 and T_0N_2. It is equal to 0 if and only if there exists a non-zero vector $u\in T_0N_1$ such that $\omega(v,u)=0$ for any $v\in T_0N_2$. It is equal to 1 if and only if the 4-form $\omega\wedge\omega$ has zero restriction to the space $W=T_0N_1+T_0N_2$.

Theorem 22 (Theorem 7.9 in [5]). *Let $\omega = \sum_{i=1}^{n} dp_i \wedge dq_i$. Let $N = N_1 \cup N_2$ be the regular union singularity with two 2-dimensional symplectic strata in the symplectic space $(\mathbb{R}^{2n}, \omega)$.*

If N_1 and N_2 are not ω-orthogonal, then N is symplectomorphic to the variety

$$N^{\alpha} = \{q_1 = p_2, p_1 = p_{\geq 3} = q_{\geq 3} = 0\} \cup \{p_2 = \alpha q_1, p_{\geq 3} = q_{\geq 2} = 0\},$$

where α is the index of non-orthogonality between N_1 and N_2.

If N_1 and N_2 are ω-orthogonal, then N has is symplectomorphic to

$$N^{\perp} = \{p_{\geq 2} = q_{\geq 2} = 0\} \cup \{p_1 = q_1 = p_{\geq 3} = q_{\geq 3} = 0\}.$$

If $n \geq 3$, then any of the normal forms is realizable and if $n = 2$, then any of the normal forms is realizable except the normal form N^1.

Theorem 22 was generalized in [6] to regular union singularities of two germs of symplectic or quasi-symplectic k-dimensional submanifolds of the symplectic space. For simplicity we present the case $k = 2$ only.

By Proposition 18 a smooth vector field-germ X is Hamiltonian on $N = N_1 \cup N_2$ if and only if X is Hamiltonian on both of symplectic submanifold-germs N_1, N_2 and $d(X \rfloor \omega)|_W = 0$.

Let $X = \sum_{i=1}^{n} f_i(p,q)\frac{\partial}{\partial p_i} + g_i(p,q)\frac{\partial}{\partial q_i}$ be a smooth vector field-germ on $\mathbb{R}^{2n}$. By Propositions 18 and direct calculations we obtain the following proposition:

Proposition 23. *The vector field-germ X is Hamiltonian on*

$$N^{\alpha} = \{q_1 = p_2, p_1 = p_{\geq 3} = q_{\geq 3} = 0\} \cup \{p_2 = \alpha q_1, p_{\geq 3} = q_{\geq 2} = 0\}$$

if and only if

$$\left(-\frac{\partial f_1}{\partial q_2} + \frac{\partial g_2}{\partial q_2} + \frac{\partial f_2}{\partial p_2} + \frac{\partial f_2}{\partial q_1}\right)|_{\{q_1 = p_2, p_1 = p_{\geq 3} = q_{\geq 3} = 0\}} = 0,$$

$$\left(\alpha\frac{\partial g_1}{\partial p_2} + \frac{\partial g_1}{\partial q_1} + \frac{\partial f_1}{\partial p_1} - \alpha\frac{\partial g_2}{\partial p_1}\right)|_{\{p_2 = \alpha q_1, p_{\geq 3} = q_{\geq 2} = 0\}} = 0,$$

$$\frac{\partial g_2}{\partial q_2}(0) + \frac{\partial f_2}{\partial p_2}(0) = \frac{\partial f_1}{\partial p_2}(0) + \frac{\partial g_2}{\partial q_1}(0) = \frac{\partial g_1}{\partial q_1}(0) + \frac{\partial f_1}{\partial p_1}(0) = \frac{\partial g_1}{\partial q_2}(0) + \frac{\partial f_2}{\partial p_1}(0) = 0,$$

$$\frac{\partial f_1}{\partial q_2}(0) - \frac{\partial f_2}{\partial q_1}(0) = \frac{\partial g_1}{\partial p_2}(0) - \frac{\partial g_2}{\partial p_1}(0) = 0.$$

Let us denote the stata of $N^{\perp}$ by

$$N_1^{\perp} = \{p_{\geq 2} = q_{\geq 2} = 0\}, \ N_2^{\perp} = \{p_1 = q_1 = p_{\geq 3} = q_{\geq 3} = 0\}.$$

In the same way we get the following result:

Proposition 24. *The vector field-germ X is Hamiltonian on $N^{\perp}=N_1^{\perp}\cup N_2^{\perp}$ if and only if*

$$\left(\frac{\partial g_1}{\partial q_1}+\frac{\partial f_1}{\partial p_1}\right)|_{N_1^{\perp}}=0, \tag{21}$$

$$\left(\frac{\partial g_2}{\partial q_2}+\frac{\partial f_2}{\partial p_2}\right)|_{N_2^{\perp}}=0, \tag{22}$$

$$\frac{\partial g_2}{\partial q_2}(0)+\frac{\partial f_2}{\partial p_2}(0)=\frac{\partial f_1}{\partial p_2}(0)+\frac{\partial g_2}{\partial q_1}(0)=\frac{\partial g_1}{\partial q_1}(0)+\frac{\partial f_1}{\partial p_1}(0)=\frac{\partial g_1}{\partial q_2}(0)+\frac{\partial f_2}{\partial p_1}(0)=0,$$

$$\frac{\partial f_1}{\partial q_2}(0)-\frac{\partial f_2}{\partial q_1}(0)=\frac{\partial g_1}{\partial p_2}(0)-\frac{\partial g_2}{\partial p_1}(0)=0.$$

The conditions (21)-(22) mean that the vector field-germ $f_i|_{N_i^{\perp}}\frac{\partial}{\partial p_i}+g_i|_{N_i^{\perp}}\frac{\partial}{\partial q_i}$ on the symplectic manifold-germ $(N_i^{\perp},\omega|_{TN_i^{\perp}})$ is Hamiltonian (in the classical sense) for $i=1,2$ (see Proposition 12).

References

[1] V. I. Arnold, A. B. Givental *Symplectic geometry*, in Dynamical systems, IV, 1-138, Encyclopedia of Matematical Sciences, vol. 4, Springer, Berlin, 2001.

[2] V. I. Arnold, S. M. Gusein-Zade, A. N. Varchenko, *Singularities of Differentiable Maps*, Vol. 1, Birhauser, Boston, 1985.

[3] P.A.M. Dirac, *Generalized Hamiltonian Dynamics* , Canadian J. Math. **2**, (1950), 129-148.

[4] W. Domitrz, S. Janeczko, M. Zhitomirskii, *Relative Poincare lemma, contractibility, quasi-homogeneity and vector fields tangent to a singular variety*, Ill. J. Math. **48**, No.3 (2004), 803-835.

[5] W. Domitrz, S. Janeczko, M. Zhitomirskii, *Symplectic singularities of varieties: the method of algebraic restrictions*, J. reine und angewandte Math. **618** (2008), 197-235.

[6] W. Domitrz, S. Janeczko, M. Zhitomirskii, *Generic singularities of symplectic and quasi-symplectic immersions*, Math. Proc. Camb. Philos. Soc. **155** (2013), no. 2, 317–329.

[7] T. Fukuda, S. Janeczko, *Singularities of implicit differential systems and their integrability,* Banach Center Publications, **65**, (2004), 23-47.

[8] T. Fukuda, S. Janeczko, *Hamiltonian systems on submanifolds*, Advanced Studies in Pure Mathematics **78**, (2018), 221-249.

[9] G. Ishikawa, S. Janeczko, *Symplectic bifurcations of plane curves and isotropic liftings*, Q. J. Math. **54**, No.1 (2003), 73-102.

[10] G. Ishikawa, S. Janeczko, *Symplectic singularities of isotropic mappings*, Geometric singularity theory, Banach Center Publications **65** (2004), 85-106.

[11] S. Janeczko, *Constrained Lagrangian submanifolds over singular constraining varieties and discriminant varieties,* Ann. Inst. Henri Poincare, Phys. Theorique, **46**, No. 1, (1987), 1-26.

[12] S. Janeczko, *On implicit Lagrangian differential systems.*, Annales Polonici Mathematici, LXXIV, (2000), 133-141.

[13] S. Janeczko, F. Pelletier, *Singularities of implicit differential systems and Maximum Principle,* Banach Center Publications, **62**, (2004), 117-132.

[14] M. Zhitomirskii, *Relative Darboux theorem for singular manifolds and local contact algebra*, Can. J. Math. **57**, No.6 (2005), 1314-1340.

Irmina Herburt

Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

INTRINSIC METRIC IN SPACES OF COMPACT SUBSETS WITH THE HAUSDORFF METRIC

Manuscript received: 24 June 2020
Manuscript accepted: 3 August 2020

Abstract: We prove that, if a metric space (X,ρ) can be endowed with the intrinsic metric ρ^* (the intrinsic distance of two points is defined as the infimum of the lengths of arcs joining these points), then the Hausdorff metric ρ_H in the space $\mathcal{C}(X)$ of compact subsets of X induces the intrinsic metric $(\rho_H)^*$, and the equality

$$(\rho_H)^* = (\rho^*)_H$$

is satisfied. This implies that $\rho_H = (\rho_H)^*$ if and only if $\rho^* = \rho$ and that each isometry between spaces X_1 and X_2 with intrinsic metrics induces an isometry between $\mathcal{C}(X_1)$ and $\mathcal{C}(X_2)$ with intrinsic metrics.
Keywords: intrinsic metric, intrinsic isometry, hyperspace of compact sets, Hausdorff metric, arcs in hyperspace of compact sets
Mathematics Subject Classification (2020): 54E40, 54E35

1. INTRODUCTION

Let (X,ρ) be a strongly arc-wise connected metric space i.e. a space in which every two points can be joined by an arc of finite length. Then X can be endowed with the intrinsic metric ρ^* in which the distance of any two points is measured as the infimum of the lengths of arcs joining these points (see Section 2). The notion of intrinsic metric was widely investigated. Let us mention only the Blumenthal notion of geodesic ([4] p. 70) and his notion of convexification of a metric space ([4], p. 72 Ex. 1-4); the Borsuk geometrically acceptable (*GA*) metric spaces i.e. spaces with metric ρ^* topologically equivalent to ρ (compare [2], [4] p. 72 Ex. 4, [10], [11],); and the Burago length spaces ([5]). The Burago length space induced by a strongly arc-wise connected metric space (X,ρ) coincides with the space (X,ρ^*) ([5], Section 2.3).

We concentrate our investigation on hyper-space $\mathcal{C}(X)$ of compact subsets of X endowed with the Hausdorff metric ρ_H. The geometry of hyper-spaces of compact subsets has been largely developed in last few decades (mostly for $X = \mathrm{R}^n$), see [16], [15], [14], [1], [12], [6]. However, very little is known about arcs in hyper-spaces of compact subsets. Only metric segments in hyper-space $\mathcal{C}(R^n)$ have been intensively investigated by many authors ([13], [17], [6], [19]). Basic proofs of the existence of arcs in $\mathcal{C}(X)$ for metric continuum X can be found in [14], Ch.V, §47, VII (consult the references therein as well). Recently, some results concerning arcs in hyper-spaces were obtained in [8].

We shall prove that, if (X,ρ) is geometrically acceptable, then so is $(\mathcal{C}(X),\rho_H)$ and the equality

$$(\rho_H)^* = (\rho^*)_H \tag{1}$$

is satisfied i.e. the following diagram is commutative:

$$\begin{array}{ccc} (X,\rho) & \xrightarrow{H} & (\mathcal{C}(X),\rho_H) \\ \downarrow * & & \downarrow * \\ (X,\rho^*) & \xrightarrow{H} & (\mathcal{C}(X),(\rho^*)_H) \end{array}$$

There are no good methods to calculate the lengths of arcs in hyper-spaces of compact sets. Our result allows us to omit this problem and find $(\rho_H)^*$ by calculating lengths of arcs in the space X.

As an immediate consequence of (1) we obtain (see Corollary 13)

$$\rho = \rho^* \text{ in } X \text{ if and only if } \rho_H = (\rho_H)^* \text{ in } \mathcal{C}(X).$$

The analogous result for the space of bounded and closed subsets of a metric space, with the Hausdorff metric was proved in [20].

2. PRELIMINARIES

In this section we shall remind the basic notions of intrinsic geometry and of geometry of hyper-space of compact sets, we shall need in the sequel.

Let (X,ρ) be a metric space.

A subset of X isometric to a closed interval in R is a *metric segment* in X (by some authors: a geodesic segment, comp. [7]). We say that (X,ρ) is *metrically convex* (by some authors: a geodesic metric space, comp. [7]), if every pair of points $x,y \in X$ can be joined by a metric segment in X .

A *path* in X is a continuous image of a closed interval. A subset of X homeomorphic to a closed interval in R is *an arc* in X. For any path L joining points x and y in X we define the length $|L|_\rho$ by

$$|L|_\rho := \sup\{\sum_{i=1}^{k-1} \rho(x_i, x_{i+1}) : i = 1, 2.., k-1\},$$

where sup is taken over all naturally ordered sequences $x_1, x_2, ...x_k$ of points in L, with $x_1 = x$ and $x_k = y$. A path is *rectifiable* if its length is finite.

We say that (X, ρ) is *strongly arc-wise connected* if for every $x, y \in X$ there exists a rectifiable arc joining x and y in X.

Remark 1. *If (X, ρ_1) and (X, ρ_2) are metric spaces, (X, ρ_1) is strongly arc-wise connected and $\rho_1 \geq \rho_2$, then (X, ρ_2) is strongly arc-wise connected.*

In a strongly arc-wise connected metric space (X, ρ) the function $\rho^* : X \times X \longrightarrow$ R given by the formula

$$\rho^*(x, y) := \inf\{|L|_\rho : L \text{ is an arc in } X \text{ and } x, y \in L\}$$

is again a metric. It is called the *intrinsic metric* in (X, ρ) ([2], [5], [11]).

Let us note that we do not require that for any two points in X an arc with the shortest length exists. Since $\rho \leq \rho^*$, the identity map from (X, ρ^*) to (X, ρ) is continuous, however the metric ρ^* need not be topologically equivalent to ρ. For instance, for a compact space (X, ρ) the space (X, ρ^*) may be non-compact (for examples see [5]). Following Borsuk, we say that a strongly arc-wise connected space (X, ρ) is geometrically acceptable ($(X, \rho) \in GA$) whenever ρ and ρ^* are topologically equivalent i.e. the identity map from (X, ρ) to (X, ρ^*) is a homeomorphism ([2]).

It can be easily shown that $(\rho^*)^* = \rho^*$ and $|L|_\rho = |L|_{\rho^*}$ for any rectifiable arc L in X (compare [5], Proposition 2.3.12).

For any strongly arc-wise connected space (X, ρ) the metric ρ is *intrinsic* if $\rho = \rho^*$. If (X, ρ) is compact, strongly arc-wise connected and ρ is intrinsic, then (X, ρ) is metrically convex (compare [4], Theorem 28.1).

An *intrinsic isometry* of two *GA* spaces is an isometry with respect to their intrinsic metrics. Intrinsic isometries between *GA* spaces are homeomorphisms preserving lengths of arcs ([3]).

We shall consider the hyper-space $\mathcal{C}_\rho(X)$ of compact subsets of X endowed with the Hausdorff metric ρ_H.

For any A, B in $\mathcal{C}_\rho(X)$,

$$\rho_H(A, B) := \max\{\sup_{x \in A} \inf_{y \in B} \rho(x, y), \sup_{x \in B} \inf_{y \in A} \rho(x, y)\}.$$

For every nonempty subset $A \subset X$ and $\varepsilon > 0$, let

$$(A)_\varepsilon = \{x \in X : \inf_{a \in A} \rho(x, a) \leq \varepsilon\}.$$

The set $(A)_\varepsilon$ is called the ε-hull of A.

It is well known that

$$\rho_H(A,B) = \inf\{\alpha > 0 : A \subset (B)_\alpha \text{ and } B \subset (A)_\alpha\}.$$

If ρ_1 and ρ_2 are topologically equivalent metrics in X, then $\mathcal{C}_{\rho_1}(X) = \mathcal{C}_{\rho_2}(X)$ and the induced Hausdorff metrics $(\rho_1)_H$ and $(\rho_2)_H$ are topologically equivalent ([18], notes for section 1.8). If it does not lead to a confusion, we shall write $\mathcal{C}(X)$ for short, instead of $\mathcal{C}_\rho(X)$.

The convergence in the sense of the Hausdorff metric can be described in terms of convergent sequences of points (compare [18], p. 69).

Theorem 2. *The convergence* $\lim_{i\to\infty} A_i = A$ *in* $\mathcal{C}(X)$ *is equivalent to the following conditions taken together:*

(i) *each point in A is a limit of a sequence* $(x_i)_{i\in\mathrm{N}}$ *with* $x_i \in A_i$ *for* $i \in \mathrm{N}$*;*

(ii) *the limit of any convergent sequence* $(x_{i_j})_{j\in\mathrm{N}}$ *with* $x_{i_j} \in A_{i_j}$ *for* $j \in \mathrm{N}$ *belongs to A, and the sequence* $(A_i)_{i\in\mathrm{N}}$ *is bounded.*

Remark 3. *If metrics* ρ_1, ρ_2 *in X are topologically equivalent and* $\rho_1 \le \rho_2$, *then* $(\rho_1)_H \le (\rho_2)_H$ *in* $\mathcal{C}(X)$.

Remark 4. *The space* (X,ρ) *is compact if and only if the space* $(\mathcal{C}(X),\rho_H)$ *is compact.*

(compare [14], p. 47).

A ball in (X,ρ) with center $x \in X$ and radius r will be denoted by $B_\rho(x,r)$.

3. ARCS IN $\mathcal{C}(X)$ APPROXIMATING DISTANCE IN $(\rho^*)_H$

In this section we shall give an algorithm which allows us to construct arcs with lengths approximating $(\rho^*)_H(A,B)$ for any A, B in $\mathcal{C}(X)$. As a consequence we shall obtain that, if (X,ρ) is geometrically acceptable, then so is $(\mathcal{C}(X),\rho_H)$ and the inequality $(\rho_H)^* \le (\rho^*)_H$ is satisfied.

Lemma 5. *Let* $(X,\rho) \in GA$ *and let A and B be finite subsets of X. Then, for any* $\varepsilon > 0$*, there exists an arc L in the space* $(\mathcal{C}(X),(\rho^*)_H)$*, with ends A and B such that*

$$|L|_{(\rho^*)_H} \le (\rho^*)_H(A,B) + \varepsilon.$$

Proof. Let $A = \{a_1, a_2, \ldots, a_n\}$ for some $n \in \mathrm{N}$, and $B = \{b_1, b_2, \ldots, b_k\}$ for some $k \in \mathrm{N}$. Take $\varepsilon > 0$. For each a_i in A ($i \in \{1, 2, \ldots, n\}$) let a_i' be a point in B with the shortest ρ^* distance to a_i. By the definition of ρ^*, there is an arc L_i with ends a_i and a_i' such that $|L_i|_\rho \leq \rho^*(a_i, a_i') + \varepsilon$. Let $|L_i|_\rho = \alpha_i$.

We repeat the same construction for points in B. For each $b_j \in B$ ($j \in \{1, 2, \ldots, k\}$) let b_j' be a point in A with the shortest ρ^* distance to b_j and let K_j be an arc with ends b_j, b_j' such that $|K_j|_\rho \leq \rho^*(b_j, b_j') + \varepsilon$. Let $|K_j|_\rho = \beta_j$. Let $\lambda = \max_{i,j}\{\alpha_i, \beta_j\}$.

By the definition of the Hausdorff metric

$$(\rho^*)_H(A, B) = \max_{i,j}\{\rho^*(a_i, a_i'), \rho^*(b_j, b_j')\}.$$

Therefore,

$$\lambda \leq (\rho^*)_H(A, B) + \varepsilon. \tag{2}$$

Let $p_i : [0, \alpha_i] \longrightarrow L_i$ be the natural parametrization of L_i for $i \in \{1, 2, \ldots, n\}$ with $p_i(0) = a_i$ and $p_i(\alpha_i) = a_i'$. Obviously,

$$\rho^*(p_i(t), p_i(t')) \leq |t - t'| \text{ for } t, t' \in [0, \alpha_i].$$

Let $\bar{p}_i : [0, \lambda] \longrightarrow L_i$ be defined by

$$\bar{p}_i(t) = p_i(\frac{\lambda_i}{\lambda} t) \text{ for } t \in [0, \lambda].$$

Let us note that

$$\rho^*(\bar{p}_i(t), \bar{p}_i(t')) \leq \frac{\lambda_i}{\lambda}|t - t'| \leq |t - t'| \text{ for } t, t' \in [0, \lambda]. \tag{3}$$

In the same manner, we take the natural parametrization

$r_j : [0, \beta_j] \longrightarrow K_j$, with $r_j(0) = b_j$ and $r_j(\beta_j) = b_j'$ and define a reparametrization

$$\bar{r}_j(t) = r_j(\frac{\beta_j}{\lambda}(\lambda - t)) \text{ for } t \in [0, \lambda].$$

Thus, $\bar{r}_j(0) = b_j'$, $\bar{r}_j(\lambda) = b_j$ and

$$\rho^*(\bar{r}_i(t), \bar{r}_i(t')) \leq |t - t'| \text{ for } t, t' \in [0, \lambda]. \tag{4}$$

Let $p : [0, \lambda] \longrightarrow \mathcal{C}(X)$ be defined by

$$p(t) = \bigcup_{i,j}\{\bar{p}_i(t), \bar{r}_j(t)\} \text{ for } t \in [0, \lambda].$$

Denote $p(t)$ by A_t. By Theorem 2, p is a continuous embedding of $[0, \lambda]$ into $(\mathcal{C}(X), (\rho^*)_H)$ with $p(0) = A$, $p(\lambda) = B$. Thus, $p([0, \lambda])$ is arc-wise connected and there is an arc L in

$p([0,\lambda])$, with ends A and B. By the definition of the Hausdorff metric, in view of (3) and (4) we get,

$$(\rho^*)_H(A_t,A'_t) \le \max_{i,j}\{\rho^*(\bar{p}_i(t),\bar{p}_i(t')),\rho^*(\bar{r}_i(t),\bar{r}_i(t'))\} \le |t-t'|$$

for $t,t' \in [0,\lambda]$.

Therefore,

$$|L|_{(\rho^*)_H} \le |p([0,\lambda]|_{(\rho^*)_H} \le \lambda.$$

By (2), we get the claim. □

Lemma 6. *Let $(X,\rho) \in GA$ and $A \in \mathcal{C}(X)$. Then, for any $\varepsilon > 0$ there exists a finite set A_1 in X and an arc L in $(\mathcal{C}(X),(\rho^*)_H)$ joining A with A_1, such that $|L|_{(\rho^*)_H} \le \varepsilon$.*

Proof. For $\varepsilon > 0$ and $i \in \mathrm{N}$, let $\delta_i = \frac{\varepsilon}{3\cdot 2^i}$. Since compact sets are limits (in the Hausdorff metric) of finite sets, there exists a finite subset A_i of A, $A_i \in B_{(\rho^*)_H}(A,\delta_i)$, for any $i \in \mathrm{N}$. By Lemma 5, there is an arc L_i in $(\mathcal{C}(X),(\rho^*)_H)$ with ends A_i and A_{i+1} such that

$$|L_i|_{(\rho^*)_H} \le (\rho^*)_H(A_i,A_{i+1}) + \delta_i.$$

The sequence (A_i) is convergent to A in $(\rho^*)_H$ metric, thus there is an arc L in $\bigcup_{i=1}^{\infty} L_i$ joining A_1 and A such that

$$|L|_{(\rho^*)_H} \le \sum_{i=1}^{\infty} |L_i|_{(\rho^*)_H} \le \sum_{i=1}^{\infty} (\rho^*)_H(A_i,A_{i+1}) + \delta_i$$

(compare [9]). Since

$$(\rho^*)_H(A_i,A_{i+1}) \le (\rho^*)_H(A_i,A) + (\rho^*)_H(A,A_{i+1}) \le 2\delta_i,$$

we obtain immediately

$$|L|_{(\rho^*)_H} \le 3\sum_{i=1}^{\infty} \delta_i = \varepsilon.$$

□

Theorem 7. *Let $(X,\rho) \in GA$ and $A,B \in \mathcal{C}(X)$. Then, for any $\varepsilon > 0$ there exists an arc L in $(\mathcal{C}(X),(\rho^*)_H)$ joining A and B, with $|L|_{(\rho^*)_H} \le (\rho^*)_H(A,B) + \varepsilon$.*

Proof. For $\varepsilon > 0$, let $\varepsilon' = \frac{\varepsilon}{5}$. By Lemma 6, there are finite sets A_1 and B_1 and arcs L_1 and L_2 joining A with A_1 and B with B_1, respectively, such that $|L_i|_{(\rho^*)_H} \le \varepsilon'$ for $i = 1,2$. By Lemma 5, there is an arc L_3 with ends A_1, B_1, such that $|L_3|_{(\rho^*)_H} \le (\rho^*)_H(A_1,B_1) + \varepsilon'$. Thus there is an arc L in $L_1 \cup L_2 \cup L_3$ joining A and B with $|L|_{(\rho^*)_H} \le |L_1|_{(\rho^*)_H} + |L_2|_{(\rho^*)_H} + |L_3|_{(\rho^*)_H} \le (\rho^*)_H(A_1,B_1) + 3\varepsilon' \le (\rho^*)_H(A,B) + 5\varepsilon' = (\rho^*)_H(A,B) + \varepsilon$. □

Let us note that for any $A,B \in \mathcal{C}(X)$ we can use Proposition 5 to construct arcs whose lengths approximate $(\rho^*)_H(A,B)$. It is enough to take sufficiently dense finite subsets of A and B and follow the algorithm (see Example 5.1 Part 4).

Corollary 8. *Let $(X,\rho) \in GA$. Then $(\mathcal{C}(X),\rho_H) \in GA$ and $(\rho_H)^* \leq (\rho^*)_H$.*

Proof. Let $(X,\rho) \in GA$. By Theorem 7 the space $(\mathcal{C}(X),(\rho^*)_H)$ is strongly arc-wise connected. From $\rho \leq \rho^*$, by Remark 3, we immediately get $\rho_H \leq (\rho^*)_H$. Hence, for any arc L in $(\mathcal{C}(X),(\rho^*)_H)$, L is an arc in $(\mathcal{C}(X),\rho_H)$ and $|L|_{\rho_H} \leq |L|_{(\rho^*)_H}$. Therefore, $(\mathcal{C}(X),(\rho^*)_H)$ is strongly arc-wise connected and $(\rho_H)^* \leq (\rho^*)_H$. Let us note that

$$\rho_H \overset{top}{\approx} (\rho^*)_H \geq (\rho_H)^* \geq \rho_H.$$

Hence $(\mathcal{C}(X),\rho_H) \in GA$. □

4. METRICS $(\rho_H)^*$ VERSUS $(\rho^*)_H$

In this section we prove that $(\rho_H)^* = (\rho^*)_H$.

Proposition 9. *Let $p : [0,\lambda] \longrightarrow \mathcal{C}(X)$ be a parametrization of an arc. Then $\bigcup_{t\in[0,\lambda]} p(t)$ is compact in (X,ρ).*

Proof. Theorem 2 makes proving easy and standard. □

Theorem 10. *Let $(X,\rho) \in GA$. Let $A,B \in \mathcal{C}(X)$ and let $\mathcal{L}$ be a rectifiable arc in $(\mathcal{C}(X),\rho_H)$, with ends A and B. Then, for every $a \in A$ there is $b \in B$ and an arc L in X, with ends a and b, such that*

$$L \subset \bigcup \mathcal{L} \subset X \text{ and } |L|_\rho \leq |\mathcal{L}|_{\rho_H}.$$

Proof. Let $\mathcal{L}$ be a rectifiable arc in $(\mathcal{C}(X),\rho_H)$ with $|\mathcal{L}|_{\rho_H} = \lambda$. Let $p : [0,\lambda] \longrightarrow \mathcal{C}(X)$ be a natural parametrization of $\mathcal{L}$, $p(0) = A$, $p(\lambda) = B$ and $p(t) = A_t$ for $t \in [0,\lambda]$. Take $a \in A$. Let $t_{i,k} = \frac{\lambda\cdot k}{2^i}$ for $k = 0,1,\ldots 2^i$ and $i \in \mathrm{N}$. For any i, we choose sets $A_0,\ldots A_{\frac{\lambda\cdot k}{2^i}},\ldots,A_\lambda$. Next, for every i we define a set $P_i = \{x_{i,0},\ldots x_{i,k},\ldots x_{i,2^i}\}$. We start with $x_{i,0} = a$. If $x_{i,k}$ has been defined, then we define $x_{i,k+1}$ to be the nearest point to $x_{i,k}$ in $A_{\frac{\lambda\cdot(k+1)}{2^i}}$ with respect to ρ^* metric (if there are more than one nearest points, then we choose one at random). By Proposition 9, $\bigcup\mathcal{L}$ is compact, thus, by Remark 4, a sequence (P_i) has a convergent subsequence in $(\mathcal{L},\rho_H)$. For simplicity, we assume that (P_i) is convergent. Now we define a sequence (γ_i) of functions $\gamma_i : \{0,\ldots,\frac{\lambda\cdot k}{2^i},\ldots,\lambda\} \longrightarrow P_i$ as follows:

$$\gamma_i\left(\frac{\lambda\cdot k}{2^i}\right) = x_{i,k} \text{ for } k = 0,1,\ldots,2^i.$$

Let $\gamma(t) = \lim_{i\to\infty}\gamma_i(t)$, for $t = \frac{\lambda\cdot k}{2^i}$. Since (P_i) is convergent in the Hausdorff metric, by Theorem 2, the sequence $(\gamma_i(t))$ is convergent in $\bigcup\mathcal{L}$. For any i and for $t_k = \frac{\lambda\cdot k}{2^i}$ we have

$$\rho(\gamma_i(t_k),\gamma_i(t_{k+1})) \leq \rho_H(A_{t_k},A_{t_{k+1}}) \leq |t_k - t_{k+1}|.$$

Therefore,

$$\rho(\gamma_i(t_{k+j}),\gamma_i(t_k)) \leq |t_{k+j}-t_k|.$$

Thus, γ is a Lipschitz function on a dense subset of interval $[0,\lambda]$ and can be extended to the whole interval. Moreover, $|\gamma([0,\lambda])|_\rho \leq \lambda$. Hence, there is an arc L in $\gamma([0,\lambda])$, with ends $\gamma(0)$ and $\gamma(\lambda)$, such that $|L|_\rho \leq \lambda$. □

By Corollary 8 and Theorem 10 we obtain the main result.

Corollary 11. *Let $(X,\rho) \in GA$. Then $(\mathcal{C}(X),\rho_H) \in GA$ and $(\rho^*)_H = (\rho_H)^*$.*

Corollary 12. *Let $(X_i,\rho_i) \in GA$ for $i = 1,2$ and let $f : (X_1,\rho_1) \longrightarrow (X_2,\rho_2)$ be an intrinsic isometry. Then the induced map $\bar{f} : (\mathcal{C}(X_1),(\rho_1)_H) \longrightarrow (\mathcal{C}(X_2),(\rho_2)_H)$, defined by $\bar{f}(A) = f(A)$, for any compact subset A of X, is an intrinsic isometry.*

Proof. Let $f : (X_1,\rho_1) \longrightarrow (X_2,\rho_2)$ be an intrinsic isometry. Then, by the definition of intrinsic isometry,

$$f : (X_1,(\rho_1)^*) \longrightarrow (X_2,(\rho_2)^*)$$

is an isometry. Thus,

$$\bar{f} : (\mathcal{C}(X_1),((\rho_1)^*)_H) \longrightarrow (\mathcal{C}(X_2),((\rho_2)^*)_H)$$

is an isometry. By Corollary 11 we get the claim. □

Corollary 13. *$(X,\rho) \in GA$. Then $\rho = \rho^*$ if and only if $\rho_H = (\rho_H)^*$.*

Proof. Let $\rho = \rho^*$. By Corollary 11, we obtain $\rho_H = (\rho^*)_H = (\rho_H)^*$.

Now, let $\rho_H = (\rho_H)^*$. By Corollary 11, for any $a,b \in X$ we obtain $\rho(a,b) = \rho_H(\{a\},\{b\})$ $= (\rho_H)^*(\{a\},\{b\}) = (\rho^*)_H(\{a\},\{b\}) = \rho^*(a,b)$. □

5. EXAMPLES

We start with a construction of metric segments in $\mathcal{C}(X)$ with ρ_H metric.

Proposition 14. *Let X be metrically convex, $A,B \in \mathcal{C}(X)$ and let $\rho_H(A,B) = \alpha$. If $(A)_t$ and $(B)_t$ are compact for every $t \in [0,\alpha]$, then $M : [0,\alpha] \longrightarrow \mathcal{C}(X)$ defined by*

$$M(t) = (A)_t \cap (B)_{\alpha-t} \text{ for } t \in [0,\alpha]$$

is an isometric embedding into $\mathcal{C}(X)$ with ρ_H metric.

Proof. The proof is an easy adaptation of proofs in [6] (Lemma 3.6, Lemma 3.7, and Proposition 3.8). □

The next example illustrates the main result of the paper (see Corollary 11).

Example 15. *Let X be the union of two equilateral triangles in R^2 with vertices a_1, a_2, o and b_1, b_2, o, where $a_1 = \left(-\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$, $a_2 = \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)$, $o = (0,0)$, $b_1 = -a_2$, $b_2 = -a_1$. Let ρ be the Euclidean metric in R^2 restricted to X. Metrics ρ^* and ρ are topologically equivalent, so compact sets in both metrics coincide i.e. $\mathcal{C}_\rho(X) = \mathcal{C}_{\rho^*}(X)$. Let A and B be segments with ends a_1, a_2 and b_1, b_2, respectively. It is evident that $\rho_H(A,B) = \sqrt{3}$.*

Part 1. (see Figure 1).

Let us consider the space $(\mathcal{C}(X), (\rho^*)_H)$. It is obvious that $A, B \in \mathcal{C}(X)$ and $(\rho^*)_H(A,B) = 1 + \frac{\sqrt{3}}{2}$.

Part 2. (see Figure 2).

We shall construct an arc in $(\mathcal{C}(X), \rho_H)$ joining A and B. Let

$$A_t = \left\{ (x_1, x_2) \in X : x_1 = \frac{\sqrt{3}}{2} t - \frac{\sqrt{3}}{2} \right\}$$

and

$$B_t = \{(-x_1, x_2) : (x_1, x_2) \in A_t\},$$

for every $t \in [0,1]$. Functions $p : [0,1] \longrightarrow \mathcal{C}(X)$ and $p' : [0,1] \longrightarrow \mathcal{C}(X)$ defined by $p(t) = A_t$ and $p'(t) = B_t$ are isometric embeddings (with respect to ρ_H), since

$$\rho_H(A_t, A_{t'}) = |t - t'| = \rho_H(B_t, B_{t'})$$

for every $t \in [0,1]$. Moreover, $p(0) = A$, $p(1) = (0,0) = p'(0)$ and $p'(1) = B$. Thus, $p([0,1]) \cup p'([0,1])$ is an arc with length equal to 2 in $(\mathcal{C}(X), \rho_H)$, joining A and B (this implies $(\rho_H)^*(A,B) \leq 2$). However, as we can see below, it is not an arc with the shortest length.

Part 3. (see Figure 3).

We shall construct an arc in $(\mathcal{C}(X), \rho_H)$ with the ends A and B and with the length $1 + \frac{\sqrt{3}}{2}$. Let us note that the space (X, ρ^*) is metrically convex, $(\rho^*)_H(A,B) = \alpha = 1 + \frac{\sqrt{3}}{2}$ and convex hulls $(A)_t$ and $(B)_t$ (with respect to ρ^* metric) are compact for every $t \in [0, \alpha]$. Thus, by Proposition 14, the function $M : [0,\alpha] \longrightarrow (\mathcal{C}(X), (\rho^*)_H)$ defined by

$$M(t) = (A)_t \cap (B)_{\alpha - t} \text{ for } t \in [0, \alpha]$$

is an isometric embedding. Therefore, $M : [0,\alpha] \longrightarrow (\mathcal{C}(X), \rho_H)$ is a homeomorphic embedding and, by Remark 3, we get

$$|M([0,\alpha])|_{\rho_H} \leq \alpha.$$

Thus, by Corollary 11,

$$|M([0,\alpha])|_{\rho_H} = \alpha.$$

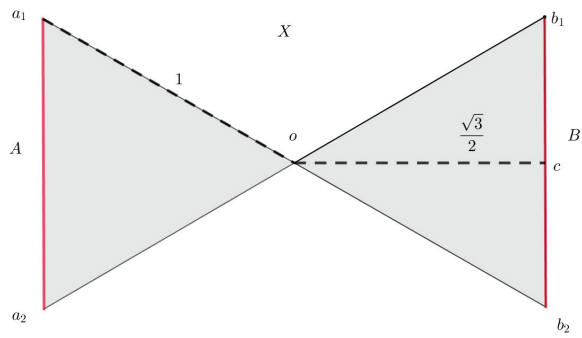


Fig. 1. $(\rho^*)_H(A,B) = 1 + \frac{\sqrt{3}}{2}$

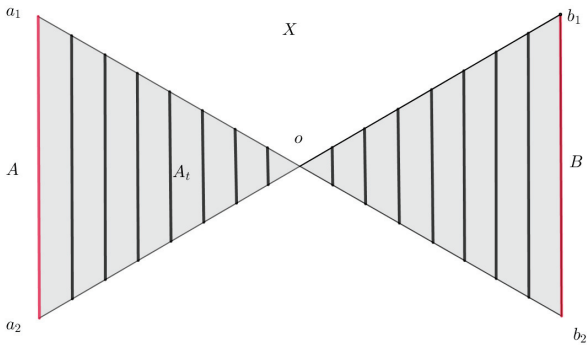


Fig. 2. Construction of an arc in $(\mathcal{C}(X), \rho_H)$ joining A and B, with length equal to 2

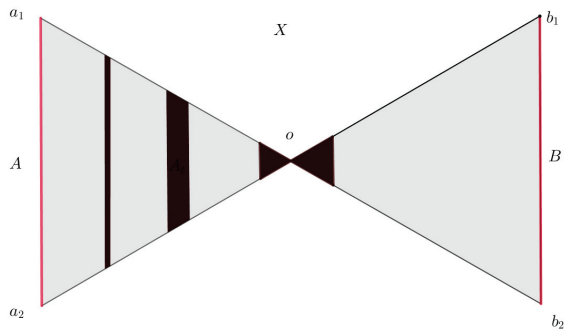


Fig. 3. Construction of an arc in $(\mathcal{C}(X),\rho_H)$ joining A and B, with the shortest length

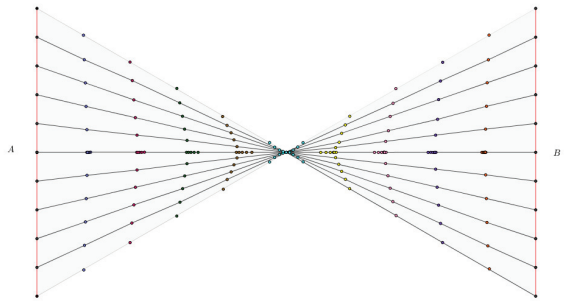


Fig. 4. Construction of an arc in $\mathcal{C}(X)$ which approximates $(\rho^*)_H(A,B)$

Part 4.

Figure 4 presents a method of constructing arcs whose lengths approximate $(\rho^*)_H(A,B)$ for sets A and B from Example 15. Here A_1 and B_1 from Lemma 5 are 10 - element subsets of A and B respectively.

Let us note that in general there are no good methods to find an arc with the shortest length in $(\mathcal{C}(X),\rho_H)$ (even if it exists). Moreover, to calculate lengths of arcs in this space may be a difficult task. The equality

$$(\rho_H)^* = (\rho^*)_H,$$

(see Corollary 11) allows us to calculate $(\rho_H)^*$ without constructing arcs in $(\mathcal{C}(X),\rho_H)$. Much simpler calculations of lengths of arcs in the space (X,ρ) are sufficient to find $(\rho_H)^*$ distance of two elements of $\mathcal{C}(X)$.

References

[1] Ch. Bandt, On the metric structures of hyperspaces with Hausdorff metric, Math. Nachr. 129, 1986, 175-183.
[2] K. Borsuk, On intrinsic isometries, Bull. Acad. Pol. Sci., 29 (1-2), 1981, 83-90.
[3] K. Borsuk, On homeomorphisms preserving lengths of arcs, Glasnik Mat. 16 (36), 1981, 307-311.
[4] L. M. Blumenthal, *Theory and applications of distance geometry*, Oxford Univ. Press, London, 1953.
[5] D. Burago, Yu. Burago, S. Ivanov, *A Course in Metric Geometry*, Graduate Studies in Math., AMS, 2001.
[6] A. Bogdewicz, Some metric properties of hyperspaces, Demonstratio Math., 32 (1), 2000, 135-149.
[7] M. R. Bridson, A. Haefliger, *Metric Spaces of Non-Positive Curvature*, Springer-Verlag, 1999.
[8] R. Dawson, Monotone arcs and Čebyšev arcs in hyperspaces, J. Geom. 98, 2010, 1-19.
[9] I. Herburt, Some remarks on the completion of uniformly geometrically acceptable spaces, Bull. Polish Acad. Sci. Math., 36 (3-4), 1988, 161-167.
[10] I. Herburt, On intrinsic isometries and rigid subsets of euclidean spaces, Demonstratio Math. 22 (4), 1989, 1205-1227.
[11] I. Herburt, M. Moszyńska, On intrinsic embeddings, Glas. Mat. 22 (42), 1987, 421-427.
[12] P. Gruber, The space of compact subsets of E^d, Geometriae Dedicata 9, 1980, 87-90.
[13] F. Jongmans, De l'art d'être a bonne distance des ensambles dont la decomposition atteint une stade avancé, Bull. Soc. Roy. Sci Liége, 48, 1979, 237-261.
[14] K. Kuratowski, *Topology*, vol II, Academic Press, New York, 1968 (French original 1961).
[15] E. Michael, Topologies on spaces of subsets, Trans. Amer. Math. Soc. 71, 1951, 152-182.
[16] S. B. Nadler, *Hyperspaces of sets*, Dekker, New York, 1978.
[17] R. Schneider, Pairs of convex bodies with unique joining metric segment, Bull. Soc. Roy. Sci. Liège 50, 1981, 5-7.
[18] R. Schneider, *Convex bodies: the Brunn-Minkowski Theory*, second expanded edition, Cambrige Univ. Press, 2013.
[19] S. Shlicker, Ch. Bay, A. Lembcke, When lines go bad in hyperspace, Demonstratio Math. 42 (2), 2009, 237-240.
[20] E. N. Sosov, On Hausdorff intrinsic metric, Lobachevskii J. Math., vol. 8, 2001, 185-189.

Jacek Jakubowski, Mariusz Niewęgłowski

Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

PRICING AND HEDGING IN LÉVY EXPONENTIAL MODEL WITH RATINGS:

FOURIER TRANSFORM APPROACH

Manuscript received: 27 July 2020
Manuscript accepted: 30 August 2020

Abstract: On a generalized Lévy exponential model with ratings we present the solutions of problems of pricing and hedging of a general payments stream process in terms of Fourier transforms. It is very important for applications of theory in practice since obtained results are easy to implement numerically.
Keywords: risk minimization, exponential Lévy model, Fourier transform
Mathematics Subject Classification (2020): 60K37, 60H30 (primary), 91G80, 91G20

1. INTRODUCTION

The problem of pricing and hedging non-attainable claims in incomplete markets is one of the fundamental problems in mathematical finance. It has been considered by many researchers, under various hedging criteria. One of the quite popular hedging criteria is risk-minimization introduced by Föllmer and Sondermann [7]. Their idea was to omit the self-financing condition and look for strategies that hedge a single claim H at time T perfectly and at the same time minimize the conditional variance of the remaining cost at each time t. They proved in [7] that there exists a unique risk minimizing hedging strategy for an arbitrary square integrable payoff at fixed maturity $T < \infty$ provided that the process of discounted price of the risky asset is a square integrable martingale. After [7], many papers have appeared dealing with the problem of finding an explicit formula for the risk minimizing strategy. Bouleau and Lamberton [1] have used the carré-du-champ operator to find a risk minimizing strategy for European options that are functions of the asset price at time T when the asset price is a function of a Markov process. Subsequently, Elliott and Föllmer [6] solved

a similar problem in the one-dimensional Markovian case as well as in the general case by using orthogonal martingale representation, i.e. the Galtchouk-Kunita-Watanabe decomposition (GKW for short). Møller [13] generalized the results of Föllmer and Sondermann to the case in which the liabilities of the hedger are described by an arbitrary square integrable and càdlàg payment process. It is also worth mentioning the papers by Dahl and Møller [5] and Dahl, Melchior and Møller [4] where explicit formulae for risk minimization of life insurance contracts that are subject to systematic mortality risk are derived. Norberg [15] (see also Norberg [14]) has proved, in a multidimensional setting and under the assumption that some predictable covariations are absolutely continuous, that finding the main component of a risk minimizing strategy can be reduced to solving a system of linear equations. Norberg [15] derives the equations for this component without referring to GKW decomposition as in [13]. In a recent paper Ceci, Cretarolla and Russo [2] found risk minimizing strategies under restricted information by solving backward stochastic differential equations driven by martingales. In [9] we derive the pricing and hedging formulae for financial contracts with a payment stream process D having some structure which also depends on the credit rating process. The formulae are given in terms of a solution of a certain Cauchy problem of integro-differential type. We note that these formulae in order to be applied in practice require first solving system of coupled integro-differential equations which is usually done numerically and then integrating numerically the solution with respect the Lévy measure. A realisation of this program can be very complicated. Therefore in this paper we proposed to use Fourier methods for solving problems of pricing and hedging in Lévy exponential model with ratings. We obtain formulae in which we first solve numerically linear matrix ODE and then integrate it numerically, so these formulae are much simpler to be applied in practice than these obtained in [9]. Our result is closely related to Tankov [18] who showed explicit formula for minimal variance portfolios in the case of European pay-offs and exponential Lévy models by using Fourier analysis techniques developed earlier by Hubalek, Kallsen, and Krawczyk [8]. Such a portfolio is closely related to a risk minimizing strategy (see discussion on page 547 in [17]) since it is solved by means of GKW decomposition. Tankov's formula is in terms of integrals of the Fourier transforms of the payoff and the characteristic function of the corresponding Lévy process. In this paper we generalize the technique introduced by Tankov [18] and apply them to pricing problem on a market with one risky asset which is subject to credit risk. The credit risk in our setup is modelled by a credit rating process which is assumed to be a finite state Markov chain whose evolution influences on the dynamic of a risky assets. We obtain formulae which are expressed by means of integral of Fourier transform of functions which appear in the payment process and a discounted conditional characteristic function of log prices. The paper is organised in the following way. In Section 2 we present, following [9], a model of market and methods of pricing and hedging using partial integro-differential equations (PIDE's). Since these results are not easy to implement in practice we develop theory giving results in terms of Fourier transforms, which are easy to implement numerically. It is done in Section 3. In Theorem 3 we give a price of a single payoff, in Theorem 5 an ex-dividend price of payments stream process D and in Theorem 7 the risk minimization strategy for D.

2. DESCRIPTION OF A MODEL

2.1. DESCRIPTION OF A (JUMP-DIFFUSION) MARKET MODEL

We consider a market on which there exists a bank account and we trade risky assets. We assume that the trading holds on interval $[0,T^*]$ with $T^* < \infty$. Our model of market takes into account ratings and jumps of price processes. These are modelled by a multidimensional process (S,C) on some filtered probability space $(\Omega,\mathcal{F},\mathbb{F},\mathbb{P})$. The process S is a process of price of tradable risky asset and the process C takes values in a finite state space $\mathcal{K} = \{1,\ldots,K\}$. It could be interpreted as a credit rating of company, so C_u represents a credit rating at time $u \leq T^*$ of a company whose stock price is given by S. We suppose that the investor can invest in a money account with the price process, denoted by B, depending on economic conditions of market which are described by the rating system C. The process B is given by a unique solution to

$$dB_u = \mathfrak{r}(C_{u-})B_u du, \quad B_0 = 1, \quad u \in [0,T^*], \tag{1}$$

where $\mathfrak{r}$ is a measurable and deterministic bounded function. Let β denote the discount factor process, i.e.,

$$\beta_t = B_t^{-1}. \quad t \in [0,T].$$

We also assume that the evolution of prices depends on credit rating or economic conditions of market which are described by C. So, we assume that our model is described by SDE in which the credit rating of company have impact on asset prices S by influence on drift and volatility. Moreover, a change in credit rating from j to k at time u causes a jump in prices of size $(e^{\Psi^{j,k}} - 1)$ in percentages. Taking into account these considerations we assume that the evolution of (S,C) is given as a unique weak solution of the following SDE

$$\begin{aligned} dS_t &= S_{t-}\Bigg(\mathfrak{r}(C_{t-})dt + \langle \Sigma(C_{t-}), dW_t\rangle + \int_{\mathbb{R}^n} (e^{\langle \Sigma(C_{t-}),x\rangle} - 1)\widetilde{\pi}(dx,dt) \\ &\qquad + \sum_{j,k\in\mathcal{K}:k\neq j} (e^{\Psi^{j,k}} - 1)\mathbb{1}_{\{j\}}(C_{t-})d\widetilde{M}_t^{j,k}\Bigg), \\ dC_t &= \sum_{j,k\in\mathcal{K}:k\neq j} (k-j)\mathbb{1}_{\{j\}}(C_{t-})dN_t^{j,k}, \end{aligned} \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbb{R}^n$, $\Sigma : \mathcal{K} \to \mathbb{R}^n$, $\Psi^{j,k} \in \mathbb{R}$, W is a standard n-dimensional Wiener process, $N^{j,k}$ are independent Poisson processes with constant intensities $\lambda^{j,k} > 0$ and $\pi(dx,dt)$ is a Poisson random measure on $\mathbb{R}^n \times [0,T^*]$ with the intensity measure $\rho(dx)dt$ satisfying, for some $m \geq 1$,

$$\int_{|x|>1} e^{2m\langle \Sigma(i),x\rangle}\rho(dx) < \infty \qquad \forall i \in \mathcal{K}. \tag{3}$$

Moreover, $\widetilde{\pi}(dx,dt)$ is the compensated Poisson random measure, i.e.,

$$\widetilde{\pi}(dx,dt) = \pi(dx,dt) - \rho(dx)dt,$$

and $\widetilde{M}^{j,k}$ are compensated Poissson processes $N^{j,k}$ given by

$$\widetilde{M}_t^{j,k} = N_t^{j,k} - \lambda^{j,k} t, \quad t \in [0,T^*].$$

Information available to the market participants is from an observation of process (S,C).

Our model generalize a regime switching model with jumps, extensively studied amongs others by Chourdakis [3], Mijatovic and Pistorius [12] or Kim et al. [11] for which $\Psi^{j,k} = 0$ in (2), so the part of dynamics driven by $M^{j,k}$ disappears. Note that the coefficients of SDE (2) satisfy standard assumptions (linear growth and Lipschitz conditions) for existence of a unique strong solution. We denote

$$H_u^j = \mathbb{1}_{\{C_u = j\}}, \qquad H_u^{j,k} = \int_0^u H_{v-}^j dH_v^k, \qquad u \in [0,T^*]$$

for $j,k \in \mathcal{K}$, $k \neq j$. The process $(H_u^{j,k})_{u\in[0,T^*]}$ counts the number of jumps of C from j to k up to time u and we have (see e.g. [9]).

$$H_u^{j,k} = \int_0^u H_{v-}^j dN_v^{j,k}.$$

This and the martingale property of $\widetilde{M}^{j,k}$ imply that the process defined by

$$M_u^{j,k} := H_u^{j,k} - \int_0^u H_{v-}^j \lambda^{j,k} dv, \qquad u \in [0,T^*], \tag{4}$$

is an $(\mathbb{F},\mathbb{P})$-martingale. Thus in view of martingale characterization theorem (see e.g. Rogers and Williams [16]) the coordinate C of solution (S,C) is a Markov chain with the state space $\mathcal{K}$ and $\lambda^{j,k}$ can be interpreted as transition intensity of C. The drift term in (2) implies that the discounted prices of tradable assets S are local martingales under the probability $\mathbb{P}$. So, $\mathbb{P}$ is a martingale measure, and hence there is no arbitrage on the market.

2.2. PRICING AND HEDGING VIA PIDE'S

We consider a contract between two parties, a seller (also called hedger) and a buyer, which specifies precisely the cash-flows between these two parties. These cash-flows are described by a càdlàg process D, i.e. D_t represents accumulated payments (outflows as well as injections of cash from the buyer) up to time t. The process D is called a payments stream process or a dividend process. Fix T, $T \leq T^*$. We consider here the dividend processes of the form

$$D_t = h(S_T,C_T)\mathbb{1}_{t\geq T} + \int_0^t g(S_u,C_u)du + \sum_{j,k\in\mathcal{K}:k\neq j} \int_0^t Z^{j,k}(S_{u-})dH_u^{j,k}, \tag{5}$$

$0 \le t \le T^*$, where h, g, $Z^{j,k}$ are continuous in s real-valued functions such that

$$\mathbb{E}\Big| \int_t^T \beta_t^{-1}\beta_u dD_u \Big|^2 < \infty \quad \text{for all } t \in [0,T]. \tag{6}$$

This payments stream process describes the following payments:

(i) The promised payment $h(S_T, C_T)$ which is paid at time T.

(ii) The promised coupons which are paid instantaneously at intensity $g(S_u, C_u)$ in $[0,T]$.

(iii) The payments $Z^{j,k}(S_{u-})$, which are paid at u provided that there is a change in the rating C from j to k at the moment u.

There are two fundamental problems that need to be studied for the financial contract introduced above, that is the pricing of the contract described by D and the hedging of this contract. We solve these problems on the market with processes of prices described by equations (1) and (2). As we know from [10], the problem of pricing of D boils down to computations of the *ex-dividend price process*, i.e.

$$V_t := \beta_t^{-1}\mathbb{E}\Big(\beta_T h(S_T,C_T) + \int_t^T \beta_u g(S_u,C_u)du + \sum_{j,k\in\mathcal{K}:k\neq j} \int_t^T \beta_u Z^{j,k}(S_{u-})dH_u^{j,k}\Big|\mathcal{F}_t\Big). \tag{7}$$

One can price and hedge such claims via associated Cauchy problems as it has been described in our paper [9]. Indeed, it is clear from the exponential form of component S of solution of (2), that $\mathfrak{D} = \mathbb{R}_+ \times \mathcal{K}$ is an invariant set for (2). So, assuming that $h, g, Z^{j,k} \in \mathcal{C}_m(\mathfrak{D})$, where

$$\mathcal{C}_m(\mathfrak{D}) = \{u : \mathfrak{D} \to \mathbb{R} :\ u(\cdot, j) \text{ is continuous and } |u(y,j)| \le K(1+|y|^m) \quad \forall j \in \mathcal{K}\},$$

and the function v is a sufficiently smooth solution of the following Cauchy problem

$$\begin{aligned}
&\partial_t v(t,s,j) + \nabla v(t,s,j)\mathfrak{r}(j)s + \frac{1}{2}s|\Sigma(j)|^2\nabla^2 v(t,s,j)\\
&\quad + \int_{\mathbb{R}^n}\Big(v(t,se^{\langle\Sigma(j),x\rangle},j) - v(t,s,j) - \nabla v(t,s,j)(e^{\langle\Sigma(j),x\rangle}-1)\Big)\rho(dx)\\
&\quad + \sum_{k\in\mathcal{K}\setminus j}\Big(v(t,se^{\Psi^{j,k}},k) - v(t,s,j) - \nabla v(t,s,j)(e^{\Psi^{j,k}}-1) + Z^{j,k}(t,s)\Big)\lambda^{j,k}\\
&\quad + g(t,s,j) = 0, \qquad (t,s,j) \in [0,T)\times\mathfrak{D},\\
&v(T,s,j) = h(s,j), \qquad (s,j) \in \mathfrak{D},
\end{aligned} \tag{8}$$

it is clear by Theorem 3.1 in [9] that the ex-dividend price process is given by

$$V_t = v(t, S_t, C_t).$$

The hedging problem is more complicated than the problem of pricing. First of all, the perfect hedging in the most cases is not possible, since the market introduced in Section 2.1 is usually incomplete. Thus, we need to specify the meaning of hedging that we wish to

execute. There are many concepts of hedging in the incomplete markets: the min-variance hedging, the indifference pricing or the (local) risk minimization amongst others. In this paper we will focus on the risk minimization approach introduced by Föllmer and Sondermann [7].

The idea of the risk minimization is to find a strategy which minimize the risk measured as a conditional variance of the remaining cost viewed from every time $t \leq T$. We denote by (φ, η) a strategy that describes the number of assets held in the portfolio at time t, i.e. φ is the number of risky assets and η describes the financial position on the bank account. By Theorem 3.2 from [9] it follows that we can find a 0-achieving risk-minimizing strategy for D by means of solving the related Cauchy problem (8) and then some system of linear equations.

Using Theorem 3.2 from [9] we see that the component φ has on the set $\{C_{t-} = j\}$ the representation

$$\varphi_t = (\widehat{G}_t^j)^{-1} \left(|\Sigma(j)|^2 \nabla v(t, S_{t-}, j) + \int_{\mathbb{R}^n} (e^{\langle \Sigma(j), x \rangle} - 1) \frac{v(t, S_{t-} e^{\langle \Sigma(j), x \rangle}, j) - v(t, S_{t-}, j)}{S_{t-}} \rho(dx) \right.$$
$$\left. + \sum_{k \in \mathcal{K}: k \neq j} (e^{\Psi^{j,k}} - 1) \frac{(v(t, S_{t-} e^{\Psi^{j,k}}, k) - v(t, S_{t-}, j) + Z^{j,k}(t, S_{t-}))}{S_{t-}} \lambda^{j,k} \right),$$

where

$$\widehat{G}_t^j := \left(|\Sigma(j)|^2 + \int_{\mathbb{R}^n} (e^{\langle \Sigma(j), x \rangle} - 1)^2 \rho(dx) + \left(\sum_{k \in \mathcal{K}: k \neq j} (e^{\Psi^{j,k}} - 1)^2 \lambda^{j,k} \right) \right).$$

and the component η of optimal strategy is given by

$$\eta_t = \beta_t \big(v(t, S_t, C_t) 1_{\{t<T\}} - \varphi_t S_t \big).$$

We can use Theorem 3.2 from [9] since condition (3) implies conditions (2.8) and (2.11) in [9].

3. PRICING AND HEDGING VIA FOURIER METHODS

In the previous section we described how problems of pricing and hedging can be solved using solutions of Cauchy problems. Now we will show, under some assumptions on functions $h, g, Z^{j,k}$, that we can find risk minimization strategy for D given by (5) via Fourier transform methods.

As in the previous section we assume that $h, g, Z^{j,k} \in \mathcal{C}_m(\mathfrak{D})$. We start with the following lemma which is of fundamental importance in this paper. The lemma gives the dynamic of

the log-prices of risky asset as well as the formula for discounted conditional characteristic function of the the log-price process in terms of matrix valued linear ODE which can be solved numerically. This will enable us, subsequently, to use Fourier transform methods to compute conditional expectation in terms of the related ODEs and the Fourier transforms of some functions.

Lemma 1. *Suppose that* (S,C) *is a unique solution of the system of SDEs* (2). *Then* $Y_t := \ln S_t$ *has the following dynamics*

$$\begin{aligned} Y_t := Y_0 &+ \int_0^t \Big(\mathfrak{r}(C_{s-}) - \mathcal{J}_1(-i\Sigma(C_{s-})) - \mathcal{J}_2(C_{s-})\Big)ds + \int_0^t \langle \Sigma(C_{s-}), dL_s\rangle \\ &+ \sum_{j,k\in\mathcal{K}, j\neq k} \Psi^{j,k} dH_s^{j,k}, \end{aligned} \tag{9}$$

where L *is an* $\mathbb{R}^K$*-valued Levy process with the characteristic triple* $(0, Id, \rho)$,

$$\begin{aligned} \mathcal{J}_1(u) &= -\frac{\langle u,u\rangle}{2} + \int_{\mathbb{R}^n} \Big(e^{i\langle u,x\rangle} - 1 - i\langle u,x\rangle \mathbb{1}_{\{|x|\leq 1\}}\Big)\rho(dx), \\ \mathcal{J}_2(j) &= \sum_{k\in\mathcal{K}, k\neq j} \Big(e^{\Psi^{j,k}} - 1\Big)\lambda^{j,k}. \end{aligned} \tag{10}$$

Moreover, for any $u \in \mathbb{R}$

$$\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{iuY_T}\mathbb{1}_{\{C_T=k\}}|\mathcal{F}_t\Big) = e^{iuY_t}\phi_{C_t,k}(t;T,u), \tag{11}$$

where $\phi(t;T,u) = [\phi_{j,k}(t;T,u)]_{j,k\in\mathcal{K}}$ *is a solution to the system of matrix valued ODEs*

$$d\phi(t;T,u)` = -(\Lambda + \Theta(u))\phi(t;T,u)dt, \qquad \phi(T;T,u) = \mathbb{I} \tag{12}$$

with a matrix function Θ *defined by*

$$\Theta_{j,k}(u) = \begin{cases} -\mathfrak{r}(j) + \mathcal{J}_1(u\Sigma(j)) + iu(\mathfrak{r}(j) - \mathcal{J}_1(-i\Sigma(j)) - \mathcal{J}_2(j)), & j = k, \\ \lambda^{j,k}(e^{iu\Psi^{j,k}} - 1), & j \neq k. \end{cases}$$

Proof. The first part follows from Itô's lemma. For the second part we use a Feynman-Kac argument. First we show that

$$v(t,y,j) = e^{iuy}\phi_{j,k}(t;T,u) \tag{13}$$

is a classical solution of a Cauchy problem

$$(\partial_t + \mathcal{A})v(t,y,j) = \mathfrak{r}(j)v(t,y,j), \qquad v(T,y,j) = e^{iuy}\mathbb{1}_{\{j=k\}}. \tag{14}$$

Indeed, under hypothesis (13) we have

$$\partial_y v(t,y,j) = iue^{iuy}\phi_{j,k}(t;T,u), \qquad \partial_{yy}^2 v(t,y,j) = -u^2 e^{iuy}\phi_{j,k}(t;T,u),$$

$$v(t,y+\langle\Sigma(j),x\rangle,j)-v(t,y,j)-\partial_y v(t,y,j)\langle\Sigma(j),x\rangle\mathbb{1}_{\{|x|\le 1\}}$$
$$=e^{iuy}\phi_{j,k}(t;T,u)\Big(e^{iu\langle\Sigma(j),x\rangle}-1-iu\langle\Sigma(j),x\rangle\mathbb{1}_{|x|\le 1}\Big),$$

$$v(t,y+\Psi^{j,m},m)-v(t,y,j)=e^{iuy}(\phi_{m,k}(t;T,u)e^{iu\Psi^{j,m}}-\phi_{j,k}(t;T,u)).$$

Hence, substituting these expressions in (14) and using (12) yield

$$(\partial_t+\mathcal{A})v(t,y,j)=e^{iuy}\Big(\partial_t\phi_{j,k}(t;T,u)+iu(\mathfrak{r}(j)-\mathcal{J}_1(-i\Sigma(j))-\mathcal{J}_2(j))\phi_{j,k}(t;T,u)$$
$$-\frac{\langle\Sigma(j),\Sigma(j)\rangle u^2}{2}\phi_{j,k}(t;T,u)+\phi_{j,k}(t;T,u)\int_{\mathbb{R}^n}\Big(e^{iu\langle\Sigma(j),x\rangle}-1-iu\langle\Sigma(j),x\rangle\mathbb{1}_{|x|\le 1}\Big)\rho(dx)$$
$$+\sum_{m\in\mathcal{K},m\neq j}(\phi_{m,k}(t;T,u)e^{iu\Psi^{j,m}}-\phi_{j,k}(t;T,u))\lambda^{j,m}\Big)=\mathfrak{r}(j)e^{iuy}\phi_{j,k}(t;T,u)=\mathfrak{r}(j)v(t,y,j),$$

which proves that v solves the Cauchy problem (14). Thus, using Itô's lemma we obtain

$$S_t^{iu}\phi_{C_t,k}(t;T,u)\beta_t=S_0^{iu}\phi_{C_0,k}(0;T,u)+iu\int_0^t\beta_s S_{s-}^{iu}\phi_{C_{s-},k}(s;T,u)\langle\Sigma(C_{s-}),dW_s\rangle$$
$$+\int_0^t\int_{\mathbb{R}^n}\beta_s S_{s-}^{iu}\phi_{C_{s-},k}(s;T,u)(e^{iu\Sigma(C_{s-})x}-1)\widetilde{\pi}(ds,dx) \tag{15}$$
$$+\int_0^t\sum_{j,m\in\mathcal{K}:m\neq j}\beta_s S_{s-}^{iu}(\phi_{m,k}(s;T,u)e^{iu\Psi^{j,m}}-\phi_{j,k}(s;T,u))dM_s^{j,m}.$$

This shows that $(S_t^{iu}\phi_{C_t,k}(t;T,u)\beta_t)_{t\in[0,T]}$ is a local martingale. Now, we note that it is uniformly integrable and hence it is a martingale. □

Remark 2. *Equality* (11) *is also valid for $u\in\mathbb{C}$ satisfying an additional assumption. Let $u=a+ib\in\mathbb{C}$ be such that*

$$b\Sigma(i)\in\Big\{v:\int_{|x|>1}(e^{2\langle v,x\rangle}+e^{\langle v,x\rangle})\rho(dx)<\infty\Big\}$$

for every $i\in\mathcal{K}$.

This assumption implies that

$$\mathbb{E}\Big(\sup_{t\in[0,T]}|S_t|^{2b}\Big)<\infty,$$

and hence $(S_t^{iu}\phi_{C_t,k}(t;T,u)\beta_t)_{t\in[0,T]}$ is still uniformly integrable for such chosen complex number u.

Let us introduce some convenient notation. For a given function $g:\mathbb{R}\times\mathcal{K}\to\mathbb{R}$ by g^e we denote the modified payoff function

$$g^e(y,c):=g(e^y,c),$$

and for a real number R we denote by g^e_R the dampened modified function g^e, i.e.

$$g^e_R(y,k) = g^e(y,k)e^{-Ry} = g(e^y,k)e^{-Ry}.$$

Moreover, by $\widehat{g}$ we denote the vector valued function $\widehat{g}(u) = (\widehat{g}(u,1),\dots\widehat{g}(u,K))^\top$, where $\widehat{g}(\cdot,c)$ is the Fourier transform of $g(\cdot,c)$, i.e.

$$\widehat{g}(u,c) := \int_{\mathbb{R}} e^{iuy} g(y,c)dy.$$

Note that we (obviously) have

$$\widehat{g^e_R}(u,c) = \widehat{g^e}(u+iR,c).$$

In the theorem below we show how the Fourier methods can be applied to solve the pricing problem for single payoff $h(S_T,C_T)$ at time $T \le T^*$.

Theorem 3. *Let h be a given payoff function $h : \mathbb{R}_+ \times \mathcal{K} \to \mathbb{R}$. Suppose that there exists R such that:*

i) the dampened modified payoff function h^e_R has properties

$$y \to h^e_R(y,k) \text{ is in } L^1(\mathbb{R}) \cap \mathcal{C}_b(\mathbb{R}), \qquad u \to \widehat{h^e_R}(u,k) \text{ is in } L^1(\mathbb{R}),$$

ii)

$$\mathbb{E}e^{RY_T} < \infty.$$

Then

$$\beta_t^{-1}\mathbb{E}(\beta_T h(S_T,C_T)|\mathcal{F}_t) = \frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\langle H_t, \phi(t;T,-u-iR)\widehat{h^e}(u+iR)\rangle du, \tag{16}$$

where ϕ is the unique solution of ODE (12) *and*

$$H_t := (\mathbb{1}_{\{1\}}(C_t),\dots,\mathbb{1}_{\{K\}}(C_t))^\top.$$

Proof. Recall that $S_t = e^{Y_t}$. We have

$$\begin{aligned}\beta_t^{-1}\mathbb{E}(\beta_T h(S_T,C_T)|\mathcal{F}_t)C_T)|\mathcal{F}_t) &= \mathbb{E}(\beta_t^{-1}\beta_T e^{RY_T} h^e_R(Y_T,C_T)|\mathcal{F}_t)\\ &= \mathbb{E}\Big(\beta_t^{-1}\beta_T e^{RY_T}\frac{1}{2\pi}\int_{\mathbb{R}} e^{-iuY_T}\widehat{h^e_R}(u,C_T)du|\mathcal{F}_t\Big).\end{aligned}$$

By assumption ii),

$$\mathbb{E}\Big(\int_{\mathbb{R}}\Big|\beta_t^{-1}\beta_T e^{RY_T}e^{-iuY_T}\widehat{h^e_R}(u,C_T)\Big|du\Big) \le K\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{RY_T}\Big) < \infty,$$

so we can apply the Fubini theorem and Lemma 1 in penultimate equality, which yield

$$\begin{aligned}
&\beta_t^{-1}\mathbb{E}(\beta_T h(S_T,C_T)|\mathcal{F}_t)\\
&=\frac{1}{2\pi}\int_{\mathbb{R}}\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{RY_T}e^{-iuY_T}\widehat{h_R^e}(u,C_T)|\mathcal{F}_t\Big)du\\
&=\frac{1}{2\pi}\int_{\mathbb{R}}\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{RY_T}e^{-iuY_T}\widehat{h^e}(u+iR,C_T)|\mathcal{F}_t\Big)du\\
&=\sum_{k\in\mathcal{K}}\frac{1}{2\pi}\int_{\mathbb{R}}\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{RY_T}e^{-iuY_T}\mathbb{1}_{\{C_T=k\}}|\mathcal{F}_t\Big)\widehat{h^e}(u+iR,k)du\\
&=\sum_{k\in\mathcal{K}}\frac{1}{2\pi}\int_{\mathbb{R}}\mathbb{E}\Big(\beta_t^{-1}\beta_T e^{i(-u-iR)Y_T}\mathbb{1}_{\{C_T=k\}}|\mathcal{F}_t\Big)\widehat{h^e}(u+iR,k)du\\
&=\sum_{k\in\mathcal{K}}\frac{1}{2\pi}\int_{\mathbb{R}}e^{-iY_t(u+iR)}\phi_{C_t,k}(t;T,-u-iR,k)\widehat{h^e}(u+iR,k)du\\
&=\sum_{k\in\mathcal{K}}\frac{1}{2\pi}\int_{\mathbb{R}}e^{Y_t(R-iu)}\phi_{C_t,k}(t;T,-u-iR,k)\widehat{h^e}(u+iR,k)du.
\end{aligned}$$

The proof is complete. □

In most cases formula (16) is applied for $t=0$, in such case (16) takes the form

$$\begin{aligned}
\mathbb{E}(\beta_T h(S_T,C_T)) &= \frac{1}{2\pi}\int_{\mathbb{R}}S_0^{R-iu}\langle H_0,\phi(0;T,-u-iR)\widehat{h^e}(u+iR)\rangle du\\
&= \frac{1}{2\pi}\int_{\mathbb{R}}e^{-iy(u+iR)}\langle H_0,\phi(0;T,-(u+iR))\widehat{h_R^e}(u)\rangle du,
\end{aligned}$$

where $y:=\ln S_0$. Note that the above integral involves $\widehat{h_R^e}$ which is the Fourier transform of modified dampened payoff function and $\phi(0;T,\cdot)$ which is the discounted characteristic function of log-prices at time T (extended to complex domain cf. Remark 2). This integral can be efficiently approximated via Fast Fourier Transform methods provided that $\widehat{h_R^e}$ is known explicitly.

The following proposition follows from Theorem 3 and shows how the pricing problem can be solved by using a Fourier integral and a linear vector valued ODE:

Proposition 4. *Let h and f be measurable functions $\mathbb{R}_+\times\mathcal{K}\mapsto\mathbb{R}$. Suppose that there exist a constant R such that:*

i) The functions

$$\begin{aligned}
&y\to h_R^e(y,k) \quad \text{and} \quad y\to f_R^e(y,k) \text{ are in } L^1(\mathbb{R})\cap\mathcal{C}_b(\mathbb{R}),\\
&u\to \widehat{h_R^e}(u,k) \quad \text{and} \quad u\to \widehat{f_R^e}(u,k) \text{ are in } L^1(\mathbb{R}).
\end{aligned}$$

ii)

$$\mathbb{E}\sup_{v\in[0,T]}e^{RY_v}<\infty.$$

Then

$$\beta_t^{-1}\mathbb{E}\Big(\beta_T h(S_T,C_T)+\int_t^T \beta_v f(S_v,C_v)dv|\mathcal{F}_t\Big)=\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\langle H_t,\Phi(t;T,u+iR)\rangle du, \tag{17}$$

where $\Phi=(\Phi_1,\dots,\Phi_K)^\top$ *is a solution of the vector valued ODE*

$$\partial_t\Phi(t,u)+(\Lambda+\Theta(-u))\Phi(t,u)=-\widehat{f^e}(u),\quad \Phi(T,u)=\widehat{h^e}(u). \tag{18}$$

Proof. Note that the solution Φ of (18) can be written as

$$\Phi(t,u)=\phi(t;T,-u)\widehat{h^e}(u)+\int_t^T \phi(t;v,-u)\widehat{f^e}(u)dv, \tag{19}$$

where ϕ is a solution of ODE (12). Hence (17) is equivalent to

$$\begin{aligned}&\beta_t^{-1}\mathbb{E}\Big(\beta_T h(S_T,C_T)+\int_t^T \beta_v f(S_v,C_v)dv|\mathcal{F}_t\Big)\\&\quad=\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\Big\langle H_t,\phi(t;T,-u-iR)\widehat{h^e}(u+iR)+\int_t^T \phi(t;v,-u-iR)\widehat{f^e}(u+iR)dv\Big\rangle du.\end{aligned}$$

Thus, in view of Theorem 3, it suffices to show that

$$\beta_t^{-1}\mathbb{E}\Big(\int_t^T \beta_v f(S_v,C_v)dv|\mathcal{F}_t\Big)=\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\Big\langle H_t,\int_t^T \phi(t;v,-u-iR)\widehat{f^e}(u+iR)dv\Big\rangle du.$$

Towards this end, note that by the Fubini theorem and Theorem 3 we have

$$\begin{aligned}\beta_t^{-1}\mathbb{E}\Big(\int_t^T \beta_v f(S_v,C_v)dv|\mathcal{F}_t\Big)&=\int_t^T \mathbb{E}\Big(\beta_t^{-1}\beta_v f(S_v,C_v)|\mathcal{F}_t\Big)dv\\&=\int_t^T\Big(\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\langle H_t,\phi(t;v,-u-iR)\widehat{f^e}(u+iR)\rangle du\Big)dv\\&=\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\Big\langle H_t,\int_t^T \phi(t;v,-u-iR)\widehat{f^e}(u+iR)dv\Big\rangle du.\end{aligned}$$

This ends the proof. □

Now we can summarize the above results and obtain the ex-dividend price of D using Propostion 4, (5) and (7).

Theorem 5. *Suppose that the payoff function* $h:\mathbb{R}_+\times\mathcal{K}\mapsto\mathbb{R}$ *and*

$$f(y,j):=g(y,j)+\sum_{k\neq j} Z^{j,k}(y)\lambda^{j,k}. \tag{20}$$

satisfy assumptions i) and ii) of Proposition 4. Then the ex-dividend price of D is given by

$$V_t=\frac{1}{2\pi}\int_{\mathbb{R}} S_t^{R-iu}\langle H_t,\Phi(t;T,u+iR)\rangle du,$$

where Φ *is the unique solution of the ODE* (18) *with the function* f *given by* (20)*.*

In the following lemma we show the dynamic of the process under the integral in formula (17). This result together with Proposition 4 will be used in the proof our second main result of the paper giving hedging strategy.

Lemma 6. *Suppose that the assumptions of Proposition 4 are in force. Let Φ be the unique solution of ODE* (18)*. Then*

$$\begin{aligned}
&d(S_t^{R-iu}\Phi_{C_t}(t,T,u+iR)) \\
&= S_{t-}^{R-iu}\Phi_{C_{t-}}(t,T,u+iR))\Big[(R-iu)\langle\Sigma(C_{t-}),dW_t\rangle + \int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1\Big)\widetilde{\pi}(ds,dy)\Big] \\
&\quad + \sum_{k,j:j\neq k} S_{t-}^{R-iu}\Big(\Phi_k(t,T,u+iR)e^{(R-iu)\Psi^{j,k}} - \Phi_j(t,T,u+iR)\Big)dM_t^{j,k} \\
&\quad + S_{t-}^{R-iu}\Big(\Phi_{C_{t-}}(t,T,u+iR))\mathfrak{r}(C_{t-}) - \widehat{f}^e(u+iR,C_{t-})\Big)dt.
\end{aligned} \tag{21}$$

Proof. Using integration by parts formula we have

$$\begin{aligned}
d(S_t^{R-iu}\Phi_{C_t}(t,T,u+iR)) &= (dS_t^{R-iu})\Phi_{C_{t-}}(t,T,u+iR)) \\
&\quad + S_{t-}^{R-iu}d\Phi_{C_t}(t,T,u+iR)) + \Delta S_t^{R-iu}\Delta\Phi_{C_t}(t,T,u+iR).
\end{aligned} \tag{22}$$

Now we calculate the components of the right hand side of (22). Using Itô's lemma note that

$$\begin{aligned}
dS_t^{R-iu} &= (R-iu)S_{t-}^{R-iu}\frac{dS_t}{S_{t-}} + \frac{1}{2}(R-iu)(R-1-iu)S_t^{R-iu}\langle\Sigma(C_{t-}),\Sigma(C_{t-})\rangle dt \\
&\quad + S_{t-}^{R-iu}\Big(\Big(1+\frac{\Delta S_t}{S_{t-}}\Big)^{R-iu} - 1 - (R-iu)\frac{\Delta S_t}{S_{t-}}\Big).
\end{aligned}$$

Then substituting (2) we can write this in the form

$$\begin{aligned}
dS_t^{R-iu} &= S_{t-}^{R-iu}\Big((R-iu)\langle\Sigma(C_{t-}),dW_t\rangle + \sum_{k,j\in\mathcal{K}:j\neq k}\Big(e^{(R-iu)\Psi^{j,k}}-1\Big)\mathbb{1}_{\{j\}}(C_{t-})dM_t^{j,k} \\
&\qquad + \int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1\Big)\widetilde{\pi}(dt,dy) + (R-iu)\mathfrak{r}(C_{t-})dt\Big) \\
&\quad + \frac{1}{2}((R-iu)^2-(R-iu))S_t^{R-iu}\langle\Sigma(C_{t-}),\Sigma(C_{t-})\rangle dt \\
&\quad + S_{t-}^{R-iu}\sum_{k,j\in\mathcal{K}:j\neq k}\Big(e^{(R-iu)\Psi^{j,k}}-1\Big)\mathbb{1}_{\{j\}}(C_{t-})\lambda^{j,k}dt \\
&\quad - (R-iu)S_{t-}^{R-iu}\sum_{k,j\in\mathcal{K}:j\neq k}(e^{\Psi^{j,k}}-1)\mathbb{1}_{\{j\}}(C_{t-})\lambda^{j,k}dt \\
&\quad + S_{t-}^{R-iu}\int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1-(R-iu)\langle\Sigma(C_{t-}),y\rangle\mathbb{1}_{|y|\leq 1}\Big)\rho(dy)dt \\
&\quad - (R-iu)S_{t-}^{R-iu}\int_{\mathbb{R}^n}\Big(e^{\langle\Sigma(C_{t-}),y\rangle}-1-\langle\Sigma(C_{t-}),y\rangle\mathbb{1}_{|y|\leq 1}\Big)\rho(dy)dt.
\end{aligned}$$

This, in turn, can be written in terms of $\mathcal{J}_1$ and $\mathcal{J}_2$ (see (10)) as

$$\begin{aligned} dS_t^{R-iu} &= S_{t-}^{R-iu}\Big[(R-iu)\langle\Sigma(C_{t-}),dW_t\rangle + \sum_{k,j\in\mathcal{K}:j\neq k}\Big(e^{(R-iu)\Psi^{j,k}}-1\Big)\mathbb{1}_{\{j\}}(C_{t-})dM_t^{j,k} \\ &+ \int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1\Big)\widetilde{\pi}(dt,dy) + \mathcal{J}_1(-(u+iR)) + \sum_{k\neq C_{t-}}\Theta_{C_{t-},k}(-(u+iR)) \\ &+ (R-iu)\Big(\mathfrak{r}(C_{t-}) - \mathcal{J}_1(-i\Sigma(C_{t-})) - \mathcal{J}_2(C_{t-})\Big)dt\Big]. \end{aligned} \tag{23}$$

Now, let us calculate differential $d\Phi_{C_{t-}}(t;T,u+iR)$. Applying Itô's lemma we obtain

$$\begin{aligned} d\Phi_{C_{t-}}(t;T,u+iR) &= \sum_{k:k\neq C_{t-}}(\Phi_k(t;T,u+iR)-\Phi_{C_{t-}}(t;T,u+iR))\lambda^{C_{t-},k}dt \\ &+ \partial_t\Phi_{C_{t-}}(t;T,u+iR)dt + \sum_{k,j:k\neq j}(\Phi_k(t;T,u+iR)-\Phi_j(t;T,u+iR))dM_t^{j,k}. \end{aligned}$$

From (18) we get

$$\begin{aligned} d\Phi_{C_{t-}}(t;T,u+iR) &= -\Big(\sum_j\Theta_{C_{t-},j}(-u-iR)\Phi_j(t;T,u+iR) + \widehat{f}^e(u+iR,C_{t-})\Big)dt \\ &+ \sum_{k,j:j\neq k}(\Phi_k(t;T,u+iR)-\Phi_j(t;T,u+iR))dM_t^{j,k}. \end{aligned} \tag{24}$$

Using this and (23) yield

$$\begin{aligned} &\Delta S_t^{R-iu}\Delta\Phi_{C_t}(t,T,u-iR) = S_{t-}^{R-iu}\Big(\Big(\frac{S_t}{S_{t-}}\Big)^{R-iu}-1\Big)\Big(\Phi_{C_t}(t,T,u-iR)-\Phi_{C_{t-}}(t,T,u-iR)\Big) \\ &= \sum_{k\neq C_{t-}}S_{t-}^{R-iu}\Big(e^{(R-iu)\Psi^{C_{t-},k}}-1\Big)\Big(\Phi_k(t,T,u-iR)-\Phi_{C_{t-}}(t,T,u-iR)\Big) \\ &= S_{t-}^{R-iu}\Big(\sum_{k\neq C_{t-}}\Theta_{C_{t-},k}(-u-iR)\Phi_k(t,T,u-iR) \\ &\qquad -\Phi_{C_{t-}}(t,T,u-iR)\sum_{k\neq C_{t-}}\Theta_{C_{t-},k}(-u-iR)\Big)dt \\ &+ \sum_{k,j:j\neq k}S_{t-}^{R-iu}\Big(e^{(R-iu)\Psi^{j,k}}-1\Big)\Big(\Phi_k(t,T,u-iR)-\Phi_j(t,T,u-iR)\Big)dM_t^{j,k}. \end{aligned}$$

Substituting the above, (23) and (24) into (22) we obtain

$$\begin{aligned} &d(S_t^{R-iu}\Phi_{C_t}(t,T,u-iR)) = \\ &= S_{t-}^{R-iu}\Phi_{C_{t-}}(t,T,u-iR))\Big((R-iu)\langle\Sigma(C_{t-}),dW_t\rangle + \sum_{k,j\in\mathcal{K}:j\neq k}\big(e^{(R-iu)\Psi^{j,k}}-1\big)dM_t^{j,k} \\ &\qquad + \int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1\Big)\widetilde{\pi}(dt,dy)\Big) \\ &+ S_{t-}^{R-iu}\Phi_{C_{t-}}(t,T,u-iR))\Big[\mathcal{J}_1(-(u+iR)\Sigma(C_{t-})) + \sum_{k\neq C_{t-}}\Theta_{C_{t-},k}(-(u+iR)) \end{aligned}$$

$$
\begin{aligned}
&+(R-iu)\Big(\mathfrak{r}(C_{t-})-\mathcal{J}_1(-i\Sigma(C_{t-}))-\mathcal{J}_2(C_{t-})\Big)\Big]dt\\
&-S_{t-}^{R-iu}\Big(\Theta_{C_{t-},C_{t-}}(-u-iR)\Phi_{C_{t-}}(t;T,u+iR)+\widehat{f}^{e}(u+iR,C_{t-})\Big)dt\\
&-S_{t-}^{R-iu}\Big(\Phi_{C_{t-}}(t,T,u-iR)\sum_{k\neq C_{t-}}\Theta_{C_{t-},k}(-u-iR)\Big)dt\\
&\quad+\sum_{k,j\in\mathcal{K}:j\neq k}S_{t-}^{R-iu}\Big(1+e^{(R-iu)\Psi^{j,k}}-1\Big)\Big(\Phi_k(t,T,u-iR)-\Phi_j(t,T,u-iR)\Big)dM_t^{j,k}.
\end{aligned}
$$

This simplifies to

$$
\begin{aligned}
&d(S_t^{R-iu}\Phi_{C_t}(t,T,u-iR))\\
&=S_{t-}^{R-iu}\Phi_{C_{t-}}(t,T,u-iR))\Big((R-iu)\langle\Sigma(C_{t-}),dW_t\rangle+\int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{t-}),y\rangle}-1\Big)\widetilde{\pi}(dt,dy)\Big)\\
&\quad+\sum_{k,j\in\mathcal{K}:j\neq k}S_{t-}^{R-iu}\Big(\Phi_k(t,T,u-iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(t,T,u-iR)\Big)dM_t^{j,k}\\
&+S_{t-}^{R-iu}\Phi_{C_{t-}}(t,T,u-iR))\Big[\mathcal{J}_1(-(u+iR)\Sigma(C_{t-}))+(R-iu)\big(\mathfrak{r}(C_{t-})\\
&\qquad\qquad-\mathcal{J}_1(-i\Sigma(C_{t-}))-\mathcal{J}_2(C_{t-})\big)\Big]dt\\
&-S_{t-}^{R-iu}\Big(\Theta_{C_{t-},C_{t-}}(-u-iR)\Phi_{C_{t-}}(t;T,u+iR)+\widehat{f}^{e}(u+iR,C_{t-})\Big)dt.
\end{aligned}
$$

From the equality

$$
\begin{aligned}
&-\Theta_{C_{t-},C_{t-}}(-u-iR)\\
&\quad=\mathfrak{r}(C_{t-})-\mathcal{J}_1((-u-iR)\Sigma(C_{t-})-i(-u-iR)\Big(\mathfrak{r}(C_{t-})-\mathcal{J}_1(-i\Sigma(C_{t-}))-\mathcal{J}_2(C_{t-})\Big),
\end{aligned}
$$

we obtain the asserted formula (21). The proof is now complete. □

As we know from [13, Theorem 2.1] the crucial role for finding the risk minimizing strategy is played by the martingale M defined by

$$
M_t:=\mathbb{E}\big(\beta_T h(S_T,C_T)+\int_0^T\beta_v f(S_v,C_v)dv|\mathcal{F}_t\big),
$$

where f is given by (20). We need to find the martingale representation of M and to do this we use the stochastic Fubini theorem. So, we need to impose some integrability assumptions. Below we give two sets of assumptions which allow to apply stochastic Fubini theorem:

(A1): Let Φ be the unique solution of ODE (18) and assume that there exists a function $\theta:\mathbb{C}\to\mathbb{R}_+$ such that $\theta(\cdot+iR)\in L^1(\mathbb{R})$ and which, for every $t\in[0,T]$, satisfies

$$
\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\frac{|\Phi_{C_{v-}}(v,T,u+iR)|^2}{\theta(u+iR)}\beta_v^2\Big|S_{v-}^{(R-iu)}\Big|^2(R-iu)^2\langle\Sigma(C_{v-}),\Sigma(C_{v-})\rangle du\Big]dv<\infty,
$$

$$
\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\int_{\mathbb{R}^d}\frac{|\Phi_{C_{v-}}(v,T,u+iR)|^2}{\theta(u+iR)}\beta_v^2\Big|S_{v-}^{(R-iu)}\Big|^2\Big|e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\Big|^2\rho(dy)du\Big]dv<\infty,
$$

$$\sum_{k,j:j\neq k}\mathbb{E}\int_0^t\int_{\mathbb{R}}\Big[\frac{\left|\Phi_k(v,T,u+iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(v,T,u+iR)\right|^2}{\theta(u+iR)}\beta_v\left|S_{v-}^{(R-iu)}\right|^2H_{v-}^j\lambda^{j,k}\Big]dudv$$
$$<\infty.$$

The second set of assumptions uses the functions which appears in the explicit form of solution (18) given by (19).

(A2): Let ϕ be the unique solution of ODE (12) and let us denote by ϕ_j the j-th row vector of the matrix ϕ. Set

$$\gamma_{j,k}(v,s,u)=e^{-iu\Psi^{j,k}}\phi_k(v,s,u)-\phi_j(v,s,u).$$

Suppose that there exist functions $\theta_1:\mathbb{C}\to\mathbb{R}_+$ and $\theta_2:[0,T]\times\mathbb{C}\to\mathbb{R}_+$ such that $\theta_1(\cdot+iR)\in L^1(\mathbb{R})$ and $\theta_2(\cdot,\cdot+iR)\in L^1([0,T]\times\mathbb{R})$ and for every $t\in[0,T]$ the following integrability conditions hold

$$\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\frac{|\phi(v,T,u+iR)\widehat{h^e}(u+iR)|^2}{\theta_1(u+iR)}\beta_v^2\left|S_{v-}^{(R-iu)}\right|^2(R-iu)^2\langle\Sigma(C_{v-}),\Sigma(C_{v-})\rangle du\Big]dv<\infty,$$

$$\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\int_{\mathbb{R}^d}\frac{|\phi(v,T,u+iR)\widehat{h^e}(u+iR)|^2}{\theta_1(u+iR)}\beta_v^2\left|S_{v-}^{(R-iu)}\right|^2\left|e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\right|^2\rho(dy)du\Big]dv$$
$$<\infty,$$

$$\sum_{k,j\in\mathcal{K}:j\neq k}\mathbb{E}\int_0^t\int_{\mathbb{R}}\Big[\frac{\left|\gamma_{j,k}(v,T,u+iR)\widehat{h^e}(u+iR)\right|^2}{\theta_1(u+iR)}\beta_v\left|S_{v-}^{(R-iu)}\right|^2H_{v-}^j\lambda^{j,k}\Big]dudv<\infty.$$

and

$$\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\int_v^T\frac{|\phi(v,s,u+iR)\widehat{f^e}(u+iR)|^2}{\theta_2(s,u+iR)}\beta_v^2\left|S_{v-}^{(R-iu)}\right|^2(R-iu)^2\langle\Sigma(C_{v-}),\Sigma(C_{v-})\rangle dsdu\Big]dv$$
$$<\infty,$$

$$\mathbb{E}\int_0^t\Big[\int_{\mathbb{R}}\int_{\mathbb{R}^d}\int_v^T\frac{|\phi(v,s,u+iR)\widehat{f^e}(u+iR)|^2}{\theta_2(s,u+iR)}\beta_v^2\left|S_{v-}^{(R-iu)}\right|^2\left|e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\right|^2ds$$
$$\rho(dy)du\Big]dv<\infty,$$

$$\sum_{k,j\in\mathcal{K}:j\neq k}\mathbb{E}\int_0^t\int_{\mathbb{R}}\int_v^T\Big[\frac{\left|\gamma_{j,k}(v,s,u+iR)\widehat{f^e}(u+iR)\right|^2}{\theta_2(s,u+iR)}\beta_v\left|S_{v-}^{(R-iu)}\right|^2H_{v-}^j\lambda^{j,k}\Big]dsdudv<\infty.$$

The next theorem is the main result of the paper and generalize results of Tankov [18] obtained for exponential Lévy models. One can verify that in the case considered by Tankov [18] the set (A1) of assumptions are satisfied.

Theorem 7. *Let Φ be the unique solution of the ODE* (18) *with the function a given by* (20). *Suppose that (A1) or (A2) holds. Then the risk minimization strategy for D exists and is determined by the following investment in risky asset*

$$\varphi_s = \frac{1}{2\pi}\int_{\mathbb{R}} S_{s-}^{R-1-iu}\frac{\langle H_{s-},(G(-u-i(R+1))-G(-u-iR)-L)\Phi(s,T,u+iR)\rangle}{\langle H_{s-},(G(-2i)-2G(-i))\mathbb{1}\rangle}du + \frac{1}{S_{s-}}\frac{\langle Z^{\top}(S_{s-})H_{s-},G(-i)^{\top}H_{s-}\rangle}{\langle H_{s-},(G(-2i)-2G(-i))\mathbb{1}\rangle}$$

where $\mathbb{1}=(1,\dots,1)^{\top}\in\mathbb{R}^K$, $G(u)$, for $u\in\mathbb{C}$, is a matrix given by

$$G_{j,k}(u)=\begin{cases}\mathcal{J}_1(u\Sigma(j)) & j=k,\\ \lambda^{j,k}\left(e^{iu\Psi^{j,k}}-1\right) & j\neq k,\end{cases}$$

and

$$L=diag(G(-i)\mathbb{1}).$$

Proof. Let M be the martingale defined by

$$M_t:=\mathbb{E}\Big(\beta_T h(S_T,C_T)+\int_0^T\beta_v g(S_v,C_v)dv+\sum_{j,k:j\neq k}\int_0^T\beta_v Z^{j,k}(S_{v-})dH_v^{j,k}|\mathcal{F}_t\Big),$$

By [13, Theorem 2.1] the position in risky asset in the risk minimization strategy is given by

$$\varphi_v=\frac{d\langle\langle M,S\beta\rangle\rangle_v}{d\langle\langle S\beta\rangle\rangle_v},$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denotes angle bracket (predictable quadratic covariation). We note that using (2) and the definition of G one can verify in a standard way that

$$d\langle\langle S\beta\rangle\rangle_v=(S_{v-}\beta_v)^2\langle H_{v-},(G(-2i)-2G(-i))\mathbb{1}\rangle dv.$$

Now we compute $d\langle\langle M,S\beta\rangle\rangle_v$. Towards this end we will find the martingale representation of M. We start with rewriting the martingale M in the form

$$\begin{aligned}M_t&=\mathbb{E}\Big(\beta_T h(S_T,C_T)+\int_t^T\beta_v g(S_v,C_v)dv+\sum_{j,k\in\mathcal{K}:j\neq k}\int_t^T\beta_v Z^{j,k}(S_{v-})dH_v^{j,k}\Big|\mathcal{F}_t\Big)\\&\quad+\int_0^t\beta_v g(S_v,C_v)dv+\sum_{j,k\in\mathcal{K}:j\neq k}\int_0^t\beta_v Z^{j,k}(S_{v-})dH_v^{j,k}\\&=\beta_t\Big(\beta_t^{-1}\mathbb{E}\Big(\beta_T h(S_T,C_T)+\int_t^T\beta_v\Big(g(S_{v-},C_{v-})+\sum_{k\neq C_{v-}}Z^{C_{v-},k}(S_{v-})\lambda^{C_{v-},k}\Big)dv|\mathcal{F}_t\Big)\Big)\\&\quad+\int_0^t\beta_v\Big(g(S_{v-},C_{v-})+\sum_{k\neq C_{v-}}Z^{C_{v-},k}(S_{v-})\lambda^{C_{v-},k}\Big)dv+\sum_{j,k\in\mathcal{K}:j\neq k}\int_0^t\beta_v Z^{j,k}(S_{v-})dM_v^{j,k}\\&=\frac{1}{2\pi}\int_{\mathbb{R}}\Big[\beta_t S_t^{R-iu}\langle H_t,\Phi(t;T,u+iR)\rangle+\int_0^t\beta_v S_v^{R-iu}\widehat{f^e}(u+iR,C_{v-})dv\Big]du\\&\quad+\sum_{j,k\in\mathcal{K}:j\neq k}\int_0^t\beta_v Z^{j,k}(S_{v-})dM_v^{j,k},\end{aligned}\tag{25}$$

where in the third equality we have used (17) and the Fourier inversion formula. Now, using (21) and integration by parts we arrive at a formula for the dynamic of the first term under the integral in (25)

$$
\begin{aligned}
&d(\beta_v S_v^{R-iu}\Phi_{C_v}(v,T,u+iR))\\
&=\beta_v S_{v-}^{R-iu}\Phi_{C_{v-}}(v,T,u+iR))\Big[(R-iu)\langle\Sigma(C_{v-}),dW_v\rangle+\int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\Big)\widetilde{\pi}(dv,dy)\Big]\\
&\quad+\sum_{j,k\in\mathcal{K}:k\neq j}\beta_v S_{v-}^{R-iu}\Big(\Phi_k(v,T,u+iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(v,T,u+iR)\Big)dM_v^{j,k}\\
&\quad-\beta_v S_{v-}^{R-iu}\widehat{fe}(u+iR,C_{v-})dv.
\end{aligned}
$$

Hence (25) takes the form

$$
\begin{aligned}
M_t&=M_0+\frac{1}{2\pi}\int_{\mathbb{R}}\int_0^t\beta_v S_{v-}^{R-iu}\Phi_{C_{v-}}(v,T,u+iR))(R-iu)\langle\Sigma(C_{v-}),dW_v\rangle du\\
&+\frac{1}{2\pi}\int_{\mathbb{R}}\int_0^t\beta_v S_{v-}^{R-iu}\Phi_{C_{v-}}(v,T,u+iR))\int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\Big)\widetilde{\pi}(dv,dy)du\\
&+\frac{1}{2\pi}\int_{\mathbb{R}}\Big[\int_0^t\sum_{j,k:j\neq k}\beta_v S_{v-}^{R-iu}\Big(\Phi_k(v,T,u+iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(v,T,u+iR)\Big)dM_v^{j,k}\Big]du\\
&+\sum_{j,k\in\mathcal{K}:j\neq k}\int_0^t\beta_v Z^{j,k}(S_{v-})dM_v^{j,k}.
\end{aligned}
\tag{26}
$$

Now, using stochastic Fubini theorems (which is allowed under our assumptions) we obtain a martingale representation of M

$$
\begin{aligned}
M_t&=M_0+\int_0^t\langle\frac{1}{2\pi}\int_{\mathbb{R}}\beta_v S_{v-}^{R-iu}\Phi_{C_{v-}}(v,T,u+iR))(R-iu)\Sigma(C_{v-})du,dW_v\rangle\\
&+\int_0^t\int_{\mathbb{R}^n}\Big[\frac{1}{2\pi}\int_{\mathbb{R}}\beta_v S_{v-}^{R-iu}\Phi_{C_{v-}}(v,T,u+iR))\Big(e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\Big)du\Big]\widetilde{\pi}(dv,dy)\\
&+\int_0^t\sum_{j,k:j\neq k}\Big[\frac{1}{2\pi}\int_{\mathbb{R}}\beta_v S_{v-}^{R-iu}\Big(\Phi_k(v,T,u+iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(v,T,u+iR)\Big)du\Big]dM_v^{j,k}\\
&+\sum_{j,k\in\mathcal{K}:j\neq k}\int_0^t\beta_v Z^{j,k}(S_{v-})dM_v^{j,k}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\frac{d\langle\langle M,S\beta\rangle\rangle_v}{dv}=(\beta_v S_{v-})^2\frac{1}{2\pi}\Big(\int_{\mathbb{R}}S_{v-}^{R-1-iu}\Phi_{C_{v-}}(v,T,u+iR))\Big[(R-iu)\langle\Sigma(C_{v-}),\Sigma(C_{v-})\rangle\Big]du\\
&+\int_{\mathbb{R}}S_{v-}^{R-1-iu}\Phi_{C_{v-}}(v,T,u+iR))\Big[\int_{\mathbb{R}^n}\Big(e^{(R-iu)\langle\Sigma(C_{v-}),y\rangle}-1\Big)\Big(e^{\langle\Sigma(C_{v-}),y\rangle}-1\Big)\rho(dy)\Big]du\\
&+\int_{\mathbb{R}}S_{v-}^{R-1-iu}\Big[\sum_{j,k:j\neq k}\Big(\Phi_k(v,T,u+iR)e^{(R-iu)\Psi^{j,k}}-\Phi_j(v,T,u+iR)\Big)\Big(e^{\Psi^{j,k}}-1\Big)H_{v-}^j\lambda^{j,k}\Big]du\Big)\\
&+\sum_{j,k\in\mathcal{K}:j\neq k}\beta_v^2 S_{v-}Z^{j,k}(S_{v-})\Big(e^{\Psi^{j,k}}-1\Big)H_{v-}^j\lambda^{j,k}
\end{aligned}
$$

$$= (\beta_v S_{v-})^2 \Bigg(\frac{1}{2\pi} \int_{\mathbb{R}} S_{v-}^{R-1-iu} \Big[\sum_{k \in \mathcal{K}} \widetilde{G}_{C_{v-},k}(R-iu) \Phi_k(v,T,u+iR) \Big] du$$
$$+ \sum_{j,k:\, j \neq k} S_{v-}^{-1} Z^{j,k}(S_{v-}) \Big(e^{\Psi^{j,k}} - 1 \Big) H_{v-}^j \lambda^{j,k} \Bigg),$$

where

$$\widetilde{G}_{j,j}(R-iu) := (R-iu)\langle \Sigma(j), \Sigma(j) \rangle + \int_{\mathbb{R}^n} \Big(e^{\langle (R-iu)\langle \Sigma(j), y \rangle} - 1 \Big) \Big(e^{\langle \Sigma(j), y \rangle} - 1 \Big) \rho(dy)$$
$$- \sum_{k \in \mathcal{K}:\, k \neq j} \Big(e^{\Psi^{j,k}} - 1 \Big) \lambda^{j,k},$$

and for $j \neq k$

$$\widetilde{G}_{j,k}(R-iu) := e^{(R-iu)\Psi^{j,k}} \Big(e^{\Psi^{j,k}} - 1 \Big) \lambda^{j,k}.$$

Letting $Z^{j,j} = 0$ we can write

$$\sum_{j,k:\, j \neq k} S_{v-}^{-1} Z^{j,k}(S_{v-}) \Big(e^{\Psi^{j,k}} - 1 \Big) H_{v-}^j \lambda^{j,k} = \frac{1}{S_{s-}} \langle Z^\top(S_{s-}) H_{s-}, G(-i)^\top H_{s-} \rangle.$$

So it remains to show that

$$\widetilde{G}(R-iu) = G(-u-i(R+1)) - G(-u-iR) - diag(G(-i)\mathbb{1}).$$

One can easily verify that

$$\Big[(R-iu)\langle \Sigma(j), \Sigma(j) \rangle + \int_{\mathbb{R}_n} (e^{\Sigma(j)x} - 1)(e^{(R-iu)\Sigma(j)x} - 1) \nu(dx) \Big]$$
$$= \mathcal{J}_1((-u-i(R+1))\Sigma(j)) - \mathcal{J}_1((-u-iR)\Sigma(j)) - \mathcal{J}_1(-i\Sigma(j)).$$

This implies, for any $j \in \mathcal{K}$, that

$$\widetilde{G}_{j,j}(R-iu) = \mathcal{J}_1((-u-i(R+1))\Sigma(j)) - \mathcal{J}_1((-u-iR)\Sigma(j)) - \mathcal{J}_1(-i\Sigma(j)) - \mathcal{J}_2(j)$$
$$= G_{j,j}(-u-i(R+1)) - G_{j,j}(-u-iR) - (G(-i)\mathbb{1})_{j,j}.$$

Now let us consider off-diagonal elements of $\widetilde{G}$. We fix $j,k \in \mathcal{K}$, $j \neq k$ and note that

$$\widetilde{G}_{j,k}(R-iu) = \Big(e^{(R+1-iu)\Psi^{j,k}} - e^{(R-iu)\Psi^{j,k}} \Big) \lambda^{j,k}$$
$$= \Big(e^{(R+1-iu)\Psi^{j,k}} - 1 \Big) \lambda^{j,k} - \Big(e^{(R-iu)\Psi^{j,k}} - 1 \Big) \lambda^{j,k}$$
$$= G_{j,k}(-u-i(R+1)) - G_{j,k}(-u-iR) - (diag(G(-i)\mathbb{1}))_{j,k}.$$

The proof is now complete. □

References

[1] N. Bouleau and D. Lamberton. Residual risks and hedging strategies in Markovian markets. *Stochastic Process. Appl.*, 33(1):131–150, 1989.

[2] C. Ceci, A. Cretarola, and F. Russo. GKW representation theorem under restricted information. An application to risk-minimization. *Stoch. Dyn.*, 14(2):1350019, 23, 2014.

[3] K. Chourdakis. Switching Lévy models in continuous time: Finite distributions and option pricing. (2005) University of Essex, Centre for Computational Finance and Economic Agents (CCFEA) Working Paper. Available at SSRN: http://ssrn.com/abstract=838924 or http://dx.doi.org/10.2139/ssrn.838924.

[4] M. Dahl, M. Melchior, and T. Møller. On systematic mortality risk and risk-minimization with survivor swaps. *Scand. Actuar. J.*, (2-3):114–146, 2008.

[5] M. Dahl and T. Møller. Valuation and hedging of life insurance liabilities with systematic mortality risk. *Insurance Math. Econom.*, 39(2):193–217, 2006.

[6] R. J. Elliott and H. Föllmer. Orthogonal martingale representation. In *Stochastic analysis*, pages 139–152. Academic Press, Boston, MA, 1991.

[7] H. Föllmer and D. Sondermann. Hedging of nonredundant contingent claims. In *Contributions to mathematical economics*, pages 205–223. North-Holland, Amsterdam, 1986.

[8] F. Hubalek, J. Kallsen, and L. Krawczyk. Variance-optimal hedging for processes with stationary independent increments. *Ann. Appl. Probab.*, 16(2):853–885, 2006.

[9] J. Jakubowski and M. Niewęgłowski. Pricing and hedging of general rating-sensitive claims in a jump-diffusion market model in the presence of stochastic factors. *Journal of Mathematical Analysis and Applications*, 476(2):737–758, 2019.

[10] J. Jakubowski and M. Niewęgłowski. Pricing and hedging of rating-sensitive claims modeled by $\mathbb{F}$-doubly stochastic Markov chains. In *Advanced mathematical methods for finance*, pages 417–453. Springer, Heidelberg, 2011.

[11] Y. Kim, F. Fabozzi, Z. Lin, and S. Rachev. Option pricing and hedging under a stochastic volatility Lévy process model. *Review of Derivatives Research*, 15(1):81–97, 2012.

[12] A. Mijatović and M. Pistorius. Exotic derivatives under stochastic volatility models with jumps. In *Advanced mathematical methods for finance*, pages 455–508. Springer, Heidelberg, 2011.

[13] T. Møller. Risk-minimizing hedging strategies for insurance payment processes. *Finance Stoch.*, 5(4):419–446, 2001.

[14] R. Norberg. Optimal hedging of demographic risk in life insurance. *Finance Stoch.*, 17(1):197–222, 2013.

[15] R. Norberg. Quadratic hedging: an actuarial view extended to solvency control. *Eur. Actuar. J.*, 3(1):45–68, 2013.

[16] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Itô calculus, Reprint of the second (1994) edition.

[17] M. Schweizer. A guided tour through quadratic hedging approaches. In *Option pricing, interest rates and risk management*, Handb. Math. Finance, pages 538–574. Cambridge Univ. Press, Cambridge, 2001.

[18] P. Tankov. Pricing and hedging in exponential Lévy models: review of recent results. In *Paris-Princeton Lectures on Mathematical Finance 2010*, volume 2003 of *Lecture Notes in Math.*, pages 319–359. Springer, Berlin, 2011.

Anna Krasnosielska-Kobos[1], **Alicja Ochędzan**[2]

[1]Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
email: akrasno@mini.pw.edu.pl
[2] Roche Global IT Solution Centre, Warsaw, Poland
email: alicjaochedzan@gmail.com

HOW INFORMATION ABOUT DISORDER TIME AFFECTS STOPPING PROBLEM

Manuscript received: 1 June 2020
Manuscript accepted: 17 August 2020

Abstract: In the paper we present two optimal stopping problems with a change of structure of rewards at random time (called disorder time). The considered problems can be formulated as follows: a decision maker observes offers which appear at jump times of a Poisson process. The decision concerning the acceptance or the rejection of a presented offer is made at the moment of its appearance. Once rejected, the offer cannot be considered again. A reward for the decision maker is equal to the discounted value of the selected offer. The distribution of offers can change at random time. In the first problem we assume that the decision maker does not know if the disorder time has appeared or not. In the second one the decision maker knows it. The aim of the decision maker in both problems is to maximize the expected reward. In the paper, an explicit solution of both problems and their comparison is presented. The influence of the knowledge of the disorder time on the value of optimal mean reward is analysed.
Keywords: optimal stopping, Elfving problem, disorder time
Mathematics Subject Classification (2020): 60G40 (primary), 62L15

1. INTRODUCTION

The paper is inspired by the Elfving problem ([6], see also [4]). The Elfving problem is an optimal stopping problem of independent, identically distributed random variables observed sequentially at jump times of a Poisson process. The problem can be interpreted as follows: we want to sell some commodity (for example a car or a house). At random times we obtain offers, one at a time. The decision about the acceptance or the rejection of an offer must be made at the time of its appearance. Once rejected, the offer cannot be taken into consideration again. If we decide to accept the offer, we obtain a reward equal to the discounted value of

the offer. Our aim is to maximize the expected reward from the sale. Contrary to the Elfving problem, we do not assume that the offers have the same distribution, that is, we allow the distribution of the offers to change at random time. Such time is called the **disorder time**.

The Elfving problem was formulated and solved in [6] and analysed subsequently in [22], where an assumption was removed from the problem (see also [4]). Next, the problem was modified and extended in different directions. In [5] renewal process and random horizon with Gamma distribution was introduced. In [19] the assumption that rewards constitute a Markov chain was considered. Random horizon in the Elfving problem was introduced in [7] and in [12], where, additionally, a cost function was taken into consideration. Random starting time of decision process was introduced in [10]. Multiperson games with rewards the same as in the Elfving problem were presented in [8] and [16]. The allocation problem inspired by the Elfving problem was considered in [1] and [9]. The Elfving problem was generalized to multiple stopping problem in [23] and [24] and to multiple stopping problem with random horizon in [14].

An optimal stopping problem in which the distribution of offers changes at random time was introduced in [21]. The author considered the sequence of rewards $\{G_i\}_{i=1}^n$, $n < \infty$ where distribution of G_i changes at random time. The offers have a uniform distribution before and after the change but the distribution before the change stochastically dominates the distribution after the change. The goal of the decision is to maximize the expected reward at all stopping times τ with respect to $\{\mathcal{F}_i\}_{i=1}^n$, where $\mathcal{F}_i = \sigma(G_1, \ldots, G_i)$. The paper [21] was generalized in [17] to the case with imperfect information about the offers. Additionally, in [17] the asymptotic behaviour of the solution was analysed. One-stopping problem with reward structures as in the Elfving problem and a change of distribution of offers at random time was considered in [15]. In [15], at each time it is known if the disorder time has appeared or not.

In [2] two optimal stopping problems with random number of offers (random horizon) are considered. In the first one it is assumed that the decision maker has full information about the random horizon, i.e. at the beginning of the decision process he knows when the random horizon will appear (in fact he knows how many offers he will receive). In the second one, the decision maker knows only if the random horizon has already appeared or not. The author analysed the dependencies between the optimal expected rewards in these problems and compared the value of the optimal expected rewards for different parameters.

Lately, the detection problem (i.e. the problem in which the objective is to find a strategy which immediately detects a change of distribution) with Poisson process, also called the disorder problem, was considered in [3], [20] and [25]. In [25] an extensive bibliography on the detection problem was presented.

In this paper we present two optimal stopping problems allowing a change of distribution of offers in the Elfving problem: Problem A, in which we do not have the information about the time of the change of the distribution of offers and Problem B, in which we have some knowledge about the disorder time, i.e. we only known if the disorder time has already appeared or not. In both cases we present differential equations which allow us to calculate the optimal expected rewards in those problems. The numerical examples are also presented.

2. FORMULATION OF TWO PROBLEMS

Let $(\Omega, \mathcal{F}, P)$ denote the basic probability space on which all random objects are considered. Let $0 < T_1 < T_2 < \ldots$ be the jump times of a homogeneous Poisson process $N(t)$, $t \geq 0$, with intensity 1 and $T_0 = 0$. Moreover, let $\{Y_n^{(1)}\}_{n=1}^{\infty}$ and $\{Y_n^{(2)}\}_{n=1}^{\infty}$ be two sequences of independent, non-negative random variables (offers) with distribution functions F_i, $i = 1,2$, respectively. We assume that $\mu_i = E(Y_1^{(i)}) < \infty$ and $Y_0^{(i)} = 0$ for $i = 1,2$. Furthermore, there is a given discount function $r : [0,\infty) \to [0,1]$ such that r is right-continuous, non-increasing, $r(0) = 1$ and

$$\int_0^{\infty} r(s)ds < \infty.$$

Additionally, assume that M is a non-negative random variable (called disorder time) with distribution function F_M such that $E(M) < \infty$. We assume that the sequences $\{Y_n^{(1)}\}_{n=1}^{\infty}$, $\{Y_n^{(2)}\}_{n=1}^{\infty}$ and $\{T_n\}_{n=1}^{\infty}$ are independent and they are independent of M.

We will need some additional technical assumptions. We assume that one of the following three conditions is satisfied:

(*i*) $Y_1^{(i)}$ has a density f_i which is continuous on $\mathbb{R}$ except for a finite number of points, where $i = 1,2$.

(*ii*) $P(Y_1^{(1)} = 0) = 1$ and $Y_1^{(2)}$ has a density f_2 which is continuous on $\mathbb{R}$ except for a finite number of points.

(*iii*) $P(Y_1^{(2)} = 0) = 1$ and $Y_1^{(1)}$ has a density f_1 which is continuous on $\mathbb{R}$ except for a finite number of points.

Moreover, we assume that the functions r and F_M are discontinuous at most at a finite number of points. Let a set $\{s_0, \ldots, s_k\}$, where $0 = s_0 < s_1 < \ldots < s_{k-1} < s_k = U$, $k < \infty$ and $U = \sup\{s > 0 : r(s) > 0\}$, contain all points of discontinuity of function r and all points of discontinuity of function F_M on $[0,U)$ and points a and U_M, where $a = \sup\{x \geq 0 : F_M(x) = 0\}$, $U_M = \inf\{s \geq 0 : r_1(s) = 0\}$,

$$r_1(s) = r(s)\bar{F}_M(s),$$

$\bar{F}_M(s) = 1 - F_M(s)$. We will also use the following notation:

$$r_2(s) = r(s)F_M(s),$$

where $s \in [0,\infty)$.

Let us introduce the following σ-fields:

$$\mathcal{F}_n = \sigma(Y_1^{(1)}, \ldots, Y_n^{(1)}, Y_1^{(2)}, \ldots, Y_n^{(2)}, T_1, T_2, \ldots, T_n), \quad n \geq 1,$$

$\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_\infty = \sigma(\bigcup_{n\in\mathbb{N}_0} \mathcal{F}_n)$,

$$\mathcal{G}_n = \sigma(\mathcal{F}_n, \sigma(\mathbb{I}(M > T_0), \ldots, \mathbb{I}(M > T_n))), \quad n \geq 0,$$

$\mathcal{G}_\infty = \sigma(\bigcup_{n\in\mathbb{N}_0} \mathcal{G}_n)$, where $\mathbb{I}(A)$ is the indicator function of an event A. Let us also introduce two sets of all stopping times $\mathcal{M}$ and $\mathcal{M}^*$, with respect to filtrations $\{\mathcal{F}_n\}_{n=0}^\infty$ and $\{\mathcal{G}_n\}_{n=0}^\infty$, respectively. Let $\mathcal{M}_n = \{\tau \in \mathcal{M} : \tau \geq n\}$ and $\mathcal{M}_n^* = \{\tau \in \mathcal{M}^* : \tau \geq n\}$, $n \geq 1$. Note that the considered stopping times can be equal to infinity with positive probability.

Let

$$G_n = (Y_n^{(1)}\mathbb{I}(M > T_n) + Y_n^{(2)}\mathbb{I}(M \leq T_n))r(T_n), \quad n \geq 0,$$

$G_\infty = \limsup_{n\to\infty} G_n$. Random variable G_n is interpreted as a reward obtained by the decision maker if the nth offer is accepted. Note that if the disorder time has not appeared yet (i.e. $M > T_n$), then the value of the offer is equal to $Y_n^{(1)}$, otherwise the value of the offer is $Y_n^{(2)}$.

In this paper, we will consider two optimal stopping problems called Problem A and Problem B. In **Problem A** we are looking for an optimal stopping time $\tau_1 \in \mathcal{M}_1$ for the sequence $\{G_n\}_{n=0}^\infty$ and the optimal expected reward $E(G_{\tau_1})$ i.e.

$$E(G_{\tau_1}) = \sup_{\tau\in\mathcal{M}_1} E(G_\tau).$$

In **Problem B** we are looking for an optimal stopping time $\tau_1^* \in \mathcal{M}_1^*$ for the sequence $\{G_n\}_{n=0}^\infty$ and the optimal expected reward $E(G_{\tau_1^*})$ i.e.

$$E(G_{\tau_1^*}) = \sup_{\tau\in\mathcal{M}_1^*} E(G_\tau).$$

Note that we are looking for the solution of the above problems in different sets of stopping times.

The motivation to consider these two problems is the following observation: it is obvious that

$$\sup_{\tau\in\mathcal{M}_1} E(G_\tau) \leq \sup_{\tau\in\mathcal{M}_1^*} E(G_\tau). \tag{1}$$

Now, the question is: Can the above inequality be replaced by equality? It is obvious that the answer for this question is yes if $P(M = m) = 1$ for some $m \in \mathbb{R}$ or $P(M > U) = 1$. In [7] (see also [15] for an alternative proof) it was proven that the answer is also yes if $P(Y_1^{(2)} = 0) = 1$. Another question is how big is the difference between the values of the optimal expected rewards in these problems? We will present examples showing that the difference can be substantial.

Since the case of $P(M > U) = 1$ (which reduces it to the original Elfving problem) and the case of $P(Y_1^2 = 0) = 1$ (which is the Elfving problem with random horizon) are already completely solved (see [6] or [4] and [7] or [15]), we will assume from now on that $P(M < U) > 0$ and one of the assumptions (i) or (ii) from this section is satisfied. Moreover, if $P(Y_1^{(1)} = 0) = 1$, we put $f_1(s) = 0$ for $s \in \mathbb{R}$.

2.1. PROBLEM A

In this section, we present the solution of Problem A. The problem was formulated in [18] and is a generalization of the one considered in [11]. The method of solving this problem is based on [4, pp. 113-118].

For $u \geq 0$ define

$$\tilde{G}_n(u) = Y_n^{(1)} r_1(u+T_n) + Y_n^{(2)} r_2(u+T_n), \quad n \in \mathbb{N}_0,$$
$$\tilde{G}_\infty(u) = \limsup_{n\to\infty} \tilde{G}_n(u).$$

Let

$$\tilde{G}_n = \tilde{G}_n(0), \quad n \geq 0.$$

Let us note that for $u \geq 0$ we have $E(\sum_{n=0}^{\infty} \tilde{G}_n(u)) \leq (\mu_1+\mu_2)\int_0^\infty r(u+x)dx < \infty$, hence $\tilde{G}_\infty(u) = \lim_{n\to\infty} \tilde{G}_n(u) = 0$, and $E(\sup_{n\geq 0} \tilde{G}_n(u)) < \infty$. Moreover, note that $E(\sum_{n=0}^{\infty} G_n) \leq (\mu_1+\mu_2)\int_0^\infty r(x)dx < \infty$, hence $G_\infty = \lim_{n\to\infty} G_n = 0$, and $E(\sup_{n\geq 0} G_n) < \infty$. Hence, all considered expectations are well defined.

In the theorem below we will show that the optimal stopping problem of the sequence $\{G_n\}_{n=0}^\infty$ in the set of stopping times $\mathcal{M}_1$ can be replaced by the optimal stopping problem of the sequence $\{\tilde{G}_n\}_{n=0}^\infty$ in the set $\mathcal{M}_1$.

Theorem 1. *In the considered problem we have*

$$\sup_{\tau\in\mathcal{M}_1} E(G_\tau) = \sup_{\tau\in\mathcal{M}_1} E(\tilde{G}_\tau).$$

Proof. It is enough to show that for each $\tau \in \mathcal{M}_1$ we have $E(G_\tau) = E(\tilde{G}_\tau)$. Let $\tau \in \mathcal{M}_1$. Then

$$E(G_\tau) = E\big(\sum_{n=1}^{\infty} G_n \mathbb{I}(\tau = n)\big) + E(G_\infty \mathbb{I}(\tau = \infty)) = \sum_{n=1}^{\infty} E(G_n \mathbb{I}(\tau = n)) + E(G_\infty \mathbb{I}(\tau = \infty)).$$

Note that G_n is a function of $\mathcal{F}_n$-measurable random variables and a random variable M which is independent of $\mathcal{F}_n$. Hence $E(G_n \mathbb{I}(\tau = n)) = E(\tilde{G}_n \mathbb{I}(\tau = n))$ for $n \in \mathbb{N}$. Additionally, we have $G_\infty = 0$ and $\tilde{G}_\infty = \tilde{G}_\infty(0) = 0$. Hence, we get the assertion. □

To solve the optimal stopping problem for the reward sequence $\{\tilde{G}_n\}$ we will find an optimal stopping time $\tau_1(u) \in \mathcal{M}_1$ for the sequence $\{\tilde{G}_n(u)\}$ and the optimal expected reward $E(\tilde{G}_{\tau_1(u)}(u))$ i.e.

$$E(\tilde{G}_{\tau_1(u)}(u)) = \sup_{\tau\in\mathcal{M}_1} E(\tilde{G}_\tau(u)).$$

Introducing u in the considered problem allows us to derive a differential equation for the optimal expected reward.

Theorem 2. *There exists a Borel function $V : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ such that*

$$E(S_{n+1}(u) \mid \mathcal{F}_n) = V(u+T_n), \quad n \in \mathbb{N}_0,$$

where $S_n(u) = \operatorname{ess\,sup}_{\tau \in \mathcal{M}_n} E(\tilde{G}_\tau(u) \mid \mathcal{F}_n)$, $n \in \mathbb{N}_0$.

Proof. Note that $\{X_{n,u}, \mathcal{F}_n\}_{n=0}^{\infty}$, where $X_{n,u} = (Y_n^{(1)}, Y_n^{(2)}, u+T_n)$, $u \geq 0$, is a homogeneous Markov chain. Moreover, $\tilde{G}_n(u)$ is a function of $X_{n,u}$ only. Hence, using [4, Thms. 4.7 and 5.2] we get the assertion. □

In the theorem below we present the form of the optimal stopping time and the optimal expected reward for the sequence $\{\tilde{G}_n(u)\}_{n=0}^{\infty}$.

Proposition 3. *Let $u \geq 0$. Then:*

(*i*) *The stopping time*

$$\tau_1(u) = \inf\{n \geq 1 : \tilde{G}_n(u) \geq V(u+T_n)\} \tag{2}$$

is optimal in $\mathcal{M}_1$ for $\{\tilde{G}_n(u)\}_{n=0}^{\infty}$.

(*ii*)

$$\sup_{\tau \in \mathcal{M}_1} E(\tilde{G}_\tau(u)) = V(u). \tag{3}$$

Proof. The first part of the theorem follows from [4, Thm. 4.5', p. 82] and Theorem 2. The second one follows from [4, Thms 4.1 and 4.7] and Theorem 2. □

Before we formulate the theorem presenting the distribution function of the random variable $T_{\tau_1(u)}$, we prove some properties of the function $V(u)$.

Proposition 4. *$V(\cdot)$ is continuous on $[0,\infty)$.*

Proof. Let $u_1, u_2 \in [0,\infty)$. Using (3) we obtain

$$\begin{aligned}|V(u_1) - V(u_2)| &\leq \sup_{\tau \in \mathcal{M}_1} |E(\tilde{G}_\tau(u_1)) - E(\tilde{G}_\tau(u_2))| \\ &\leq \sup_{\tau \in \mathcal{M}_1} |E(Y_\tau^{(1)}(r_1(u_1+T_\tau) - r_1(u_2+T_\tau)))| + \sup_{\tau \in \mathcal{M}_1} |E(Y_\tau^{(2)}(r_2(u_1+T_\tau) - r_2(u_2+T_\tau)))|.\end{aligned}$$

For $i = 1,2$ we have

$$\begin{aligned}\sup_{\tau \in \mathcal{M}_1} |E(Y_\tau^{(i)}(r_i(u_1+T_\tau) - r_i(u_2+T_\tau)))| &\leq \sum_{n=1}^{\infty} E(Y_n^{(i)} |r_i(u_1+T_n) - r_i(u_2+T_n)|) \\ &= \mu_i \int_0^{\infty} |r_i(u_1+x) - r_i(u_2+x)| dx.\end{aligned}$$

From monotonicity and boundedness of r_1 we get

$$\int_0^\infty |r_1(u_1+x)-r_1(u_2+x)|dx \le |u_1-u_2|.$$

Additionally, from the definition of r_2 we obtain

$$\begin{aligned}&|r_2(u_1+x)-r_2(u_2+x)|\\ &\quad=|r(u_1+x)-r(u_2+x)+r(u_2+x)\bar{F}_M(u_2+x)-r(u_1+x)\bar{F}_M(u_1+x)|\\ &\quad\le|r(u_1+x)-r(u_2+x)|+|r_1(u_2+x)-r_1(u_1+x)|.\end{aligned}$$

Hence using monotonicity and boundedness of the functions r and r_1 we get

$$\int_0^\infty |r_2(u_1+x)-r_2(u_2+x)|dx \le 2|u_1-u_2|.$$

Therefore,

$$|V(u_1)-V(u_2)| \le (\mu_1+2\mu_2)|u_1-u_2|.$$

Thus, the Lipschitz condition is satisfied. □

Fact 5. *$V(s)>0$ for $s\in[0,U)$. Moreover, if $U<\infty$, then $V(s)=0$ for $s\ge U$; if $U=\infty$, then* $\lim_{s\to\infty}V(s)=0$.

Proof. Note that $V(s)\ge E(\tilde{G}_1(s))>0$ for $s\in[0,U)$. If $U<\infty$ and $s\ge U$, then $r(x)=0$ for $x\ge s$, hence $V(s)=0$ for $s\ge U$. The last part follows from the inequality $V(s)\le(\mu_1+\mu_2)\int_s^\infty r(x)dx$. □

To find $\tau_1(u)$ and $V(u)$, $u\in[0,U)$, we need to find the distribution of $T_{\tau_1(u)}$. Let

$$f_u(v)=P(T_{\tau_1(u)}>v),\quad v\in[0,\infty).$$

Fact 6. *The function f_u is continuous on $[0,\infty)$ and $f_u(0)=1$.*

For $x\ge 0$ define

$$g(x,t)=P(Y_1^{(1)}r_1(x)+Y_1^{(2)}r_2(x)<t).$$

Then, for $x\ge 0$ we have $g(x,t)=0$ for $t\le 0$. Moreover, for $x\ge 0$ and $t>0$,

$$g(x,t)=\begin{cases}F_1(\frac{t}{r_1(x)}), & x\in[0,a)\vee(x=a\wedge F_M(a)=0),\\ \int_0^\infty F_2\big(\frac{t-sr_1(x)}{r_2(x)}\big)f_1(s)ds+F_2\big(\frac{t}{r_2(x)}\big)P(Y_1^{(1)}=0), & x\in(a,U)\vee(x=a\wedge F_M(a)>0),\\ 1, & x\ge U.\end{cases}\tag{4}$$

Note that $0\le a<U$ from the assumption of the problem.

In the theorem below we present the form of the tail of distribution of the random variable $T_{\tau_1(u)}$.

Theorem 7. *For $u \in [0,U)$ the function $f_u(\cdot)$ has the following form:*

$$f_u(v) = \begin{cases} 1 & \text{for} \quad v < 0, \\ \exp\Big(-\int_u^{u+v} 1 - g(t,V(t))dt\Big) & \text{for} \quad v \in [0,a-u), \\ f_u(a-u)\exp\Big(-\int\limits_a^{u+v} 1 - g(t,V(t))dt\Big) & \text{for} \quad v \in [\max\{0,a-u\},U-u), \\ f_u(U-u)\exp(U-u-v) & \text{for} \quad v \in [U-u,\infty). \end{cases}$$

Proof. Before we start the main part of the proof, for $x_1,x_2 \geq 0$ and $t \in \mathbb{R}$ let $g_1(x_1,x_2,t) = P(Y_1^{(1)} r_1(x_1) + Y_1^{(2)} r_2(x_2) < t)$. Note that $g_1(x_1,x_2,t) = 0$ for $t \leq 0$ and $x_1,x_2 \geq 0$. Moreover, for $t > 0$ and $x_1,x_2 \geq 0$ we have

$$g_1(x_1,x_2,t) = \begin{cases} \int\limits_0^\infty F_2\big(\frac{t-sr_1(x_1)}{r_2(x_2)}\big) f_1(s)ds \\ +F_2\big(\frac{t}{r_2(x_2)}\big)P(Y_1^{(1)}=0), & x_1 \geq 0 \wedge (x_2 \in (a,U) \vee (x_2 = a \wedge F_M(a) > 0)), \\ F_1(\frac{t}{r_1(x_1)}), & x_1 < U_M \wedge (x_2 \in [0,a) \cup [U,\infty) \vee (x_2 = a \wedge F_M(a) = 0), \\ 1, & x_1 \geq U_M \wedge (x_2 \in [0,a) \cup [U,\infty) \vee (x_2 = a \wedge F_M(a) = 0)). \end{cases}$$

Now we can derive the formula for $f_u(v)$. Let $u \in [0,U)$. If $v < 0$, then $f_u(v) = 1$. Now assume that $v \geq 0$. For $h > 0$ we have

$$f_u(v+h) = P_1 + P_2 + P_3, \tag{5}$$

where

$$\begin{aligned} &P_1 = P(T_{\tau_1(u)} > v+h, \Delta N(v,v+h) = 0), \\ &P_2 = P(T_{\tau_1(u)} > v+h, \Delta N(v,v+h) = 1), \\ &0 \leq P_3 \leq P(\Delta N(v,v+h) \geq 2) = 1 - (1+h)\exp(-h) \end{aligned}$$

and $\Delta N(v,v+h) = N(v+h) - N(v)$. Let $\mathcal{H}_v = \sigma(N(s), s \leq v, Y_1^{(1)}, \ldots, Y_{N(v)}^{(1)}, Y_1^{(2)}, \ldots, Y_{N(v)}^{(2)})$. Then

$$\begin{aligned} P_1 &= P(T_{\tau_1(u)} > v, \Delta N(v,v+h) = 0) = P(\tau_1(u) > N(v), \Delta N(v,v+h) = 0) \\ &= E(P(\tau_1(u) > N(v), \Delta N(v,v+h) = 0 \mid \mathcal{H}_v)) \\ &= E(\mathbb{I}(\tau_1(u) > N(v)) P(\Delta N(v,v+h) = 0)) = \exp(-h) f_u(v). \end{aligned}$$

For $i = 1,2$ define x_i and $\tilde{x}_i$ such that $x_i \in (u+v, u+v+h]$, $r_i(x_i) = \inf_{x \in (0,h]} r_i(u+v+x)$, $\tilde{x}_i \in (u+v, u+v+h]$ and $r_i(\tilde{x}_i) = \sup_{x \in (0,h]} r_i(u+v+x)$. Then

$$\begin{aligned} P_2 &= P(T_{\tau_1(u)} > v, \Delta N(v,v+h) = 1, \tilde{G}_{N(v)+1}(u) < V(u + T_{N(v)+1})) \\ &\geq P(\tau_1(u) > N(v), \Delta N(v,v+h) = 1, Y_{N(v)+1}^{(1)} r_1(\tilde{x}_1) + Y_{N(v)+1}^{(2)} r_2(\tilde{x}_2) < \inf_{x \in (0,h]} V(u+v+x)) \\ &= g_1(\tilde{x}_1, \tilde{x}_2, \inf_{x \in (0,h]} V(u+v+x)) P(T_{\tau_1(u)} > v, \Delta N(v,v+h) = 1) \\ &= g_1(\tilde{x}_1, \tilde{x}_2, \inf_{x \in (0,h]} V(u+v+x)) h \exp(-h) f_u(v). \end{aligned}$$

Similarly, we get

$$P_2 \le g_1(x_1,x_2,\ \sup_{x\in(0,h]} V(u+v+x))h\exp(-h)f_u(v).$$

Hence,

$$\begin{aligned}
&\frac{\exp(-h)f_u(v)-f_u(v)+h\exp(-h)g_1(\tilde{x}_1,\tilde{x}_2,\inf_{x\in(0,h]}V(u+v+x))f_u(v)}{h}\\
&\le \frac{f_u(v+h)-f_u(v)}{h} \le\\
&\le \frac{\exp(-h)f_u(v)-f_u(v)+h\exp(-h)g_1(x_1,x_2,\sup_{x\in(0,h]}V(u+v+x))f_u(v)}{h}\\
&+\frac{1-\exp(-h)(1+h)}{h}.
\end{aligned}$$

Now, we will consider three cases. First, if $a>0$ and $0\le u+v<a$, then take $h>0$ such that $[u+v-h,u+v+h]\subset(s_i,s_{i+1})\cap[0,a)$ for some $i\in\{0,1,\ldots,k-1\}$. Note that $\inf_{x\in(0,h]}V(u+v+x)>0$ and $\sup_{x\in(0,h]}V(u+v+x)>0$. Hence, g_1 is continuous in the considered interval as a result of selecting x_i and $\tilde{x}_i$. Letting h converge to zero and using the observation $g(x,t)=g_1(x,x,t)$ for $x\ge 0$ and $t>0$ we get $f'_u(v^+)=f_u(v)(g(u+v,V(u+v))-1)$. Similarly, estimating $f_u(v)$ with $f_u(v-h)$ for $h>0$ we get $f'_u(v^-)=f_u(v)(g(u+v,V(u+v))-1)$. Hence,

$$f'_u(v)=f_u(v)(g(u+v,V(u+v))-1)$$

for $u+v\in(s_i,s_{i+1})\cap[0,a)$, $i\in\{0,1,\ldots,k-1\}$. Solving the above differential equation in each interval (s_i,s_{i+1}) and using the boundary condition $f_u(0)=1$ and continuity of f_u we get the assertion in this case.

Now, assume that $a\le u+v<U$. Take $h>0$ such that $[u+v-h,u+v+h]\subset(s_i,s_{i+1})\cap[a,U)$ for some $i\in\{0,1,\ldots,k-1\}$. As a result of the selection of u,v,h we obtain that g_1 is also continuous in this case. Hence, similarly to the previous case, we get $f'_u(v)=f_u(v)(g(u+v,V(u+v))-1)$ for $u+v\in(s_i,s_{i+1})\cap[a,U)$, $i\in\{0,1,\ldots,k-1\}$. If $a>0$, then using continuity of function f_u and the formula for $f_u(v)$ for $u+v\in[0,a)$ we get the boundary condition $f_u(a-u)$. If $a=0$, then the boundary condition becomes $f_u(a-u)=1$. Next, we solve the above differential equation with this boundary condition. Hence, using continuity of function f_u we get f_u for $a\le u+v<U$.

Now, let us consider the last case. Assume that $u+v\ge U$. Take $h>0$ such that $u+v-h\ge U$. Note that $T_{N(v)+1}>v$, so $u+T_{N(v)+1}>U$. Hence, $r_1(u+T_{N(v)+1})=0$ and $r_2(u+T_{N(v)+1})=0$. Therefore, $V(u+T_{N(v)+1})=0$ and $\tilde{G}_{N(v)+1}(u)=0$. Consequently, $P_2=0$. Hence, using (5), similarly to the first case, we get $f'_u(v)=-f_u(v)$. This differential equation is solved with the boundary condition $f_u(U-u)$ given in the second case. □

To find the function $V(u)$, we define a function $H(x,s)$ and prove two lemmas.

For $s>0$ and $x\in[0,\infty)$ define

$$H(x,s)=E((Y_1^{(1)}r_1(x)+Y_1^{(2)}r_2(x))\mathbb{I}(Y_1^{(1)}r_1(x)+Y_1^{(2)}r_2(x)\geq s))$$

and for $s\in\mathbb{R}$

$$H_i(s)=E(Y_1^{(i)}\mathbb{I}(Y_1^{(i)}\geq s)),\quad i=1,2.$$

Lemma 8. *Let $s>0$ and $x\in[0,\infty)$. Then*

$$H(x,s)=\begin{cases} r_1(x)H_1\big(\frac{s}{r_1(x)}\big), & x\in[0,a)\vee(x=a\wedge F_M(a)=0),\\ \int\limits_0^\infty f_1(y)\Big(r_2(x)H_2\big(\frac{s-yr_1(x)}{r_2(x)}\big)+yr_1(x)\bar F_2\big(\frac{s-yr_1(x)}{r_2(x)}\big)\Big)dy & \\ +r_2(x)P(Y_1^{(1)}=0)H_2\big(\frac{s}{r_2(x)}\big), & x\in(a,U)\vee(x=a\wedge F_M(a)>0),\\ 0, & x\geq U.\end{cases}$$

Proof. For $x\in[0,U)$ and $t>0$ we have

$$\frac{\partial g(x,t)}{\partial t}=\begin{cases} f_1\big(\frac{t}{r_1(x)}\big)\frac{1}{r_1(x)}, & x\in[0,a)\vee(x=a\wedge F_M(a)=0),\\ \int\limits_0^\infty f_2\big(\frac{t-sr_1(x)}{r_2(x)}\big)\frac{f_1(s)}{r_2(x)}ds+f_2\big(\frac{t}{r_2(x)}\big)\frac{P(Y_1^{(1)}=0)}{r_2(x)}, & x\in(a,U)\vee(x=a\wedge F_M(a)>0),\\ 0, & x\geq U.\end{cases}\tag{6}$$

Therefore,

$$H(x,s)=\int\limits_s^\infty t\cdot\frac{\partial g(x,t)}{\partial t}dt.$$

Hence, using (6) for $x<U$ and the fact that $r_1(x)=r_2(x)=0$ for $x\geq U$ we get the assertion. □

Lemma 9. *For $u\in[0,U)$ on the event $\{T_{\tau_1(u)}\in[0,U-u)\}$ we have*

$$E(\tilde G_{\tau_1(u)}(u)\mid\tau_1(u),T_{\tau_1(u)})=\frac{H(u+T_{\tau_1(u)},V(u+T_{\tau_1(u)}))}{1-g(u+T_{\tau_1(u)},V(u+T_{\tau_1(u)}))}.$$

Moreover, if $U<\infty$, then on the event $\{T_{\tau_1(u)}\in[U-u,\infty)\}$ we have $E(\tilde G_{\tau_1(u)}(u)\mid\tau_1(u),T_{\tau_1(u)})=0$.

Proof. First, we prove the first part of the lemma. Let $A=\{\tau_1(u)=k,T_k\in C\}$, where $C\in\mathcal{B}(\mathbb{R}_0^+)$, $k\in\mathbb{N}$. Let $D=C\cap[0,U-u)$ and $Z_{k-1}(u)=\{\tilde G_1(u)<V(u+T_1),\tilde G_2(u)<V(u+T_2),\ldots,\tilde G_{k-1}<V(u+T_{k-1})\}$. It is enough to show that

$$\int_A\tilde G_{\tau_1(u)}(u)\mathbb{I}(T_{\tau_1(u)}\in[0,U-u))dP=\int_A\frac{H(u,V(u+T_{\tau_1(u)}))}{1-g(u,V(u+T_{\tau_1(u)}))}\mathbb{I}(T_{\tau_1(u)}\in[0,U-u))dP.\tag{7}$$

Using (2), we get that the left hand side of (7) is equal to

$$\begin{aligned}
&E(\tilde{G}_k(u)\mathbb{I}(T_k \in D, Z_{k-1}(u), \tilde{G}_k(u) \geq V(u+T_k))) \\
&= E(E(\tilde{G}_k(u)\mathbb{I}(T_k \in D, Z_{k-1}(u), \tilde{G}_k(u) \geq V(u+T_k)) \mid \tilde{G}_1(u), ..., \tilde{G}_{k-1}(u), T_1, ..., T_k)) \\
&= E(\mathbb{I}(T_k \in D, Z_{k-1}(u))E(\tilde{G}_k(u)\mathbb{I}(\tilde{G}_k(u) \geq V(u+T_k)) \mid T_k)) \\
&= E(\mathbb{I}(T_k \in D, Z_{k-1}(u))H(u+T_k, V(u+T_k))).
\end{aligned}$$

Moreover, the right hand side of (7) is equal to

$$\begin{aligned}
&E\Big(\frac{H(u+T_k,V(u+T_k))}{1-g(u+T_k,V(u+T_k))}\mathbb{I}(T_k \in D, Z_{k-1}(u), \tilde{G}_k(u) \geq V(u+T_k))\Big) \\
&= E\Big(E\Big(\frac{H(u+T_k,V(u+T_k))}{1-g(u+T_k,V(u+T_k))}\mathbb{I}(T_k \in D, Z_{k-1}(u), \tilde{G}_k(u) \geq V(u+T_k)) \\
&\qquad\qquad \mid \tilde{G}_1(u), ..., \tilde{G}_{k-1}(u), T_1, ..., T_k\Big)\Big) \\
&= E\Big(\frac{H(u+T_k,V(u+T_k))}{1-g(u+T_k,V(u+T_k))}\mathbb{I}(T_k \in D, Z_{k-1}(u))E(\mathbb{I}(\tilde{G}_k(u) \geq V(u+T_k)) \mid T_k)\Big) \\
&= E(H(u+T_k,V(u+T_k))\mathbb{I}(T_k \in D, Z_{k-1}(u))).
\end{aligned}$$

Hence, we get (7).

The second part of the lemma follows from the assumption that $r(s) = 0$ for $s \geq U$. □

In the theorem below we present an integral equation satisfied by the function $V(u)$.

Theorem 10. *Function $V(u)$ for $u \in [0,U)$ satisfies the following equation:*

$$V(u) = \int_u^U H(v,V(v))f_u(v-u)dv. \tag{8}$$

Proof. Let $u \in [0,U)$. Using Proposition 3 and Lemma 9 we get

$$V(u) = \int_0^\infty \frac{H(u+v,V(u+v))}{1-g(u+v,V(u+v))} \cdot \frac{d(1-f_u(v))}{dv}dv.$$

Hence, using the definition of $H(x,s)$ and Fact 5 we obtain $H(u+v,V(u+v)) = 0$ for $v \geq U-u$. Therefore, from Theorem 7 we get the assertion. □

In the theorem below we show that the function $V(u)$ is uniquely determined by (8).

Theorem 11. *Let $\tilde{V}(u)$ satisfy (8) for $u \in [0,U)$, and if $U < \infty$, let $\tilde{V}(u) = 0$ for $u \geq U$. Then $\tilde{V}(u) = V(u)$ for $u \in [0,\infty)$.*

Proof. Let $\tilde{V}$ satisfy the assumptions of the theorem. Define $\tilde{\tau}(u) = \inf\{n \geq 1 : \tilde{G}_n(u) \geq \tilde{V}(u+T_n)\}$, $u \geq 0$. Following the methods in Theorems 7 and 10 we get

$$E(\tilde{G}_{\tilde{\tau}(u)}(u)) = \tilde{V}(u). \tag{9}$$

Note that $\mathbb{I}(\tilde{\tau}(u) \geq 2)E(\tilde{G}_{\tilde{\tau}(u)}(u) \mid \mathcal{F}_1) = \mathbb{I}(\tilde{G}_1(u) < \tilde{V}(u+T_1))\tilde{V}(u+T_1)$. Therefore,

$$\begin{aligned}\tilde{V}(u) &= E(\mathbb{I}(\tilde{\tau}(u) = 1)\tilde{G}_1(u) + \mathbb{I}(\tilde{\tau}(u) \geq 2)E(\tilde{G}_{\tilde{\tau}(u)}(u) \mid \mathcal{F}_1)) \\ &= E(\mathbb{I}(\tilde{G}_1(u) \geq \tilde{V}(u+T_1))\tilde{G}_1(u) + \mathbb{I}(\tilde{G}_1(u) < \tilde{V}(u+T_1))\tilde{V}(u+T_1)) \\ &= E(\max\{\tilde{G}_1(u), \tilde{V}(u+T_1)\}).\end{aligned}$$

Let $\tilde{S}(y_1, y_2, u) = \max\{y_1 r_1(u) + y_2 r_2(u), \tilde{V}(u)\}$ for $y_1, y_2, u \in [0, \infty)$. Hence, $E(\tilde{S}(Y_1^{(1)}, Y_1^{(2)}, u+T_1)) = \tilde{V}(u)$. Consequently,

$$E(\tilde{S}(Y_{n+1}^{(1)}, Y_{n+1}^{(2)}, T_{n+1}) \mid \mathcal{F}_n) = \tilde{V}(T_n).$$

Define $\tilde{S}_n = \tilde{S}(Y_n^{(1)}, Y_n^{(2)}, T_n)$, $n \in \mathbb{N}_0$, and $\tilde{S}_\infty = \limsup_{n\to\infty} \tilde{S}_n$. Then

$$\tilde{S}_n = \max\{\tilde{G}_n, \tilde{V}(T_n)\} \tag{10}$$

and

$$E(\tilde{S}_{n+1} \mid \mathcal{F}_n) = \tilde{V}(T_n). \tag{11}$$

From (10) we obtain $\tilde{S}_n \geq \tilde{G}_n$.

Note that $\tilde{V}(u) = 0$ for $u \geq U$ if $U < \infty$. Moreover, if $U = \infty$, then from (9) we have $\tilde{V}(u) \leq (\mu_1 + \mu_2) \int_u^\infty r(x)dx$. Hence, $\lim_{u\to\infty} \tilde{V}(u) = 0$. Therefore, for $U \in [0, \infty]$ we have $\lim_{n\to\infty} \tilde{V}(T_n) = 0$. Consequently, from the fact that $\lim_{n\to\infty} \tilde{G}_n = 0$ and from (10) we have $\tilde{S}_\infty = \lim_{n\to\infty} \tilde{S}_n = 0$. Additionally, we have $\tilde{V}(T_n) \leq (\mu_1 + \mu_2) \int_0^\infty r(T_n + x)dx$. Therefore, from (10) for $n \in \mathbb{N}_0$ we have

$$\tilde{S}_n \leq \tilde{G}_n + \tilde{V}(T_n) \leq \sum_{k=0}^{\infty} \tilde{G}_k + (\mu_1 + \mu_2) \int_0^\infty r(T_n + x)dx =: W. \tag{12}$$

Note that $E(W) < \infty$. Since $\tilde{S}_n$ is $\mathcal{F}_n$-measurable, using (12) we get $\tilde{S}_n \leq E(W \mid \mathcal{F}_n)$. Hence, from [4, Lem. 4.9]

$$\tilde{S}_n \leq S_n, \tag{13}$$

where $S_n = S_n(0)$, $n \geq 0$. Additionally, from [4, Thm. 4.6] it follows that $\{S_n, \mathcal{F}_n\}_{n=0}^\infty$ is the minimal supermartingale dominating $\{\tilde{G}_n, \mathcal{F}_n\}_{n=0}^\infty$. Moreover, from (10) and (11) we get that $\{\tilde{S}_n, \mathcal{F}_n\}_{n=0}^\infty$ is the supermartingale dominating $\{\tilde{G}_n, \mathcal{F}_n\}_{n=0}^\infty$. Hence, from (13) we obtain $\tilde{S}_n = S_n$. Therefore, from (11) and Theorem 2 we get

$$\tilde{V}(T_n) = V(T_n). \tag{14}$$

Now, we only need to show that $\tilde{V}(u) = V(u)$ for $u \in [0, U)$. Let $A = \{u \in [0, U) : \tilde{V}(u) \neq V(u)\}$. Assume that the Lebesgue measure $\mu(A) > 0$. Then $P(\{\omega : T_n(\omega) \in A\}) > 0$, which contradicts (14). Hence, using continuity of V and $\tilde{V}$ we get the assertion. Note that the continuity of $\tilde{V}$ follows from (8) and from the fact that the functions H and f_u are bounded. □

In the theorem below we show that the problem of finding the solution of integral equation (8) is equivalent to the problem of finding the solution of differential equation (15).

Theorem 12. *Let function $\tilde{V}(\cdot)$ be continuous on $[0,\infty)$. Moreover, let*

- $\tilde{V}(u) = 0$ *for* $u \geq U$, *if* $U < \infty$, *and*
- $\lim_{s\to\infty} \tilde{V}(s) = 0$, *if* $U = \infty$.

Then $\tilde{V}(u)$, $u \in [0,U)$, satisfies (8) if and only if

$$\frac{d}{du}\tilde{V}(u) = \tilde{V}(u)(1 - g(u,\tilde{V}(u))) - H(u,\tilde{V}(u)) \tag{15}$$

for $u \in (s_i, s_{i+1})$, $i \in \{0,1,\ldots,k-1\}$.

Proof. ($\Rightarrow$) It is enough to differentiate both sides of (8) with respect to u in each of the intervals (s_i, s_{i+1}). Then we see that $\tilde{V}$ satisfies (15) in the intervals (s_i, s_{i+1}), $i \in \{0,1,\ldots,k-1\}$.

($\Leftarrow$) Assume that $\tilde{V}$ satisfies the assumptions of the theorem and (15) in each of the intervals (s_i, s_{i+1}), $i \in \{0,1,\ldots,k-1\}$. Define $V_1(u)$ as follows:

$$V_1(u) = \int_u^U H(v,\tilde{V}(v))\tilde{f}_u(v-u)dv \quad \text{for} \quad u \in [0,U) \tag{16}$$

and if $U < \infty$, then $V_1(u) = 0$ for $u \geq U$, where

$$\tilde{f}_u(s) = \begin{cases} 1 & \text{for} \quad s < 0, \\ \exp\Big(-\int_u^{u+s} 1 - g(t,\tilde{V}(t))dt\Big) & \text{for} \quad s \in [0, a-u), \\ \tilde{f}_u(a-u)\exp\Big(-\int\limits_a^{u+s} 1 - g(t,\tilde{V}(t))dt\Big) & \text{for} \quad s \in [\max\{0,a-u\}, U-u), \\ \tilde{f}_u(U-u)\exp(U-u-s) & \text{for} \quad s \in [U-u,\infty). \end{cases}$$

If $U < \infty$, then $V_1(u) = V(u) = 0$ for $u \geq U$ (Fact 5). If $U = \infty$, then $\lim\limits_{s\to\infty} V_1(s) = \lim\limits_{s\to\infty} V(s) = 0$ (Fact 5). Hence, we need to show that $V_1(u) = \tilde{V}(u)$ for $u \in [0,U)$. Differentiating (16) in each of the intervals (s_i, s_{i+1}) we get

$$\frac{d}{du}V_1(u) = V_1(u)(1 - g(u,\tilde{V}(u))) - H(u,\tilde{V}(u)).$$

Hence,

$$\frac{d(V_1(u) - \tilde{V}(u))}{du} = (1 - g(u,\tilde{V}(u)))(V_1(u) - \tilde{V}(u)).$$

Assume that $V_1(u_0) \neq \tilde{V}(u_0)$ for some $u_0 \in (s_{k-1}, U)$. Let $u_1 = \inf\{s \in (u_0,U) : V_1(s) = \tilde{V}(s)\}$. If such u_1 does not exist, we take $u_1 = U$. Hence, for $u \in (u_0, u_1)$

$$\ln|V_1(u) - \tilde{V}(u)| = \ln|V_1(u_0) - \tilde{V}(u_0)| + \int_{u_0}^u (1 - g(t,\tilde{V}(t)))dt. \tag{17}$$

Note that $\lim_{u\to u_1} \ln|V_1(u) - \tilde{V}(u)| = -\infty$ while for $u \to u_1$ the right hand side of (17) is finite. From the contradiction we get $V_1(u) = \tilde{V}(u)$ for $u \in (s_{k-1}, U)$. Using continuity of functions V_1 and $\tilde{V}$ we get $V_1(u) = \tilde{V}(u)$ for $u \in [s_{k-1}, U]$. Using recursion we show that $V_1(u) = \tilde{V}(u)$ in each of the intervals $[s_i, s_{i+1}]$, $i = k-2,\ldots,0$. □

2.2. PROBLEM B

The presented problem can be found in [15] (see also [13]).

In this section we additionally assume that $U_M < \infty$. Moreover, we assume that if $Y_1^{(1)} = 0$, then F_M is increasing on (u_0, U_M), where $u_0 \in [0, U_M)$ or $U_M < U$. Let $\{\tilde{s}_0, \tilde{s}_1, \ldots, \tilde{s}_{\tilde{k}}\}$, where $0 = \tilde{s}_0 < \tilde{s}_1 < \ldots < \tilde{s}_{\tilde{k}-1} < \tilde{s}_{\tilde{k}} = U$, contain all points of discontinuity of the function r. Moreover, let $\{a_0, a_1, \ldots, a_l\}$, where $0 = a_0 < a_1 < \ldots < a_{l-1} < a_l = U_M$, contain all points of discontinuity of function r on $[0, U_M]$ and points of indifferentiability of F_M on $[0, U_M]$. We assume that $\tilde{k} < \infty$ and $l < \infty$.

Theorem 13 ([4]). *There exists exactly one function $V_2(\cdot)$ satisfying the following differential equation:*

$$\frac{d}{du}V_2(u) = \bar{F}_2(y_2(u))V_2(u) - r(u)H_2(y_2(u)), \tag{18}$$

in each of the intervals $(\tilde{s}_i, \tilde{s}_{i+1})$, $i \in \{0, \ldots, \tilde{k}-1\}$, such that $V_2(u)$ is continuous for $u \in [0, U]$ and

$$\lim_{u \to U} V_2(u) = 0,$$

where

$$y_2(u) = \begin{cases} \frac{V_2(u)}{r(u)}, & u < U, \\ 0, & u \geq U. \end{cases}$$

Note that $V_2(0)$ is the optimal expected reward for the sequence $\{Y_n^{(2)} r(T_n)\}_{n=1}^{\infty}$ in the original Elfving problem.

In all theorems below we assume that $V_2(\cdot)$ is as in Theorem 13.

In the first part of Theorem 14 we introduce a function $V_1(u,1)$ which is uniquely determined by differential equation (19). Both $V_1(u,1)$ and $V_2(u)$ are used in the second and the third part of the theorem to compute the optimal expected reward and the optimal stopping time for the sequence $\{G_n\}_{n=0}^{\infty}$ in the set of stopping times $\mathcal{M}_1^*$.

Theorem 14 ([15]). (*i*) *There exists exactly one function $V_1(\cdot,1)$ satisfying the following differential equation:*

$$\frac{d}{du}(\bar{F}_M(u)V_1(u,1)) + \frac{dF_M(u)}{du}V_2(u) = \bar{F}_M(u)(\bar{F}_1(y_1(u,1))V_1(u,1) - r(u)H_1(y_1(u,1))) \tag{19}$$

in each of the intervals (a_i, a_{i+1}), $i \in \{0, \ldots, l-1\}$, such that $\bar{F}_M(u)V_1(u,1) + F_M(u)V_2(u)$ is continuous for $u \in [0, U_M]$ and

$$\lim_{u \to U_M^-} \bar{F}_M(u)V_1(u,1) = V_2(U_M)P(M = U_M), \tag{20}$$

where for $u \in [0, U_M)$

$$y_1(u,1) = \frac{V_1(u,1)}{r(u)}.$$

(*ii*) *The optimal expected reward for the sequence* $\{G_n\}_{n=0}^{\infty}$ *has the form*

$$\sup_{\tau\in\mathcal{M}_1^*} E(G_\tau) = V_1(0,1)\bar{F}_M(0) + V_2(0)F_M(0).$$

(*iii*) *An optimal stopping time in* $\mathcal{M}_1^*$ *for the sequence of rewards* $\{G_n\}_{n=0}^{\infty}$ *has the form*

$$\tau_1^* = \inf\{n \geq 1 : (M > T_n, Y_n^{(1)} \geq y_1(T_n,1)) \text{ or } (M \leq T_n, Y_n^{(2)} \geq y_2(T_n))\}.$$

Note that the optimal stopping time in $\mathcal{M}_1^*$ for the sequence $\{G_n\}$ has the following interpretation: assume that we are at the time of the arrival of the nth offer. If the disorder time M has not appeared yet (i.e. $M > T_n$), then we stop if the value of the offer $Y_n^{(1)}$ is greater than or equal to $y_1(T_n,1)$; if the disorder time has already appeared (i.e. $M \leq T_n$), then we stop if the value of the offer $Y_n^{(2)}$ is greater than or equal to $y_2(T_n)$.

Let us present the solutions of two special cases of the problem. The first one is called the optimal stopping problem with random starting time.

Theorem 15 ([10]). *If* $P(Y_1^{(1)} = 0) = 1$, *then*

$$\sup_{\tau\in\mathcal{M}_1^*} E(G_\tau) = \int_0^U r(v)H_2(y_2(v))f_0(v)dv - \int_0^{U_M} r_1(v)H_2(y_2(v))dv,$$

where

$$f_0(v) = f_0^{(2)}(v) + \int_0^v \bar{F}_M(t)\bar{F}_2(y_2(t))f_t^{(2)}(v-t)dt$$

and

$$f_u^{(2)}(v) = \exp\Big(-\int_u^{u+v} \bar{F}_2(y_2(s))ds\Big).$$

Now, we will present the second special case of the problem, i.e. the case when M has a discrete distribution.

Theorem 16 ([15]). *Assume that* $r(s) = \mathbb{I}(s \in [0,U))$, $U < \infty$, *and* $P(M = a_n) = p_n$, *where* $0 \leq a_0 < a_1 < a_2 < \ldots$ *and* $\sum_{n=0}^{\infty} p_n = 1$. *Then* $V_1(u,1)$ *is uniquely determined by the following conditions:*

(*i*) *In each interval* (a_i, a_{i+1}), $i \in \{0,1,\ldots,l-1\}$, *the function* $V_1(\cdot,1)$ *satisfies the differential equation*

$$\frac{d}{du}V_1(u,1) = \bar{F}_1(V_1(u,1))V_1(u,1) - H_1(V_1(u,1)).$$

(*ii*) $\bar{F}_M(u)V_1(u,1) + F_M(u)V_2(u)$ *is continuous for* $u \in [0,U_M]$ *and* $\lim_{s\to U_M^-} V_1(s,1) = V_2(U_M)$.

3. EXAMPLES

In all the examples below we assume that

$$r(s) = \begin{cases} 1, & s \le 10, \\ 0, & s > 10. \end{cases}$$

Hence, $U = 10$. Moreover, we assume that $Y_1^{(2)}$ has the exponential distribution with parameter $\beta_2 > 0$. Then $F_2(x) = 1 - \exp(-\beta_2 x)$ for $x \ge 0$ and $F_2(x) = 0$ for $x < 0$. Moreover, $H_2(x) = (x + \frac{1}{\beta_2})\exp(-\beta_2 x)$ for $x \ge 0$ and $H_2(x) = \frac{1}{\beta_2}$ for $x < 0$. Hence, (18) has the form

$$\frac{dV_2(u)}{du} = -\frac{1}{\beta_2}\exp(-\beta_2 V_2(u)).$$

Solving the above differential equation with the boundary condition $V_2(U) = 0$ and using continuity of V_2 we get

$$V_2(u) = \frac{1}{\beta_2}\ln(1 + U - u),\ u \in [0, U]. \tag{21}$$

Note that $V_2(0)$ is the optimal expected reward for the sequence $\{Y_n^{(2)} r(T_n)\}_{n=0}^{\infty}$ in the original Elfving problem.

To simplify the notation in all the examples below we will write $\mathbb{A} = \sup_{\tau \in \mathcal{M}_1} E(G_\tau)$ and $\mathbb{B} = \sup_{\tau \in \mathcal{M}_1^*} E(G_\tau)$.

Example 17. *Let $Y_1^{(1)}$ have the exponential distribution with the parameter $\beta_1 > 0$. Moreover, let M have two point distribution, i.e. $P(M = m_1) = p$, $P(M = m_2) = q$, $q = 1 - p$, where $0 < m_1 < m_2 < U$ and $p \in (0,1)$. First, let us find the optimal expected reward in Problem A. Note that $a = m_1$. We consider two cases. First, assume that $q\beta_2 - p\beta_1 \ne 0$. Then, using the definition of r_1 and r_2, we get that for $u \in [0, U)$*

$$g(u, V(u)) = \begin{cases} F_1(V(u)), & 0 \le u < m_1, \\ \frac{1}{q\beta_2 - p\beta_1}\big(q\beta_2 F_1\big(\frac{V(u)}{q}\big) - p\beta_1 F_2\big(\frac{V(u)}{p}\big)\big), & m_1 \le u < m_2, \\ F_2(V(u)), & m_2 \le u < U \end{cases}$$

and

$$H(u, V(u)) = \begin{cases} H_1(V(u)), & 0 \le u < m_1, \\ \frac{1}{q\beta_2 - p\beta_1}\big(q^2\beta_2 H_1\big(\frac{V(u)}{q}\big) - p^2\beta_1 H_2\big(\frac{V(u)}{p}\big)\big), & m_1 \le u < m_2, \\ H_2(V(u)), & m_2 \le u < U. \end{cases}$$

Hence, if $q\beta_2 - p\beta_1 \ne 0$, then the function $V(u)$ satisfies

$$\frac{dV(u)}{du} = \begin{cases} -\frac{1}{\beta_1}\exp(-\beta_1 V(u)), & 0 < u < m_1, \\ \frac{1}{\beta_2 q - p\beta_1}\big(\frac{p^2\beta_1}{\beta_2}\exp\big(-\frac{\beta_2 V(u)}{p}\big) - \frac{q^2\beta_2}{\beta_1}\exp\big(-\frac{\beta_1 V(u)}{q}\big)\big), & m_1 < u < m_2, \\ -\frac{1}{\beta_2}\exp(-\beta_2 V(u)), & m_2 < u < U. \end{cases} \tag{22}$$

The above differential equation should be solved numerically in consecutive intervals: (m_2,U), (m_1,m_2), $(0,m_1)$. *Let* $u \in (m_2,U)$. *Solving the above differential equation we get* $V(u) = \frac{1}{\beta_2}\ln(1+U-u)$ *for* $u \in (m_2,U)$. *Now, using continuity of* V *we obtain* $V(m_2) = \frac{1}{\beta_2}\ln(1+U-m_2)$. *So we solve (22) for* $u \in (m_1,m_2)$ *with boundary condition* $V(m_2) = \frac{1}{\beta_2}\ln(1+U-m_2)$. *Hence, we get* $V(m_1)$, *which is the boundary condition for (22) when* $u \in (0,m_1)$. *Hence, for* $u \in [0,m_1]$ *we get*

$$V(u) = \frac{1}{\beta_1}\ln\Big(\exp(\beta_1 V(m_1)) + m_1 - u\Big).$$

Therefore, if $q\beta_2 - p\beta_1 \neq 0$, *then the optimal expected reward in Problem A is equal to* $\mathbb{A} = V(0) = \frac{1}{\beta_1}\ln(\exp(\beta_1 V(m_1)) + m_1)$.

Now, assume that $q\beta_2 - p\beta_1 = 0$. *Then*

$$g(u,V(u)) = \begin{cases} F_1(V(u)), & 0 \le u < m_1, \\ \frac{\beta_1 V(u)}{q}\bar{F}_2\big(\frac{V(u)}{p}\big), & m_1 \le u < m_2, \\ F_2(V(u)), & m_2 \le u < U \end{cases}$$

and

$$H(u,V(u)) = \begin{cases} H_1(V(u)), & 0 \le u < m_1, \\ \frac{\beta_1 V(u) p}{q} H_2\big(\frac{V(u)}{p}\big) + \frac{p}{\beta_2}\bar{F}_1\big(\frac{V(u)}{q}\big) + qH_1\big(\frac{V(u)}{q}\big), & m_1 \le u < m_2, \\ H_2(V(u)), & m_2 \le u < U. \end{cases}$$

Hence, if $q\beta_2 - p\beta_1 = 0$, *then the function* $V(u)$ *satisfies:*

$$\frac{dV(u)}{du} = \begin{cases} -\frac{1}{\beta_1}\exp(-\beta_1 V(u)), & 0 < u < m_1, \\ V(u) - \frac{\beta_1 V(u)}{q}\Big(2V(u) + \frac{p}{\beta_2}\Big)\bar{F}_2\big(\frac{V(u)}{p}\big) & \\ \qquad -\Big(\frac{p}{\beta_2} + \frac{q}{\beta_1} + V(u)\Big)\bar{F}_1\big(\frac{V(u)}{q}\big), & m_1 < u < m_2, \\ -\frac{1}{\beta_2}\exp(-\beta_2 V(u)), & m_2 < u < U. \end{cases}$$

The above differential equations can be solved in the same way as equation (22). Note that $\mathbb{A} = V(0)$.

Now, let us find the optimal expected reward in Problem B. Note that $U_M = m_2$. *From Theorem 16(i), we get that in each of the intervals* $(0,m_1)$ *and* (m_1,m_2) *the function* $V_1(\cdot,1)$ *satisfies*

$$\frac{dV_1(u,1)}{du} = -\frac{1}{\beta_1}\exp(-\beta_1 V_1(u,1)). \tag{23}$$

Solving the above differential equation with boundary condition $V_1(m_2,1) = V_2(m_2) = \frac{1}{\beta_2}\ln(1+U-m_2)$ *we get that for* $u \in (m_1,m_2)$

$$V_1(u,1) = \frac{1}{\beta_1}\ln\Big((1+U-m_2)^{\frac{\beta_1}{\beta_2}} + m_2 - u\Big).$$

Hence, using continuity of the function $\bar{F}_M(u)V_1(u,1)+F_M(u)V_2(u)$, we obtain the following boundary condition:

$$\lim_{u\to m_1^-} V_1(u,1) = \frac{q}{\beta_1}\ln\left((1+U-m_2)^{\frac{\beta_1}{\beta_2}}+m_2-m_1\right)+\frac{p}{\beta_2}\ln(1+U-m_1).$$

Solving (23) with the above boundary condition we get for $u\in(0,m_1)$

$$V_1(u,1) = \frac{1}{\beta_1}\ln\left(m_1+\left((1+U-m_2)^{\frac{\beta_1}{\beta_2}}+m_2-m_1\right)^q(1+U-m_1)^{\frac{p\beta_1}{\beta_2}}-u\right).$$

Hence, using Theorem 14 and the fact that $F_M(0)=0$, we get that the optimal expected reward in Problem B is equal to

$$\mathbb{B} = V_1(0,1) = \frac{1}{\beta_1}\ln\left(m_1+\left((1+U-m_2)^{\frac{\beta_1}{\beta_2}}+m_2-m_1\right)^q(1+U-m_1)^{\frac{p\beta_1}{\beta_2}}\right).$$

The numerical comparison of the optimal expected rewards in Problems A and B is presented in Table 1. We assume that $p=\frac{1}{2}$ in the comparison.

Table 1

Numerical results for Example 17

no.	β_1	β_2	m_1	m_2	μ_1	μ_2	Var(M)	$\mathbb{A}$	$\mathbb{B}$	$\frac{\mathbb{B}-\mathbb{A}}{\mathbb{A}}\cdot 100\%$
1	10^{-1}	1	1	9	10	1	16	14.2747	14.7692	3.46 %
2	10^{-1}	1	2	8	10	1	9	15.7129	16.0490	2.14 %
3	10^{-1}	1	3	7	10	1	4	16.8655	17.0796	1.27 %
4	10^{-1}	1	4	6	10	1	1	17.7459	17.8571	0.63%
5	1	1	1	9	1	1	16	1.2883	2.3979	86.14 %
6	1	1	2	8	1	1	9	1.5318	2.3979	56.54%
7	1	1	3	7	1	1	4	1.7279	2.3979	38.78%
8	1	1	4	6	1	1	1	1.9024	2.3979	26.04%
9	1	1	4.99	5.01	1	1	10^{-4}	2.3821	2.3979	0.66%
10	1	10^{-1}	0.1	9.9	1	10	21.34	8.6942	9.6407	10.89 %
11	1	10^{-1}	1	9	1	10	16	12.7818	14.9826	17.22%
12	1	10^{-1}	2	8	1	10	9	13.7577	16.4792	19.78%
13	1	10^{-1}	3	7	1	10	4	15.0886	17.3287	14.85%
14	1	10^{-1}	4	6	1	10	1	16.5202	17.7767	7.61%

Let us analyse the numerical results presented in Table 1. Note that in the table we chose m_1 and m_2 in such a way that $E(M)=5=\frac{1}{2}U$. First of all, note that the optimal expected rewards in Problems A and B are different and $\mathbb{A}<\mathbb{B}$ even for $\beta_1=\beta_2=1$ (lines 5-9). It means that the optimal expected reward in Problem A is smaller than the optimal expected reward in Problem B (for considered parameters), even if the distribution of offers does not change at the disorder time. Moreover, for $\beta_1=\beta_2=1$ we get that the optimal expected reward in the original Elfving problem (see (21)) is equal to $V_2(0)=\ln(11)\approx 2.3979=\mathbb{B}$, so it is also equal to the optimal expected reward in Problem B, while $\mathbb{A}<V_2(0)$. Now look at

the last column in the lines 10-14. We can see that there is no relation between the variance and percentage difference between the optimal expected rewards in both problems. We see that the optimal expected rewards in Problem A in lines 1-4 are greater than the optimal expected rewards in Problem A in lines 11-14 for the same m_1 and m_2. This observation suggests that in Problem A the case when we obtain offers with larger expectation ($\mu_1 = 10$) before the disorder time and offers with smaller expectation ($\mu_2 = 1$) after the disorder time is more profitable (gives larger optimal expected reward) than the case when we obtain offers with expectation $\mu_1 = 1$ before the disorder time and $\mu_2 = 10$ after the disorder time. Such a situation does not take place in Problem B (see lines 3 and 13). Finally, note that the difference in the optimal expected rewards in Problems A and B can be as big as 86% (see line 5), which shows that in general we should not approximate one problem by the other. Therefore, it is important to have the explicit solution of each of these problems. Such a substantial difference in the optimal expected rewards also shows how important the information of the disorder time is and how this information can change the optimal expected reward, for example from selling a commodity.

Example 18. *Let $Y_1^{(1)}$ be as in Example 17. Moreover, let M has the exponential distribution with parameter $\delta > 0$. Hence, $U = U_M$ and $a = 0$. First, note that for $u \geq 0$ we have $\beta_2 r_1(u) - \beta_1 r_2(u) = 0$ if $u = \frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1})$ or $u \geq U$. Note that $\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}) > 0$. Hence, using identity (4) for $u \in [0, \min\{\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}), U\})$ and $u \in (\min\{\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}), U\}, U)$ we obtain*

$$g(u,V(u)) = 1 - exp\Big(-\frac{\beta_1 V(u)}{r_1(u)}\Big) - \frac{\beta_1 r_2(u)}{\beta_2 r_1(u) - \beta_1 r_2(u)}\Big(\exp\Big(-\frac{\beta_1 V(u)}{r_1(u)}\Big) - \exp\Big(-\frac{\beta_2 V(u)}{r_2(u)}\Big)\Big)$$

and

$$H(u,V(u)) = \Bigg(\frac{r_2(u)}{\beta_2} + V(u) + \frac{r_1(u)}{\beta_1} + \Big(V(u) + \frac{r_2(u)}{\beta_2}\Big)\frac{\beta_1 r_2(u)}{\beta_2 r_1(u) - \beta_1 r_2(u)}\Bigg)\exp\Big(-\frac{\beta_1 V(u)}{r_1(u)}\Big) - \Big(V(u) + \frac{r_2(u)}{\beta_2}\Big)\frac{\beta_1 r_2(u)}{\beta_2 r_1(u) - \beta_1 r_2(u)}\exp\Big(-\frac{\beta_2 V(u)}{r_2(u)}\Big).$$

Therefore, using Theorem 12 we get that the function $V(u)$ for $u \in (0, \min\{\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}), U\})$ and $u \in (\min\{\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}), U\}, U)$ satisfies

$$\frac{dV(u)}{du} = -\Big(\frac{F_M(u)}{\beta_2} + \frac{\bar{F}_M(u)}{\beta_1} + \frac{\beta_1 (F_M(u))^2}{\beta_2(\beta_2 \bar{F}_M(u) - \beta_1 F_M(u))}\Big)\exp\Big(-\frac{\beta_1 V(u)}{\bar{F}_M(u)}\Big) + \frac{\beta_1 (F_M(u))^2}{\beta_2(\beta_2 \bar{F}_M(u) - \beta_1 F_M(u))}\exp\Big(-\frac{\beta_2 V(u)}{F_M(u)}\Big). \tag{24}$$

If $\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}) \geq U$ *we solve the above differential equation in the interval* $(0,U)$ *with the boundary condition* $V(U)=0$. *Otherwise, we solve the above differential equation in the interval* $(\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}),U)$ *with the boundary condition* $V(U)=0$. *Next, we use continuity of* $V(u)$ *at* $u=\frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1})$ *to find the solution of (24) for* $u \in [0, \frac{1}{\delta}\ln(\frac{\beta_1+\beta_2}{\beta_1}))$.

Now, we present the solution of Problem B. Note that using (19) we get that for $u \in (0,U)$, $V_1(u,1)$ *satisfies*

$$\frac{dV_1(u,1)}{du} = -\frac{1}{\beta_1}\exp(-\beta_1 V_1(u,1)) + \delta V_1(u,1) - \frac{\delta}{\beta_2}\ln(1+U-u).$$

We solve the above differential equation with the boundary condition given in (20). In our example, the boundary condition has the form: $V_1(U,1)=0$. *Moreover, note that* $F_M(0)=0$, *hence the optimal expected reward in Problem B is equal to* $\mathbb{B}=V_1(0,1)$.

The comparison of the optimal expected rewards in both problems is presented in Table 2. In the table, we chose δ *such that* $E(M)$ *is equal to* 1, 5 *and* 9.

Table 2

The numerical results for Example 18

β_1	β_2	δ	$\mathbb{A}$	$\mathbb{B}$	$\frac{\mathbb{B}-\mathbb{A}}{\mathbb{A}}\cdot 100\%$
10^{-1}	1	1	6.6320	7.2001	8.57%
10^{-1}	1	5^{-1}	13.6176	14.1051	3.58%
10^{-1}	1	9^{-1}	16.5869	16.9666	2.29%
1	1	1	2.2817	2.3979	5.09%
1	1	5^{-1}	2.0571	2.3979	16.57 %
1	1	9^{-1}	2.0806	2.3979	15.25%
1	10^{-1}	1	22.0662	22.9653	4.07%
1	10^{-1}	5^{-1}	14.7837	16.8374	13.89%
1	10^{-1}	9^{-1}	10.7414	12.5211	16.57%

The expected optimal rewards in both models are different, even for $\beta_1=\beta_2$. *More precisely, in the considered cases we have* $\mathbb{A}<\mathbb{B}$. *However, the difference is not as big as in Example 17.*

Example 19. ***Random starting time.*** *Let* $P(Y_1^{(1)}=0)=1$, $\beta_2=1$. *Additionally, assume that M has the exponential distribution with parameter* $\delta>0$. *Hence,* $a=0$ *and* $U_M=U$. *Using (4) we get that for* $u \in [0,U)$

$$g(u,V(u)) = F_2\Big(\frac{V(u)}{r_2(u)}\Big) = 1-\exp\Big(-\frac{V(u)}{1-\exp(-\delta u)}\Big).$$

Moreover, from Lemma 8 we obtain that for $u \in [0,U)$

$$\begin{aligned} H(u,V(u)) &= r_2(u)H_2\Big(\frac{V(u)}{r_2(u)}\Big) \\ &= (1-\exp(-\delta u))\Big(1+\frac{V(u)}{1-\exp(-\delta u)}\Big)\exp\Big(-\frac{V(u)}{1-\exp(-\delta u)}\Big). \end{aligned}$$

From (15) it follows that for $u \in (0,U)$ the function $V(u)$ satisfies the differential equation

$$\frac{d}{du}V(u) = -(1-\exp(-\delta u))\exp\Big(-\frac{V(u)}{1-\exp(-\delta u)}\Big)$$

with the boundary condition $V(U) = 0$. The differential equation should be solved numerically. Let us recall that the optimal expected reward in Problem A is equal to $\mathbb{A} = V(0)$.

To find the optimal expected reward in Problem B note that using Theorem 15 we obtain

$$f_u^{(2)}(v) = \frac{1+U-u-v}{1+U-u}.$$

Hence,

$$f_0(v) = \frac{1+U-v}{1+U} + (1+U-v)\int_0^v \frac{\exp(-\delta t)}{(1+U-t)^2}dt.$$

Therefore,

$$\mathbb{B} = \int_0^U (1+\ln(1+U-v))\Big(\frac{1}{1+U} + \int_0^v \frac{\exp(-\delta t)}{(1+U-t)^2}dt\Big)dv$$
$$- \int_0^U \frac{\exp(-\delta v)}{1+U-v}(1+\ln(1+U-v))dv.$$

The numerical comparison of the optimal expected rewards in Problems A and B is presented in Table 3.

Table 3

Numerical results for Example 19

δ	$\mathbb{A}$	$\mathbb{B}$	$\frac{\mathbb{B}-\mathbb{A}}{\mathbb{A}} \cdot 100\%$
10	2.3795	2.3887	0.39%
1	2.2037	2.2965	4.21%
5^{-1}	1.4543	1.6717	14.95 %
10^{-1}	0.9615	1.1366	18.17%
10^{-2}	0.1298	0.1574	21.32%
10^{-6}	0.000013	0.000016	21.77%

Note that the optimal expected rewards in both problems increase as δ increases (equivalently $E(M)$ decreases). Moreover, the larger δ, the smaller the percentage difference between the optimal expected rewards. In all considered cases we have $\mathbb{A} < \mathbb{B}$.

References

[1] A. S. Christian. 1974. *Optimal sequential assignments with random arrival times*, Management Science 21, 60–67.

[2] P. C. Allaart. 2004. *An application of prophet regions to optimal stopping with a random number of observations*, Optimization 53, 331–338.
[3] E. Bayraktar, S. Dayanik, I. Karatzas, 2005. *The standard Poisson disorder problem revisited*, Stochastic Process. Appl. 115, 1437–1450.
[4] Y. S. Chow, H. Robbins, D. Siegmund. 1971. *Great Expectations: The Theory of Optimal Stopping*, Houghton Mifflin, Boston 1971.
[5] I. David, U. Yechiali. 1985. *A time-dependent stopping problem with application to live organ transplants*, Oper. Res. 33, 491–504.
[6] G. Elfving. 1967. *A persistency problem connected with a point process*, J. Appl. Probab. 4, 77–89.
[7] E. Z. Ferenstein, A. Krasnosielska. 2009a. *A version of the Elfving optimal stopping time problem with random horizon*, Game Theory and Applications 14, Petrosjan L., Mazalov V. (eds.), Nova Science Publishers, NY, 40–53.
[8] E. Ferenstein, A. Krasnosielska. 2009b. *Nash equilibrium in a game version of Elfving problem*, Advances in Dynamic Games and Their Applications, Analytical and Numerical Developments, (eds. Pierre B., Gaitsgory V., Pourtallier O.), Birkhäuser, 399–414.
[9] A. Gershkov, B. Moldovanu. 2010. *Efficient sequential assignment with incomplete information*. Games Econom. Behav. 68, 144–154.
[10] A. Krasnosielska. 2009a. *A version of the Elfving problem with random starting time*, Stat. Probab. Lett. 79, 2429–2436.
[11] A. Krasnosielska. 2009b. *On some optimal stopping time problem with a change of the distribution*, Proceedings of 4th International PhD Students and Young Scientists Conference: Young Scientists Towards the Challenges of Modern Technology, Warsaw, 293–298.
[12] A. Krasnosielska. 2010. *A time dependent best choice problem with costs and random lifetime in organ transplants*, Appl. Math. 37, 257–274.
[13] A. Krasnosielska. 2011. *Zagadnienia optymalnego stopowania inspirowane problemem Elfvinga*. PhD thesis, Faculty of Mathematics and Information Science, Warsaw University of Technology.
[14] A. Krasnosielska-Kobos. 2015. *Multiple stopping problems with random horizon*, Optimization 64, 1625–1645.
[15] A. Krasnosielska-Kobos. 2020. *Optimal stopping problem with Poisson process and change of distribution of offers at random time*, In preparation.
[16] A. Krasnosielska-Kobos, E. Ferenstein. 2013. *Construction of Nash equilibrium in game version of Elfving's multiple stopping problem*, Dyn. Games Appl. 3, 220–235.
[17] V. Mazalov, E. Ivashko. 2012. *Bayes' model of the best-choice problem with disorder*, Int. J. Stoch. Anal. Article ID 697458, doi:10.1155/2012/697458.
[18] A. Ochędzan. 2012. *Wpływ informacji o momencie rozregulowania w zagadnieniach optymalnego stopowania*, MSc thesis, Faculty of Mathematics and Information Science, Warsaw University of Technology.
[19] M. Parlar, D. Perry, W. Stadje. 2007. *Optimal shopping when the sales are on - A Markovian full-information best-choice problem*, Stoch. Models 23, 351–371.
[20] G. Peskir, A. N. Shiryaev. 2002. *Solving the Poisson disorder problem*, Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann, Sandmann K., Schönbucher P. (eds.), Springer, Berlin, 295–312.
[21] M. Sakaguchi. 2001. *A best-choice problem for a production system which deteriorates at a disorder time*, Sci. Math. Jpn. 54, 125–134.
[22] D. O. Siegmund. 1967. *Some problems in the theory of optimal stopping*, Ann. Math. Statist. 38, 1627–1640.
[23] W. Stadje. 1987. *An optimal k-stopping problem for the Poisson process*, Proceedings of the 6th Pannonian Symposium on Mathematical Statistics, Bauer B. P., Konecny F., Wertz W. (eds.), 231–244.
[24] W. Stadje. 1990. *A full information pricing problem for the sale of several identical commodities*, ZOR - Methods Model. Oper. Res. 34, 161–181.
[25] K. Szajowski. 2011. *On a random number of disorders*, Probab. Math. Statist. 31, 17–45.

Agata Pilitowska, Anna Zamojska-Dzienio

Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

SEMILATTICE ORDERED ALGEBRAS WITH CONSTANTS

Manuscript received: 2 June 2020
Manuscript accepted: 17 July 2020

Abstract: We continue our studies on semilattice ordered algebras. This time we accept constants in the type of algebras. We investigate identities satisfied by such algebras and describe the free objects in varieties of semilattice ordered algebras with constants.
Keywords: free ordered structures, power algebras, idempotent semirings
Mathematics Subject Classification (2020): 06F05 (primary), 08B20, 08A30

1. INTRODUCTION

Ordered algebras such as Boolean algebras, Heyting algebras, lattice-ordered groups and MV-algebras played a decisive role in logic, both as the models of theories of first (or higher) order logic, as well as the algebraic semantics for the plethora of non-classical logics emerging in the twentieth century from linguistics, philosophy, mathematics, and computer science. For example, lattice-ordered groups play a fundamental role in the study of algebras of logic, while MV-algebras are the algebraic counterparts of the infinite-valued Łukasiewicz propositional logic.

Another important and widely investigated class is given by *quantales* [25, 26] - complete semilattices with an additional associative multiplication that distributes over arbitrary joins. They were introduced in the 1980s as a non-commutative generalization of *locales*, to capture the non-commutative logic arising in quantum mechanics. Quantales are examples of *semilattice ordered algebras* (*SLO algebras*, for short) which we discuss in this paper.

In the series of papers [17]–[21] we investigated SLO algebras $(A, \Omega, \leq)$, where $(A, \leq)$ is a (join) semilattice, (A, Ω) is an algebra (where Ω is a set of operations of any finitary positive arity, and, moreover, Ω is not necessarily finite) and each operation from Ω distributes over

the join. Obviously, examples are provided not only by already mentioned quantales or well known *additively idempotent semirings* [3, 7, 27]; SLO algebras are much more general structures. The basic role in the theory was played by extended power algebras of non-empty subsets and extended algebras of (non-empty) subalgebras. The main aim of the present paper is to describe the properties of SLO algebras with constants, i.e. we allow operations of the arity equal to zero (or in the case of power constructions we allow the empty subset and the empty subalgebra). We study the relations between the SLO algebras with the signatures including and excluding constants. Our motivation is very natural and came from applications in logic, where constants 0 and 1 play a significant role. Similar research in case of *commutative doubly-idempotent semirings* has been recently described in [1, 3].

The paper is organized as follows. In Section 2 we provide basic definitions, results and examples concerning semilattice ordered algebras with and without various types of constants in the signature. In Section 3 we investigate identities satisfied by SLO algebras and we present a necessary and sufficient condition for a SLO algebra to satisfy some non-linear identity. In Section 4 we describe the free objects in an arbitrary variety $\mathcal{S}$ of semilattice ordered algebras (with various types of constants in the signature) and in the quasivariety of Ω-subreducts of SLO algebras in $\mathcal{S}$. In Section 5 we apply the results to some particular idempotent varieties of SLO algebras.

2. SLO ALGEBRAS

Let $\mho$ be the variety of all algebras (A,Ω) of a (fixed) finitary type $\tau\colon \Omega \to \mathbb{N}^+$ and let $\mathcal{V} \subseteq \mho$ be a subvariety of $\mho$. In [20] we introduced the following definition of a semilattice ordered algebra.

Definition 1. *An algebra $(A,\Omega,+)$ is called a **semilattice ordered $\mathcal{V}$-algebra** (or briefly **semilattice ordered algebra**) if (A,Ω) belongs to a variety $\mathcal{V}$, $(A,+)$ is a (join) semilattice (with semilattice order $\leq$, i.e. $x \leq y \Leftrightarrow x+y=y$) and the operations from the set Ω* distribute *over the operation $+$, i.e. for each n-ary operation $\omega \in \Omega$, and $x_1,\ldots,x_i,y_i,\ldots,x_n \in A$*

$$\omega(x_1,\ldots,x_i+y_i,\ldots,x_n) = \omega(x_1,\ldots,x_i,\ldots,x_n) + \omega(x_1,\ldots,y_i,\ldots,x_n)$$

for any $1 \leq i \leq n$.

Definition 1 can be also formulated for semilattice ordered algebras with constants. Such constants may be of two types. The first one may consist of some special elements in the semilattice $(A,+)$ and the second one may refer to the algebra $(A,\Omega) \in \mathcal{V}$. In particular, we can consider semilattice algebras with neutral element with respect to the operation $+$ or with unit elements with respect to operations in Ω.

Definition 2. *An algebra* $(A,\Omega,+,0)$ *is called a* 0-***semilattice ordered*** $\mathcal{V}$***-algebra*** *if* $(A,\Omega,+)$ *is a semilattice ordered* $\mathcal{V}$*-algebra,* $(A,+,0)$ *is a semilattice with the least element* 0 *and for each* $\omega\in\Omega$ *and* $x_1,\dots,x_i,\dots,x_n\in A$

$$\omega(x_1,\dots,x_i,\dots,x_n)=0$$

whenever there is $1\leq i\leq n$ *such that* $x_i=0$.

Definition 3. *Let A be a non-empty set and let n be a positive integer. An element* $\alpha\in A$ *is called a* ***unit for an*** *n****-ary operation*** $\omega\colon A^n\to A$ *if for every* $x\in A$

$$\omega(x,\alpha,\dots,\alpha)=\omega(\alpha,x,\alpha,\dots,\alpha)=\dots=\omega(\alpha,\dots,\alpha,x)=x.$$

We say that α *is a* ***unit for an algebra*** (A,Ω) *if it is a unit for each operation* $\omega\in\Omega$.

Remark 4. *Let* (A,Ω) *be an algebra. Denote by* $\mathcal{E}$ *the set of all units for* (A,Ω). *If there is a binary operation in* Ω *then* $|\mathcal{E}|\leq 1$, *but in general* $0\leq|\mathcal{E}|\leq|A|$. *In this paper we assume that whenever a unit exists it is unique, i.e. we consider algebras* (A,Ω,α) *of a (fixed) finitary type* $\tau\colon\Omega\cup\{\alpha\}\to\mathbb{N}$ *with the unit* $\alpha\in\mathcal{E}$. *We denote this unique unit by* 1.

Let $\mho_1$ be the variety of all algebras $(A,\Omega,1)$ of a (fixed) finitary type $\tau\colon\Omega\cup\{1\}\to\mathbb{N}$ with the unit 1 and such that $(A,\Omega)\in\mho$ and let $\mathcal{V}_1\subseteq\mho_1$ be a subvariety of $\mho_1$.

Definition 5. *An algebra* $(A,\Omega,+,1)$ *is a* ***semilattice ordered*** $\mathcal{V}_1$***-algebra with a unit*** 1 *if* $(A,\Omega,+)$ *is a semilattice ordered* $\mathcal{V}$*-algebra and* 1 *is a unit for* (A,Ω).

Note that we do not assume that 1 is the greatest element in the semilattice $(A,+)$.

Definition 6. *An algebra* $(A,\Omega,+,0,1)$ *is a* 0-***semilattice ordered*** $\mathcal{V}_1$***-algebra with a unit*** *if* $(A,\Omega,+,0)$ *is a* 0*-semilattice ordered* $\mathcal{V}$*-algebra and* 1 *is a unit for* (A,Ω).

A direct consequence of distributivity is that in a semilattice ordered algebra $(A,\Omega,+)$ for each n-ary operation $\omega\in\Omega$ and $x_{ij}\in A$ for $1\leq i\leq n$, $1\leq j\leq r$ we have

$$\begin{aligned}\omega(x_{11},\dots,x_{n1})+\dots+\omega(x_{1r},\dots,x_{nr})\\ \leq\omega(x_{11}+\dots+x_{1r},\dots,x_{n1}+\dots+x_{nr}).\end{aligned}\tag{1}$$

It is also easy to notice that in semilattice ordered algebras all Ω-operations are *monotone* with respect to the semilattice order $\leq$. Namely if $x_i\leq y_i\in A$ for each $1\leq i\leq n$, then

$$\omega(x_1,\dots,x_n)\leq\omega(y_1,\dots,y_n).\tag{2}$$

This means that such algebras form a subclass of a class of ordered algebras in the sense of [4] (see also [5] and [2]). Basic examples are given by additively idempotent semirings, distributive lattices, semilattice ordered semigroups [6], semilattice ordered idempotent, entropic algebras (modals) [17], extended power algebras [20] or semilattice modes [13].

We start with a few natural examples of semilattice ordered algebras.

Example 7. Semilattice ordered semigroups. *An algebra $(A,\cdot,+)$, where $(A,\cdot)$ is a semigroup, $(A,+)$ is a semilattice and for any $a,b,c \in A$, $a\cdot(b+c) = a\cdot b + a\cdot c$ and $(a+b)\cdot c = a\cdot c + b\cdot c$ is a* ***semilattice ordered semigroup****. In particular, semirings with an idempotent additive reduct ([27], [16]),* ***dissemilattices*** *(called also* ***$\cdot$-distributive bisemilattices*** *in [15]) - algebras $(M,\cdot,+)$ with two semilattice structures $(M,\cdot)$ and $(M,+)$ in which the operation $\cdot$ distributes over the operation $+$, and distributive lattices are semilattice ordered $\mathcal{SG}$-algebras, where $\mathcal{SG}$ denotes the variety of all semigroups.*

Another important class here is given by ***quantales*** *[25, 26], i.e. semilattice ordered semigroups $(A,\cdot,+)$, where $(A,+)$ is a* ***complete*** *semilattice and the operation $\cdot$ distributes over arbitrary joins.*

Example 8. Extended power algebras of algebras. *[20] For a given set A denote by $\mathcal{P}A$ the family of all subsets of A and by $\mathcal{P}_{>0}A$ the family of all non-empty subsets of A. For any $0 \neq n$-ary operation $\omega\colon A^n \to A$ we define* ***the complex operation*** *$\omega\colon (\mathcal{P}A)^n \to \mathcal{P}A$ in the following way:*

$$\omega(A_1,\ldots,A_n) := \{\omega(a_1,\ldots,a_n) \mid a_i \in A_i\}, \tag{3}$$

where $\emptyset \neq A_1,\ldots,A_n \subseteq A$ and

$$\omega(A_1,\ldots,A_n) := \emptyset,$$

if there is $A_i = \emptyset$ for some $1 \leq i \leq n$.

The set $\mathcal{P}A$ also carries a join semilattice structure under the set-theoretical union $\cup$. In [11] B. Jónsson and A. Tarski proved that complex operations distribute over the union $\cup$. Hence, for any algebra $(A,\Omega) \in \mathcal{U}$, the ***extended power algebra*** *$(\mathcal{P}_{>0}A,\Omega,\cup)$ is a semilattice ordered $\mathcal{U}$-algebra and the* ***$\emptyset$-extended power algebra*** *$(\mathcal{P}A,\Omega,\cup,\emptyset)$ is a 0-semilattice ordered $\mathcal{U}$-algebra.*

Notice that the algebras $(\mathcal{P}^{<\omega}A,\Omega,\cup,\emptyset)$ and $(\mathcal{P}^{<\omega}_{>0}A,\Omega,\cup)$ of all finite (non-empty) subsets of A are subalgebras of $(\mathcal{P}A,\Omega,\cup,\emptyset)$ and $(\mathcal{P}_{>0}A,\Omega,\cup)$, respectively. Moreover, the power algebra of all subsets of A can also be viewed as *the Boolean algebra* $(\mathcal{P}A,\cup,\cap,-,A,\emptyset,\Omega)$ *with operators* Ω. This concept was introduced and studied by B. Jónsson and A. Tarski [11, 12].

Example 9. Extended power algebras of algebras with a unit. *Let $(A,\Omega,1)$ be an algebra with the unit $1 \in A$. It is clear that for any non-empty subset $X \subseteq A$ and n-ary operation $\omega \in \Omega$*

$$\omega(\{1\},\ldots,\underbrace{X}_{i},\ldots,\{1\}) = \{\omega(1,\ldots,x_i,\ldots,1) \mid x_i \in X\} = \{x_i \mid x_i \in X\} = X.$$

Then the algebra $(\mathcal{P}_{>0}A,\Omega,\cup,\{1\})$ is a semilattice ordered algebra with the unit $\{1\}$ and the algebra $(\mathcal{P}A,\Omega,\cup,\emptyset,\{1\})$ is a 0-semilattice ordered algebra with the unit $\{1\}$.

For a semigroup $(A,\cdot)$ its $\emptyset$-extended power algebra $(\mathcal{P}A,\cdot,\cup,\emptyset)$ is a basic example of a quantale. Similarly, the $\emptyset$-extended power algebra with a unit $(\mathcal{P}A,\cdot,\cup,\emptyset,\{1\})$ for a monoid $(A,\cdot,1)$ is a unital quantale.

An algebra (A,Ω) is ***idempotent*** if each singleton is a subalgebra, i.e. for every n-ary operation $\omega \in \Omega$ and $x \in A$ the following identity is satisfied:

$$\omega(x,\dots,x) = x.$$

A variety $\mathcal{V}$ of algebras is called *idempotent* if every algebra in $\mathcal{V}$ is idempotent.

An algebra (A,Ω) is ***entropic*** if any two of its operations commute. This property may also be expressed by means of identities: for every m-ary $\omega \in \Omega$ and n-ary $\varphi \in \Omega$ operations and $x_{11},\dots,x_{n1},\dots,x_{1m},\dots,x_{nm} \in A$

$$\begin{aligned}\omega(\varphi(x_{11},\dots,x_{n1}),\dots,\varphi(x_{1m},\dots,x_{nm})) = \\ \varphi(\omega(x_{11},\dots,x_{1m}),\dots,\omega(x_{n1},\dots,x_{nm})).\end{aligned}$$

Remark 10. *If there is a unit* 1 *in an entropic algebra* (A,Ω) *then the algebra is* ***symmetric****, i.e. for every n-ary operation* $\omega \in \Omega$ *and* $x_1,\dots,x_n \in A$ *the following identity holds:*

$$\omega(x_1,\dots,x_n) = \omega(x_{\pi(1)},\dots,x_{\pi(n)})$$

for each permutation π *of the set* $\{1,\dots,n\}$.

Example 11. Modals. *A* ***modal*** $(M,\Omega,+)$ *is a semilattice ordered algebra in which the algebra* (M,Ω) *is idempotent and entropic. Examples of modals include semilattice ordered semilattices (dissemilattices) and the algebra* $(\mathbb{R},\underline{I}^0,max)$ *defined on the set of real numbers, where* $\underline{I}^0$ *is the set of the binary operations:*

$$\underline{p} : \mathbb{R}\times\mathbb{R} \to \mathbb{R},\ (x,y) \mapsto (1-p)x + py,$$

for each $p \in (0,1) \subset \mathbb{R}$.

Idempotent and entropic algebras (called ***modes****) and also modals were introduced and investigated in detail by A. Romanowska and J.D.H. Smith ([22]-[24]). In particular, they showed that for a given idempotent and entropic algebra* (M,Ω)*, the sets* $S_{>0}(M)$ *of non-empty subalgebras and* $P_{>0}(M)$ *of finitely generated non-empty subalgebras under the complex operations* $\omega \in \Omega$ *and ordered by set-theoretic inclusion are modals. In the case we allow empty subalgebras, one obtains* 0*-semilattice ordered modes:* $(S(M),\Omega,\cup,\emptyset)$ *and* $(P(M),\Omega,\cup,\emptyset)$.

If a modal $(M,\Omega,+)$ *is entropic, then it is an example of a* ***semilattice mode****. Semilattice modes were described by K. Kearnes in [13].*

3. IDENTITIES IN SLO ALGEBRAS

As we will see in Section 4 the extended power algebras of algebras and their Ω-reducts play a special role in the context of semilattice ordered algebras. Let us recall some fundamental results referring to such algebras.

Let $\mathcal{V} \subseteq \mathcal{U}$ be a variety of algebras and let

$$\mathcal{V}\Sigma := \mathrm{HSP}(\{(\mathcal{P}A, \Omega) \mid (A, \Omega) \in \mathcal{V}\}) \text{ and}$$

$$\mathcal{V}\Sigma_{>0} := \mathrm{HSP}(\{(\mathcal{P}_{>0}A, \Omega) \mid (A, \Omega) \in \mathcal{V}\}).$$

Let us consider their subvarieties

$$\mathcal{V}\Sigma^{<\omega} := \mathrm{HSP}(\{(\mathcal{P}^{<\omega}A, \Omega) \mid (A, \Omega) \in \mathcal{V}\}) \text{ and}$$
$$\mathcal{V}\Sigma^{<\omega}_{>0} := \mathrm{HSP}(\{(\mathcal{P}^{<\omega}_{>0}A, \Omega) \mid (A, \Omega) \in \mathcal{V}\})$$

of power algebras of finite subsets.

We call a term t of the language of a variety $\mathcal{V}$ ***linear***, if every variable occurs in t at most once. An identity $t \approx u$ is called ***linear***, if both terms t and u are linear.

Note that the definition (3) of a complex operation extends to each linear derived operation t:

$$t(A_1, \ldots, A_n) := \{t(a_1, \ldots, a_n) \mid a_i \in A_i\}. \tag{4}$$

Each non-linear term t can be obtained from a linear one t^* by identification of some variables. Let $t^*(x_{11}, \ldots, x_{1k_1}, \ldots, x_{m1}, \ldots, x_{mk_m})$ be a linear term such that

$$t(x_1, \ldots, x_m) = t^*(\underbrace{x_1, \ldots, x_1}_{k_1-times}, \ldots, \underbrace{x_m, \ldots, x_m}_{k_m-times}).$$

Then for any subsets $A_1, \ldots, A_m$

$$\begin{aligned} \{t(a_1, \ldots, a_m) \mid a_i \in A_i\} &\subseteq t(A_1, \ldots, A_m) \\ &= \{t^*(a_{11}, \ldots, a_{1k_1}, \ldots, a_{m1}, \ldots, a_{mk_m}) \mid a_{ij} \in A_i\} \\ &= t^*(\underbrace{A_1, \ldots, A_1}_{k_1-times}, \ldots, \underbrace{A_m, \ldots, A_m}_{k_m-times}). \end{aligned}$$

G. Grätzer and H. Lakser proved in [8] that for any subvariety $\mathcal{V} \subseteq \mathcal{U}$ the following result holds.

Theorem 12 ([8, Theorem 1])**.** *Let $\mathcal{V}$ be a variety of algebras. The variety $\mathcal{V}\Sigma_{>0}$ satisfies precisely those identities which can be obtained from the linear identities true in $\mathcal{V}$ through identification of variables.*

Corollary 13 ([19])**.** *Let $\mathcal{V}$ be a variety of algebras. The varieties $\mathcal{V}\Sigma_{>0}$ and $\mathcal{V}\Sigma^{<\omega}_{>0}$ coincide.*

An identity $t \approx u$ is called ***regular*** if the set of variable symbols occurring in t equals the set of variable symbols occurring in u. The following results are analogues to the ones above formulated for varieties of power algebras including the empty set.

Theorem 14 ([8, Theorem 2])**.** *Let $\mathcal{V}$ be a variety of algebras. The variety $\mathcal{V}\Sigma$ satisfies precisely those regular identities which can be obtained from the linear identities true in $\mathcal{V}$ through identification of variables.*

Corollary 15. *Let $\mathcal{V}$ be a variety of algebras. The varieties $\mathcal{V}\Sigma$ and $\mathcal{V}\Sigma^{<\omega}$ coincide.*

Corollary 16 ([8, Corollary 2],[21, Theorem 4.6])**.** *Let $\mathcal{V}$ be a variety of algebras. Then $\mathcal{V} = \mathcal{V}\Sigma^{<\omega}_{>0}$ if and only if $\mathcal{V}$ is defined by a set of linear identities and $\mathcal{V} = \mathcal{V}\Sigma^{<\omega}$ if and only if $\mathcal{V}$ is defined by a set of linear regular identities.*

Let (A,Γ) be an algebra of a given type $\tau : \Gamma \to \mathbb{N}$. Denote by $\mathfrak{B}\Gamma$ a set of derived (or term) operations of Γ and let $\Omega \subseteq \mathfrak{B}\Gamma$. An algebra (A,Ω) is said to be a *reduct* (Ω-*reduct*) of the algebra (A,Γ). A subalgebra of a reduct of (A,Γ) is called a *subreduct*.

Let $(A,\Omega,+)$ be a semilattice ordered algebra generated by a set $X \subseteq A$. Denote by $(\langle X\rangle_\Omega,\Omega)$ the subalgebra of the Ω-reduct (A,Ω) generated by the set X. The algebra $(\langle X\rangle_\Omega,\Omega)$ contains all elements from $(A,\Omega,+)$ obtained as results of *derived* (or *term*) operations from Ω on the set X. We will call it the *full Ω-algebra subreduct* (*of a semilattice ordered algebra* $(A,\Omega,+)$) *relative to* X.

An element $r \in A$ is said to be in ***disjunctive form*** if it is a join of a finite number of elements from their full Ω-subreduct $\langle X\rangle_\Omega$.

The following theorem shows that each element in a semilattice ordered algebra may be expressed in such form.

Lemma 17 (Disjunctive Form Lemma)**.** *Let $(A,\Omega,+)$ be a semilattice ordered algebra generated by a set $X \subseteq A$. For each $r \in A$, there exist $r_1,\ldots,r_p \in \langle X\rangle_\Omega$ such that*

$$r = r_1 + \ldots + r_p.$$

Proof. The proof is done by induction on the minimal number m of occurrences of the semilattice operation $+$ in the expression of r as a semilattice ordered algebra word in the alphabet X.

Consider $r = r_1$ with $r_1 \in \langle X\rangle_\Omega$. Hence, the result holds for $m = 0$.

Now suppose that the hypothesis is established for $m > 0$ and let $r \in A$ be an element in which the semilattice operation $+$ occurs $m+1$ times. Let $r = r_1 + r_2$, for some $r_1, r_2 \in A$. By induction hypothesis there are $r_{11},\ldots,r_{1k},r_{21},\ldots,r_{2n} \in \langle X\rangle_\Omega$ such that

$$r = r_1 + r_2 = r_{11} + \ldots + r_{1k} + r_{21} + \ldots + r_{2n}.$$

Otherwise, $r = \omega(r_1,\ldots,r_k + s_k,\ldots,r_n)$ for some $\omega \in \Omega$ and $r_1,\ldots,r_k,\ldots,r_n,s_k \in A$. Then, by distributivity we have

$$r = \omega(r_1,\ldots,r_k + s_k,\ldots,r_n) = \omega(r_1,\ldots,r_k,\ldots,r_n) + \omega(r_1,\ldots,s_k,\ldots,r_n).$$

Because $\omega(r_1,\ldots,r_k,\ldots,r_n), \omega(r_1,\ldots,s_k,\ldots,r_n) \in A$, this completes the proof. □

Corollary 18. *Let $(A,\Omega,+)$ be a semilattice ordered algebra generated by a set $X \subseteq A$. There is a set $Y \subseteq A$ of generators of the semilattice $(A,+)$ such that $Y \subseteq \langle X\rangle_\Omega$.*

Let $\mathcal{SV}$ denote the class of all semilattice ordered algebras such that for each $(A,\Omega,+)\in\mathcal{SV}$ there exists a set of generators such that their full Ω-subreduct lies in $\mathcal{V}$.

Theorem 19. *Let $\mathcal{V}$ be a variety of Ω-algebras satisfying an identity $t\approx u$ for some $0\neq$ n-ary terms and let $\mathcal{S}\subseteq\mathcal{SV}$ be a variety of semilattice ordered algebras $(A,\Omega,+)$ such that the word operation $t:A^n\to A$ distributes over the operation $+$.*

Then the identity $t\approx u$ is satisfied in $\mathcal{S}$ if and only if the word operation $u:A^n\to A$ distributes over the operation $+$.

Proof. Let $(A,\Omega,+)\in\mathcal{S}\subseteq\mathcal{SV}$ and let the word operation $t:A^n\to A$ distribute over the operation $+$.

Because the variety $\mathcal{S}$ is, by assumption, included in $\mathcal{SV}$, there exists a set X of generators of $(A,\Omega,+)$ such that its full Ω-algebra subreduct relative to X belongs to the variety $\mathcal{V}$. Hence, the identity $t\approx u$ is also true in $(\langle X\rangle_\Omega,\Omega)$.

First, suppose that the word operation $u:A^n\to A$ distributes over the operation $+$ and let $r_1,\ldots,r_n\in A$. By the Disjunction Form Lemma 17 there exist $r_{11},\ldots,r_{1k_1},\ldots,r_{n1},\ldots,r_{nk_n}\in\langle X\rangle_\Omega$ such that for each $1\leq i\leq n$, $r_i=r_{i1}+\ldots+r_{ik_i}$. Then, by distributivity of operations $t:A^n\to A$ and $u:A^n\to A$ we obtain

$$t(r_1,\ldots,r_n)=t(r_{11}+\ldots+r_{1k_1},\ldots,r_{n1}+\ldots+r_{nk_n})=$$
$$\sum_{\substack{1\leq i\leq n\\ a_i\in\{r_{i1},\ldots,r_{ik_i}\}}} t(a_1,\ldots,a_n)=\sum_{\substack{1\leq i\leq n\\ a_i\in\{r_{i1},\ldots,r_{ik_i}\}}} u(a_1,\ldots,a_n)=$$
$$u(r_{11}+\ldots+r_{1k_1},\ldots,r_{n1}+\ldots+r_{nk_n})=u(r_1,\ldots,r_n).$$

The converse implication is obvious. □

Example 20. *Let $(A,\Omega,+)$ be a semilattice ordered algebra and let $\omega\in\Omega$ be an n-ary operation. The unary operation $t(x):=\omega(x,\ldots,x):A\to A$ distributes over the operation $+$ if and only if for any $x,y\in A$*

$$t(x)+t(y)=\sum_{x_i\in\{x,y\}}\omega(x_1,\ldots,x_n).$$

In particular, if $\omega\in\Omega$ is a binary idempotent operation then the operation $t(x)=\omega(x,x):A\to A$ distributes over the operation $+$ if and only if for any $x,y\in A$

$$x+y=x+y+\omega(x,y)+\omega(y,x).$$

Lemma 21. *Let $(A,\Omega,+)$ be a semilattice ordered algebra and let t be an $0\neq$ n-ary linear Ω-term. Then the word operation $t:A^n\to A$ distributes over the operation $+$.*

Proof. The proof is done by induction on the minimal number m of occurrences of (symbols of) the basic Ω-operations in the corresponding linear Ω-term.

By definition of a semilattice ordered algebra, the lemma is certainly true for $m = 1$. Now suppose that the hypothesis is established for $m > 1$. Let

$$t(x_{11},\dots,x_{kp_k}) = \omega(\nu_1(x_{11},\dots,x_{1p_1}),\dots,\nu_k(x_{k1},\dots,x_{kp_k}))$$

be a linear Ω-term, for some $\omega \in \Omega$, different variable symbols $x_{11},\dots,x_{1p_1},\dots,x_{k1},\dots,x_{kp_k}$ and linear Ω-words $\nu_1,\dots,\nu_k$, in which the basic Ω-operations occur $m+1$ times.

By the induction hypothesis, the Ω-word operations $\nu_i : A^{p_i} \to A$, for $1 \le i \le k$, distribute over the operation $+$. This implies that for any $x_{11},\dots,x_{1p_1},\dots,x_{i1},\dots,x_{ij},y_{ij},\dots,x_{ip_i}$, $\dots,x_{k1},\dots,x_{kp_k} \in A$,

$$\begin{aligned}
&t(x_{11},\dots,x_{ij}+y_{ij},\dots,x_{kp_k})\\
&\quad= \omega(\nu_1(x_{11},\dots,x_{1p_1}),\dots,\nu_i(x_{i1},\dots,x_{ij}+y_{ij},\dots,x_{ip_i}),\dots,\nu_k(x_{k1},\dots,x_{kp_k}))\\
&\quad= \omega(\nu_1(x_{11},\dots,x_{1p_1}),\dots,\nu_i(x_{i1},\dots,x_{ij},\dots,x_{ip_i})\\
&\qquad+ \nu_i(x_{i1},\dots,y_{ij},\dots,x_{ip_i}),\dots,\nu_k(x_{k1},\dots,x_{kp_k}))\\
&\quad= \omega(\nu_1(x_{11},\dots,x_{1p_1}),\dots,\nu_i(x_{i1},\dots,x_{ij},\dots,x_{ip_i}),\dots,\nu_k(x_{k1},\dots,x_{kp_k}))\\
&\qquad+ \omega(\nu_1(x_{11},\dots,x_{1p_1}),\dots,\nu_i(x_{i1},\dots,y_{ij},\dots,x_{ip_i}),\dots,\nu_k(x_{k1},\dots,x_{kp_k}))\\
&\quad= t(x_{11},\dots,x_{ij},\dots,x_{kp_k}) + t(x_{11},\dots,y_{ij},\dots,x_{kp_k}),
\end{aligned}$$

which completes the proof. □

Corollary 22. *Let $\mathcal{V}$ be a variety of Ω-algebras satisfying an identity $t \approx u$ for some $0 \neq n$-ary terms, where t is linear. The identity $t \approx u$ is true in a variety $\mathcal{S} \subseteq \mathcal{SV}$ of semilattice ordered algebras if and only if the word operation $u : A^n \to A$ distributes over the operation $+$.*

Corollary 23. *A variety $\mathcal{S} \subseteq \mathcal{SV}$ of semilattice ordered algebras satisfies each linear identity true in $\mathcal{V}$.*

Corollary 24. *Let $\mathcal{V}$ be a variety of Ω-algebras satisfying an identity $\omega(x,\dots,x) = x$, for $\omega \in \Omega$. The identity $\omega(x,\dots,x) = x$ is true in a variety $\mathcal{S} \subseteq \mathcal{SV}$ of semilattice ordered algebras if and only if the following identity*

$$x + y = \sum_{x_i \in \{x,y\}} \omega(x_1,\dots,x_n)$$

is true in $\mathcal{S}$.

Let ${}^{\cup}\mathcal{V}\Sigma^{<\omega}_{>0}$ denote the variety of semilattice ordered algebras generated by extended power algebras of finite non-empty subsets of algebras from $\mathcal{V}$, i.e.,

$${}^{\cup}\mathcal{V}\Sigma^{<\omega}_{>0} := \mathrm{HSP}(\{(P^{<\omega}_{>0}A,\Omega,\cup) \mid (A,\Omega) \in \mathcal{V}\}).$$

Theorem 25. *Let $\mathcal{V}$ be a variety of Ω-algebras. The variety $\mathcal{V}\Sigma^{<\omega}_{>0}$ is locally finite if and only if the variety ${}^{\cup}\mathcal{V}\Sigma^{<\omega}_{>0}$ is locally finite.*

Proof. Let $(C,\Omega,\cup) \in^{\cup} \mathcal{V}\Sigma_{>0}^{<\omega}$ be the algebra generated by a finite set $X \subseteq C$. By Disjunctive Form Lemma 17, for each $a \in C$, there exist $a_1,\ldots,a_p \in \langle X\rangle_\Omega$ such that

$$a = a_1 \cup \ldots \cup a_p. \tag{5}$$

If the variety $\mathcal{V}\Sigma_{>0}^{<\omega}$ is locally finite, then the algebra $\langle X\rangle_\Omega \in \mathcal{V}\Sigma_{>0}^{<\omega}$ is finite. Hence, there are only finitely many elements of the form (5). Consequently, the algebra $(C,\Omega,\cup)$ is finite.

Let $(F_{\cup\mathcal{V}\Sigma_{>0}^{<\omega}}(X),\Omega,\cup)$ be the free algebra in the variety ${}^{\cup}\mathcal{V}\Sigma_{>0}^{<\omega}$ generated by a set X. It is known that the free algebra over X in the variety generated by Ω-subreducts of algebras in ${}^{\cup}\mathcal{V}\Sigma_{>0}^{<\omega}$ is isomorphic to the Ω-subreduct $(\langle X\rangle,\Omega)$, generated by X, of the free algebra $(F_{\cup\mathcal{V}\Sigma_{>0}^{<\omega}}(X),\Omega,\cup)$. (See e.g. [17, Theorem 3.9]). The free algebra $(F_{\mathcal{V}\Sigma_{>0}^{<\omega}}(X),\Omega)$ is then a homomorphic image of $(\langle X\rangle_\Omega,\Omega)$. Consequently, the variety $\mathcal{V}\Sigma_{>0}^{<\omega}$ is locally finite if the variety ${}^{\cup}\mathcal{V}\Sigma_{>0}^{<\omega}$ is locally finite. □

Note that the same is true also for varieties generated by power algebras of all finite subsets (i.e. including the empty set).

4. FREE SLO ALGEBRAS WITH CONSTANTS

Let $(F_{\mathcal{V}}(X),\Omega)$ be the free algebra over a set X in the variety $\mathcal{V} \subseteq \mho$ and $(F_{\mathcal{V}_1}(X),\Omega)$ be the free algebra over a set X in the variety $\mathcal{V}_1 \subseteq \mho_1$. Let $\mathcal{S}_{\mathcal{V}}$ denote the variety of all semilattice ordered $\mathcal{V}$-algebras, $\mathcal{S}_{\mathcal{V}}^0$ denote the variety of all 0-semilattice ordered $\mathcal{V}$-algebras, $\mathcal{S}_{\mathcal{V}_1}$ denote the variety of all semilattice ordered $\mathcal{V}_1$-algebras and $\mathcal{S}_{\mathcal{V}_1}^0$ denote the variety of all 0-semilattice ordered $\mathcal{V}_1$-algebras.

Theorem 26 (Universality Property for Semilattice Ordered Algebras)**.** *[20] Let X be an arbitrary set and $(A,\Omega,+) \in \mathcal{S}_{\mathcal{V}}$. Each mapping $h\colon X \to A$ can be extended to a unique homomorphism $\overline{\overline{h}}\colon (\mathcal{P}_{>0}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup) \to (A,\Omega,+)$ such that $\overline{\overline{h}}|_X = h$.*

Corollary 27 (Universality Property for 0-Semilattice Ordered Algebras)**.** *Let X be an arbitrary set and $(A,\Omega,+,0) \in \mathcal{S}_{\mathcal{V}}^0$. Each mapping $h\colon X \to A$ can be extended to a unique homomorphism $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset) \to (A,\Omega,+,0)$ such that $\overline{\overline{h}}|_X = h$.*

Proof. Let $(A,\Omega,+,0) \in \mathcal{S}_{\mathcal{V}}^0$. Since $(A,\Omega) \in \mathcal{V}$ then any mapping $h\colon X \to A$ may be uniquely extended to an Ω-homomorphism $\overline{h}\colon (F_{\mathcal{V}}(X),\Omega) \to (A,\Omega)$.

Let us define the mapping $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset) \to (A,\Omega,+,0)$ by

$$\overline{\overline{h}}(T) = \sum_{t\in T} \overline{h}(t),$$

if T is a non-empty finite subset of $F_{\mathcal{V}}(X)$ and

$$\overline{\overline{h}}(\emptyset) = 0.$$

By Theorem 26 the mapping $\overline{\overline{h}}|_{\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)}$ is the unique $\{\Omega,\cup\}$-homomorphism such that $\overline{\overline{h}}|_X = h$. But obviously if $T_i = \emptyset$ for some i, we also have:

$$\begin{aligned}\overline{\overline{h}}(\omega(T_1,\ldots,\emptyset,\ldots,T_n)) = \overline{\overline{h}}(\emptyset) = 0 &= \omega(\overline{\overline{h}}(T_1),\ldots,0,\ldots,\overline{\overline{h}}(T_n)) \\ &= \omega(\overline{\overline{h}}(T_1),\ldots,\overline{\overline{h}}(\emptyset),\ldots,\overline{\overline{h}}(T_n)).\end{aligned}$$

Moreover, for $T_1 = \emptyset$

$$\overline{\overline{h}}(\emptyset \cup T_2) = \overline{\overline{h}}(T_2) = 0 + \overline{\overline{h}}(T_2) = \overline{\overline{h}}(\emptyset) + \overline{\overline{h}}(T_2),$$

which shows that the mapping $\overline{\overline{h}}|_{\mathcal{P}^{<\omega}F_{\mathcal{V}}(X)}$ is an $\{\Omega,\cup,\emptyset\}$-homomorphism. This completes the proof. □

Corollary 28 (Universality Property for Semilattice Ordered Algebras with a unit)**.** *Let X be an arbitrary set and $(A,\Omega,+,1) \in \mathcal{S}_{\mathcal{V}_1}$. Each mapping $h\colon X \to A$ can be extended to a unique homomorphism $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}_1}(X),\Omega,\cup,\{1\}) \to (A,\Omega,+,1)$ such that $\overline{\overline{h}}|_X = h$.*

Proof. Let $(A,\Omega,1)$ be a $\mathcal{V}_1$-algebra with the unit $1 \in A$. Then for the $\{\Omega,1\}$-homomorphism $\overline{h}\colon (F_{\mathcal{V}_1}(X),\Omega,1) \to (A,\Omega,1)$ which is an extension of a mapping $h\colon X \to A$ one has that $\overline{h}(1) = 1$. Further, by Theorem 26, the mapping $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}_1}(X),\Omega,\cup) \to (A,\Omega,+)$,

$$\overline{\overline{h}}(T) = \sum_{t\in T}\overline{h}(t)$$

for a non-empty finite subset T of $F_{\mathcal{V}_1}(X)$, is a homomorphism such that $\overline{\overline{h}}|_X = h$. In particular, for $T = \{1\}$

$$\overline{\overline{h}}(\{1\}) = \overline{h}(1) = 1.$$

□

Directly from Corollaries 27-28 we obtain the following corollary:

Corollary 29 (Universality Property for 0-Semilattice Ordered Algebras with a unit)**.** *Let X be an arbitrary set and $(A,\Omega,+,0,1) \in \mathcal{S}^0_{\mathcal{V}_1}$. Each mapping $h\colon X \to A$ can be extended to a unique homomorphism $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}F_{\mathcal{V}_1}(X),\Omega,\cup,\emptyset,\{1\}) \to (A,\Omega,+,0,1)$ such that $\overline{\overline{h}}|_X = h$.*

By Theorem 26 and Corollaries 27-29, for an arbitrary variety $\mathcal{V} \subseteq \mathfrak{V}$ or $\mathcal{V}_1 \subseteq \mathfrak{V}_1$, algebras $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$, $(\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset)$ or $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}_1}(X),\Omega,\cup,\{1\})$ have the universality property for semilattice ordered algebras in $\mathcal{S}_{\mathcal{V}}$, $\mathcal{S}^0_{\mathcal{V}}$ or $\mathcal{S}_{\mathcal{V}_1}$, respectively, but in general, the algebras themselves need not belong to these varieties.

Example 30. *Let $\mathcal{V}$ be a variety of semilattices $(A,\cdot)$ and $\mathcal{V}_1$ be a variety of semilattices $(A,\cdot,1)$ with the greatest element* 1.

Consider the free semilattice $(F_{\mathcal{V}}(X),\cdot)$ over a set X in the variety $\mathcal{V}$, the free algebra $(F_{\mathcal{V}_1}(X),\cdot,1)$ over a set X in the variety $\mathcal{V}_1$ and their two generators $x,y\in X$. One can easily see that

$$\{x,y\}\cdot\{x,y\}=\{x,x\cdot y,y\}\neq\{x,y\}.$$

This shows that the algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\cdot,\cup)$ does not belong to the variety $\mathcal{S}_{\mathcal{V}}$. This also immediately implies that algebras $(\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\cdot,\cup,\emptyset)$ and $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)F_{\mathcal{V}}(X),\cdot,\cup,\{1\})$ do not belong to varieties $\mathcal{S}^0_{\mathcal{V}}$ and $\mathcal{S}_{\mathcal{V}_1}$, respectively.

Corollary 31. *[20] The semilattice ordered algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$ is free over a set X in the variety $\mathcal{S}_{\mathcal{V}}$ if and only if $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)\in\mathcal{S}_{\mathcal{V}}$.*

Corollary 32. *The semilattice ordered algebra $(\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset)$ is free over a set X in the variety $\mathcal{S}^0_{\mathcal{V}}$ if and only if $(\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset)\in\mathcal{S}^0_{\mathcal{V}}$.*

Corollary 33. *The semilattice ordered algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}_1}(X),\Omega,\cup,\{1\})$ is free over a set X in the variety $\mathcal{S}_{\mathcal{V}_1}$ if and only if $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}_1}(X),\Omega,\cup,\{1\})\in\mathcal{S}_{\mathcal{V}_1}$.*

Corollary 34. *The semilattice ordered algebra $(\mathcal{P}^{<\omega}F_{\mathcal{V}_1}(X),\Omega,\cup,\emptyset,\{1\})$ is free over a set X in the variety $\mathcal{S}^0_{\mathcal{V}_1}$ if and only if $(\mathcal{P}^{<\omega}F_{\mathcal{V}_1}(X),\Omega,\cup,\emptyset,\{1\})\in\mathcal{S}^0_{\mathcal{V}_1}$.*

Corollary 35. *[20] Let $(F_{\mathfrak{V}}(X),\Omega)$ be the free algebra over a set X in the variety $\mathfrak{V}$. The extended power algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathfrak{V}}(X),\Omega,\cup)$ is free over X in the variety $\mathcal{S}_{\mathfrak{V}}$ of all semilattice ordered $\mathfrak{V}$-algebras.*

Note that, by Corollary 16, the same holds also for any variety defined by a set of linear identities.

Theorem 36. *[20] Let $\mathcal{V}$ be a variety defined by a set of linear identities. The extended power algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$ is free over X in the variety $\mathcal{S}_{\mathcal{V}}$ of all semilattice ordered $\mathcal{V}$-algebras.*

Theorem 37. *Let $\mathcal{V}$ be a variety defined by a set of linear regular identities. The $\emptyset$-extended power algebra $(\mathcal{P}^{<\omega}F_{\mathcal{V}}(X),\Omega,\cup,\emptyset)$ is free over X in the variety $\mathcal{S}^0_{\mathcal{V}}$ of all* 0*-semilattice ordered $\mathcal{V}$-algebras.*

For a variety $\mathcal{V}$ let $\mathcal{V}^*$ be its ***linearization***, the variety defined by all linear identities satisfied in $\mathcal{V}$. Obviously, $\mathcal{V}^*$ contains $\mathcal{V}$ as a subvariety.

Since by Theorem 12 and Corollary 13 for any subvariety $\mathcal{V}\subseteq\mathfrak{V}$ the algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega)$ satisfies only those identities which are obtained from the linear identities true in $\mathcal{V}$ through identification of variables, then for each subvariety $\mathcal{V}\subseteq\mathfrak{V}$, the algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega)$ belongs to $\mathcal{V}^*$, but it does not belong to any of its proper subvarieties.

Corollary 38. *Let X be an infinite set. For any subvariety $\mathcal{V}\subseteq\mathfrak{V}$ we have*

$$\mathcal{S}_{\mathcal{V}^*}=\mathrm{HSP}((\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}^*}(X),\Omega,\cup)).$$

Let $\mathcal{S}$ be a non-trivial subvariety of $\mathcal{S}_{\mathcal{V}}$ and X be a set. By [24, Chapter 3.3] the congruence

$$\Phi_{\mathcal{S}}(X) := \bigcap\{\phi \in Con(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup) \mid (\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)/\phi,\Omega,\cup) \in \mathcal{S}\}$$

is the $\mathcal{S}$-***replica congruence*** of $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$ and $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)/\Phi_{\mathcal{S}}(X),\Omega,\cup)$ is called the $\mathcal{S}$-***replica*** of $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$.

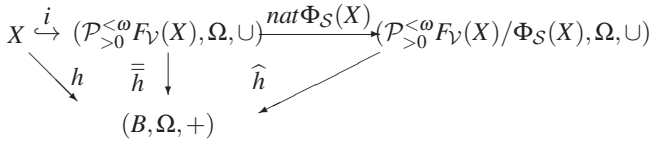
Let $(B,\Omega,+) \in \mathcal{S}$. By the universality property of replication (see [24, Lemma 3.3.1.]), for each homomorphism $\overline{\overline{h}}\colon (\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup) \to (B,\Omega,+)$, there is a unique homomorphism

$$\widehat{h}\colon (\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)/\Phi_{\mathcal{S}}(X),\Omega,\cup) \to (B,\Omega,+)$$

such that

$$\overline{\overline{h}} = \widehat{h} \circ nat\Phi_{\mathcal{S}}(X),$$

where $nat\Phi_{\mathcal{S}}(X)$ is the natural projection onto the quotient $\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X)/\Phi_{\mathcal{S}}(X)$. Hence, the universality property for $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$ yields the following commuting diagram for any mapping $h\colon X \to B$:



As a result, we obtain the following theorem

Theorem 39. *The $\mathcal{S}$-replica of the algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup)$ is free over a set X in the variety $\mathcal{S} \subseteq \mathcal{S}_{\mathcal{V}}$.*

Corollary 40. *Let $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{V}}(X),\Omega,\cup) \in \mathcal{S}$. Then it is free in $\mathcal{S} \subseteq \mathcal{S}_{\mathcal{V}}$ over a set X.*

Additionally, by Theorem [17, Theorem 3.9] one can obtain a characterization of free algebras in the quasivariety $\mathcal{U}_{\mathcal{S}}$ of Ω-subreducts of semilattice ordered algebras in a given variety $\mathcal{S}$.

Corollary 41. *The free algebra $(F_{\mathcal{U}_{\mathcal{S}}}(X),\Omega)$ over X in $\mathcal{U}_{\mathcal{S}}$ is isomorphic to the full Ω-subreduct $\langle X\rangle_{\Omega}$ of the free semilattice ordered algebra $(F_{\mathcal{S}}(X),\Omega,+)$ in $\mathcal{S}$.*

Free algebras in a quasivariety $\mathcal{U}_{\Omega}$ are also free in the variety $V(\mathcal{U}_{\Omega})$ generated by $\mathcal{U}_{\Omega}$ (see [14]). But note that, even if we have a free semilattice ordered algebra $(F_{\mathcal{S}}(X),\Omega,+)$ in a given quasivariety $\mathcal{S} \subseteq \mathcal{SV}$, its full Ω-subreduct $(\langle X\rangle_{\Omega},\Omega)$ need not be a free algebra in $\mathcal{V}$.

5. APPLICATIONS

5.1. IDEMPOTENT SLO ALGEBRAS

As we have already shown in Section 4, in varieties of semilattice ordered algebras true identities are determined by appropriate extended power algebras of algebras and their homomorphic images. Entropic and symmetric identities are both linear and regular. In this section we focus on the idempotent identities which are linear only if the operation ω occuring there is unary.

Extended power algebras are very rarely idempotent. Note that if the power algebra $(\mathcal{P}_{>0}A,\Omega)$ of (A,Ω) is idempotent then the algebra (A,Ω) must be idempotent too. Furthermore, if (A,Ω) is idempotent then for any non-empty subset B of A and $\omega\in\Omega$, we have $B\subseteq\omega(B,\dots,B)$. Moreover, as an easy consequence of results of A. Romanowska and J.D.H. Smith [23, Proposition 2.1] for an idempotent algebra (A,Ω), a non-empty subset $B\in\mathcal{P}_{>0}A$ is a subalgebra of (A,Ω) if and only if $\omega(B,\dots,B)=B$ for each $\omega\in\Omega$.

Corollary 42. *[21] The power algebra $(\mathcal{P}_{>0}A,\Omega)$ of an idempotent algebra (A,Ω) is idempotent if and only if each non-empty subset B of A is a subalgebra of (A,Ω).*

Example 43. *[21] An algebra (A,Ω) such that $\omega(a_1,\dots,a_n)\in\{a_1,\dots,a_n\}$, for each n-ary $\omega\in\Omega$ and $a_1,\dots,a_n\in A$, is called **conservative**. By Corollary 42, the power algebra of any conservative algebra is idempotent. In particular, the power algebra of a chain, the power algebra of a left zero-semigroup [24], the power algebra of an equivalence algebra [9] and the power algebra of a tournament [10] are all idempotent.*

Let θ be a congruence on an idempotent algebra (A,Ω). Obviously, $a\ \theta\ \omega(a,\dots,a)$ for each $a\in A$ and $\omega\in\Omega$. On the other hand, it is not always true that $X\ \theta\ \omega(X,\dots,X)$ for a subset X of A, if $(\mathcal{P}_{>0}A,\Omega)$ is not idempotent. It is enough to consider the equality relation on (A,Ω) in such a case.

Let (M,Ω) be an idempotent and entropic algebra. Denote by $\mathcal{I}$ the variety of all idempotent τ-algebras of type $\tau:\Omega\cup\{\cup\}\to\mathbb{N}^+$. Then $Con_{\mathcal{I}}(\mathcal{P}^{<\omega}_{>0}M)$ is the set of all congruence relations γ on $(\mathcal{P}^{<\omega}_{>0}M,\Omega,\cup)$, such that the quotient $(\mathcal{P}^{<\omega}_{>0}M^{\gamma},\Omega)$ is idempotent. By [24, Section 1.4.3] $Con_{\mathcal{I}}(\mathcal{P}^{<\omega}_{>0}M)$ is an algebraic subset of the lattice of all congruences of $(\mathcal{P}^{<\omega}_{>0}M,\Omega,\cup)$. Recall that the least element in $(Con_{\mathcal{I}}(\mathcal{P}^{<\omega}_{>0}M),\subseteq)$ is the $\mathcal{I}$-replica congruence of $(\mathcal{P}^{<\omega}_{>0}M,\Omega,\cup)$.

Let $(M,\Omega,1)$ be an idempotent and entropic algebra with the unit $1\in M$ and let the algebra $(\mathcal{P}^{<\omega}M,\Omega,\cup,\emptyset,\{1\})$ be the $\emptyset$-extended power algebra with the unit $\{1\}$. Let us define a binary relation ρ on the set $\mathcal{P}^{<\omega}M$ in the following way:

$$\begin{aligned} A\ \rho\ B \quad\Leftrightarrow\quad & \text{there exist a } k\text{-ary term } t \text{ and an } m\text{-ary term } s \\ & \text{both of type } \Omega \text{ such that} \\ & A\subseteq t(B,B,\dots,B) \ \text{ and } \ B\subseteq s(A,A,\dots,A). \end{aligned} \tag{6}$$

It was proved in [19] that the relation $\rho|_{\mathcal{P}^{<\omega}_{<0}M}$ is the $\mathcal{I}$-replica congruence of $(\mathcal{P}^{<\omega}_{>0}M,\Omega,\cup)$ and is equal to the relation:

$$A\ \alpha\ B \quad \Leftrightarrow \quad \langle A\rangle = \langle B\rangle, \tag{7}$$

where $\langle A\rangle$ is the subalgebra of (M,Ω) generated by the set A. Therefore, $(\mathcal{P}^{<\omega}_{>0}M/\rho,\Omega,\cup)\cong(\{\langle A\rangle\colon A\in\mathcal{P}^{<\omega}_{>0}M\},\Omega,+)$, where for each n-ary complex operation $\omega\in\Omega$ and non-empty subsets $A_1,\ldots,A_n$ of M

$$\omega(\langle A_1\rangle,\ldots,\langle A_n\rangle)=\langle\omega(A_1,\ldots,A_n)\rangle \text{ and} \tag{8}$$

$$\langle A_1\rangle+\langle A_2\rangle=\langle A_1\cup A_2\rangle. \tag{9}$$

It is easy to notice that

$$\emptyset\ \rho\ A \quad \Leftrightarrow \quad A=\emptyset$$

and

$$\{1\}\ \rho\ A \quad \Leftrightarrow \quad A=\{1\}.$$

Hence, assuming that $\langle\emptyset\rangle=\emptyset$, ρ is also the $\mathcal{I}$-replica congruence of $(\mathcal{P}^{<\omega}M,\Omega,\cup)$ and $(\mathcal{P}^{<\omega}M/\rho,\Omega,\cup)\cong(\{\langle A\rangle\colon A\in\mathcal{P}^{<\omega}M\},\Omega,+)$.

By Theorem 39 we have the following theorem

Theorem 44. *Let $\mathcal{M}$ be the variety of all idempotent and entropic Ω-algebras (M,Ω). The 0-semilattice ordered algebra $(\{\langle A\rangle\colon A\in\mathcal{P}^{<\omega}F_{\mathcal{M}}(X)\},\Omega,+,\emptyset)$ is free over a set X in the variety $\mathcal{S}^0_{\mathcal{M}}$.*

Moreover, for any k-ary term t and a subset $1\notin S\subset M$

$$\{1\}\cup S\subseteq t(S,\ldots,S) \quad \Leftrightarrow \quad \exists(s_1,\ldots,s_k\in S) \quad 1=t(s_1,\ldots,s_k).$$

This shows that for a non-empty subset $S\subset M$ such that $1\notin S$

$$\{1\}\cup S\ \rho\ S \quad \Leftrightarrow \quad \begin{array}{l}\text{there exists a } k\text{-ary term } t \text{ of type } \Omega \text{ and } s_1,\ldots,s_k\in S\\ \text{such that } 1=t(s_1,\ldots,s_k).\end{array}$$

Lemma 45. *Let $(M,\Omega,1)$ be an idempotent and entropic algebra with the unit $1\in M$. Let us assume that the algebra $(M,\Omega,1)$ satisfies the following condition:*

$$\forall(\omega\in\Omega)\,\forall(x_1,\ldots,x_n\in M) \quad \omega(x_1,\ldots,x_n)=1 \quad \Rightarrow \quad \forall(1\le i\le n) \quad x_i=1. \tag{10}$$

Then the relation ρ is the $\mathcal{I}$-replica congruence of $(\mathcal{P}^{<\omega}_{>0}M,\Omega,\cup,\{1\})$ and the quotient $(\mathcal{P}^{<\omega}_{>0}M/\rho,\Omega,\cup,\{1\})$ is isomorphic to the semilattice ordered algebra:

$$(\{\langle A\rangle\colon A\in\mathcal{P}^{<\omega}_{>0}(M\setminus\{1\})\}\cup\{\langle A\cup\{1\}\rangle\colon A\in\mathcal{P}^{<\omega}_{>0}M\},\Omega,+,\{1\}).$$

Theorem 46. *Let $\mathcal{M}_1$ be the variety of all idempotent and entropic Ω-algebras $(M,\Omega,1)$ with the unit 1 which satisfy Condition* (10)*.*

Then the semilattice ordered algebra $(\{\langle A\rangle\colon 1\notin A \text{ and } A\in \mathcal{P}^{<\omega}_{>0}F_{\mathcal{M}}(X)\}\cup \{\langle A\cup\{1\}\rangle\colon 1\notin A \text{ and } A\in \mathcal{P}^{<\omega}_{>0}F_{\mathcal{M}}(X)\}\cup\{\{1\}\},\Omega,+,\{1\})$ is free over a set X in the variety $\mathcal{S}_{\mathcal{M}_1}$.

Corollary 47. *Let $\mathcal{M}_1$ be the variety of all idempotent and entropic Ω-algebras $(M,\Omega,1)$ with the unit 1 which satisfy Condition* (10)*.*

Then the 0-semilattice ordered algebra $(\{\langle A\rangle\colon 1\notin A \text{ and } A\in \mathcal{P}^{<\omega}F_{\mathcal{M}}(X)\}\cup \{\langle A\cup\{1\}\rangle\colon 1\notin A \text{ and } A\in \mathcal{P}^{<\omega}F_{\mathcal{M}}(X)\},\Omega,+,\emptyset,\{1\})$ is free over a set X in the variety $\mathcal{S}^0_{\mathcal{M}_1}$.

5.2. COMMUTATIVE DOUBLE IDEMPOTENT SEMIRINGS

Definition 48. *A **semiring** is an algebra $(S,\cdot,+)$ such that*

1. *$(S,\cdot)$ is a semigroup,*
2. *$(S,+)$ is a commutative semigroup,*
3. *for $a,b,c\in S$, $a\cdot(b+c)=a\cdot b+a\cdot c$ and $(b+c)\cdot a=b\cdot a+c\cdot a$.*

A semiring is said to be ***commutative*** if the semigroup $(S,\cdot)$ is commutative. A semiring is ***additively idempotent*** if the semigroup $(S,+)$ is idempotent and it is ***multiplicatively idempotent*** if $(S,\cdot)$ is idempotent. Hence, additively idempotent semirings are simply semilattice ordered semigroups.

Remark 49. *Notice that in the literature of semirings there are several definitions depending on whether the algebra contains an identity and/or a zero element. See e.g. [7].*

Let $\mathcal{SG}$ denote the variety of all semigroups. By Theorem 36, since associativity is a linear identity, the extended power algebra $(\mathcal{P}^{<\omega}_{>0}F_{\mathcal{SG}}(X),\cdot,\cup)$ is free over X in the variety $\mathcal{S}_{\mathcal{SG}}$ of all additively idempotent semirings. Free additively idempotent semirings, where $(S,\cdot)$ belongs to a subvariety of $\mathcal{SG}$, defined by a set of linear identities, can be described in a similar way.

On the other hand, idempotency is not a linear identity so the extended power algebra of the free algebra in the variety of all idempotent semigroups (*bands*) need not be idempotent. As a consequence, such algebra is not a free algebra in the variety of all additively and multiplicatively idempotent semirings. Double idempotent semirings were called *distributive ·-bisemilattices* by R. McKenzie and A. Romanowska and studied in [15].

If the semigroup $(S,\cdot)$ is also entropic (*normal band*), i.e. it satisfies for $a,b,c,d\in S$

$$a\cdot b\cdot c\cdot d=a\cdot c\cdot b\cdot d,$$

then by Theorem 36 the algebra $(\{\langle A\rangle : A \in \mathcal{P}^{<\omega}_{>0} F_{\mathcal{NB}}(X)\}, \cdot, +)$ of all finitely generated subalgebras of free algebra $F_{\mathcal{NB}}(X)$ in the variety $\mathcal{NB}$ of all normal bands is free in the variety of double idempotent semirings with entropic multiplication reduct. This result coincides with a construction given by Zhao in [27] where he applied so called *closed subsets*, since each non-empty subset of a normal band is closed if and only if it is a subband.

Quite recently Chajda and Langer [3] investigated commutative double idempotent semirings $(S, \cdot, +, 0, 1)$ with two constants 0 and 1, such that $(S, +, 0)$ and $(S, \cdot, 1)$ are semilattices with the least element 0 and the greatest element 1, respectively, and for each $x \in S$,

$$x \cdot 0 = 0 \cdot x = 0.$$

Clearly, such semirings are exactly 0-semilattice ordered semilattices with a unit 1. In particular, Chajda and Langer described free algebras in the variety $\mathcal{Z}$ of all commutative double idempotent semirings with two constants. Since commutative semigroups are trivially entropic some results in [3] immediately follow by general ones.

Let $\mathcal{SL}_1$ be the variety of all semilattices with a unit 1 and let $(F_{\mathcal{SL}}(X), \cdot)$ be the free semilattice in $\mathcal{SL}$ generated by a set X. Obviously, condition (10) is satisfied in any idempotent monoid so by Corollary 47 we obtain:

Theorem 50. *The* 0*-semilattice ordered algebra* $(\{\langle A\rangle : 1 \notin A \in \mathcal{P}^{<\omega} F_{\mathcal{SL}}(X)\} \cup \{\langle A\rangle \cup \{1\} : 1 \notin A \in \mathcal{P}^{<\omega} F_{\mathcal{SL}}(X)\}, \cdot, +, \emptyset, \{1\})$ *is free over a set* X *in the variety* $\mathcal{Z} = \mathcal{S}^0_{\mathcal{SL}_1}$.

Therefore, directly by Disjunctive Form Lemma 17, every term $t(x_1, \ldots, x_n) \in F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)$ is a sum of some products of variables $x_1, \ldots, x_n$ (see [3, Lemma 4]).

Furthermore, it is well known that the free algebra generated by X in the variety $\mathcal{SL}_0$ is isomorphic to the semilattice $(\mathcal{P}X, \cup)$ of all subsets of X. Then the number of different n-ary terms in $F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)$ is less than or equal to 2^{2^n} (see [3, Corollary 5]). The local finiteness of $\mathcal{Z} = F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)$ follows also by Theorem 25.

If X is a finite set, then by Theorem 50 the cardinality of $F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)$ is twice that of the set of all subalgebras of the free algebra in the variety $\mathcal{SL}$ including the empty set:

$$|F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)| = 2|\{(A, \cdot) : (A, \cdot) \leq F_{\mathcal{SL}}(X)\}|.$$

In particular, for $X = \emptyset$ there is only one subalgebra of $F_{\mathcal{SL}}(\emptyset)$: the empty set. Then $F_{\mathcal{S}^0_{\mathcal{SL}_1}}(\emptyset) \cong (\{\emptyset, \{1\}\}, \cdot, \cup, \emptyset, \{1\})$. Furthermore, for $X = \{x\}$ we obtain

$$F_{\mathcal{S}^0_{\mathcal{SL}_1}}(\{x\}) \cong (\{\emptyset, \{x\}, \{1\}, \{x, 1\}\}, \cdot, \cup, \emptyset, \{1\}).$$

For $X = \{x, y\}$, the free semilattice $F_{\mathcal{SL}}(X)$ on two generators has three elements: x, y, xy and 7 subalgebras (including the empty set): $\emptyset, \{x\}, \{y\}, \{xy\}, \{x, xy\}, \{y, xy\}, \{x, y, xy\}$. Hence $|F_{\mathcal{S}^0_{\mathcal{SL}_1}}(\{x, y\})| = 14$.

Referring to the notion introduced in [3] we say that a subset A of $F_{\mathcal{SL}}(X)$ is ***reduced*** if

$$\forall(a \in A)\ \forall(k \in \mathbb{N}^+)\ \forall(b_1,\ldots,b_k \in A \setminus \{a\}) \quad a \neq b_1 \cdots b_k.$$

It is evident that for each finitely generated subalgebra $(C,\cdot)$ of $F_{\mathcal{SL}}(X)$ there exists exactly one finite reduced subset $A_r \subseteq F_{\mathcal{SL}}(X)$ such that $(C,\cdot) = \langle A_r \rangle$. Hence, the cardinality of the free algebra $F_{\mathcal{S}_{\mathcal{SL}}}(X)$ in the variety $\mathcal{S}_{\mathcal{SL}}$ of all semilattice ordered semilattices is equal to the cardinality of all reduced subsets of $F_{\mathcal{SL}}(X)$. This implies that the cardinality of $F_{\mathcal{S}^0_{\mathcal{SL}_1}}(X)$ is twice that of the set of all reduced subsets of $F_{\mathcal{SL}}(X)$ including the empty set.

Acknowledgements

While working on this paper, the authors were supported by the Grant of Warsaw University of Technology 504/04259/1120.

References

[1] N. Alpay, P. Jipsen, *Commutative Doubly-Idempotent Semirings Determined by Chains and by Preorder Forests*, in: Fahrenberg U., Jipsen P., Winter M. (eds) Relational and Algebraic Methods in Computer Science RAMiCS 2020, Lecture Notes in Computer Science, vol 12062, Springer, Cham (2020).
[2] S.L. Bloom, *Varieties of ordered algebras*, J. Comput. Syst. Sci. 13 (1976) 200–212.
[3] I. Chajda, H. Länger, *The variety of commutative additively and multiplicatively idempotent semigroups*, Semigroup Forum 96 (2018) 409–415.
[4] G. Czédli, A. Lenkehegyi, *On classes of ordered algebras and quasiorder distributivity*, Acta Sci. Math. 46 (1983) 41–54.
[5] L. Fuchs, *On partially ordered algebras I*, Colloq. Math. 14 (1966) 113–130.
[6] S. Ghosh, F. Pastijn, X.Z. Zhao, *Varieties generated by ordered bands I*, Order 22 (2005) 109–128.
[7] J.S. Golan, *Semirings and Their Applications*, Kluwer, Dordrecht (1999).
[8] G. Grätzer, H. Lakser, *Identities for globals (complex algebras) of algebras*, Colloq. Math. 56 (1988) 19–29.
[9] J. Ježek, R. McKenzie, *The variety generated by equivalence algebras*, Algebra Universalis, 2001, 45, 211–219.
[10] J. Ježek, P. Marković, M. Maróti, R. McKenzie, *The variety generated by tournaments*, Acta Univ. Carolin. Math. Phys., 1999, 40, no. 1, 21–41.
[11] B. Jónsson, A. Tarski, *Boolean algebras with operators I*, Am. J. Math. 73 (1951) 891–939.
[12] B. Jónsson, A. Tarski, *Boolean algebras with operators II*, Am. J. Math. 74 (1952), 127–167.
[13] K. Kearnes, *Semilattice modes I: the associated semiring*, Algebra Univers. 34 (1995) 220–272.
[14] A.I. Mal'cev, *Algebraičeskie sistemy*, [in Russian], Sovremennaja Algebra, Nauka, Moscow, 1970. English translation: *Algebraic systems*, Springer Verlag, Berlin, 1973.
[15] R. McKenzie, A. Romanowska, *Varieties of · - distributive bisemilattices*, Contrib. Gen. Algebra (Proc. Klagenfurt Conf. 1978) 213–218.
[16] F. Pastijn, X.Z. Zhao, *Varieties of idempotent semirings with commutative addition*, Algebra Univers. 54 (2005), 301–321.
[17] A. Pilitowska, A. Zamojska-Dzienio, *Representation of modals*, Demonstr. Math. 44(3) (2011) 535–556.
[18] A. Pilitowska, A. Zamojska-Dzienio, *On some congruences of power algebras*, Cent. Eur. J. Math. 10 (2012) 987–1003.

[19] A. Pilitowska, A. Zamojska-Dzienio, *Varieties generated by modes of submodes*, Algebra Univers. 68 (2012) 221–236.
[20] A. Pilitowska, A. Zamojska-Dzienio, *The lattice of subvarieties of semilattice ordered algebras*, Order 31 (2014) 217–238.
[21] A. Pilitowska, A. Zamojska-Dzienio, *Closure operators on algebras*, Internat.J.Algebra Comput. 25(6) (2015) 1055–1074.
[22] A.B. Romanowska, J.D.H. Smith, *Modal Theory*, Heldermann Verlag, Berlin (1985).
[23] A.B. Romanowska, J.D.H. Smith, *Subalgebra systems of idempotent entropic algebras*, J. Algebra, 1989, 120, 247–262.
[24] A.B. Romanowska, J.D.H. Smith, *Modes*, World Scientific, Singapore (2002).
[25] K.I. Rosenthal, *Quantales and their applications*, Pitman Res. Notes Math. Ser. 234, Longman Scientific & Technical, Harlow (copublished in the United States with John Wiley & Sons Inc., New York), 1990.
[26] W. Rump, Y.C. Yang, *Non-commutative logical algebras and algebraic quantales*, Ann. Pure Appl. Logic **165** (2014), 759–785.
[27] X.Z. Zhao, *Idempotent semirings with a commutative additive reducts*, Semigroup Forum 64 (2002) 289–296.

Leszek Pysiak[1], **Wiesław Sasin**[2]

[1] Institute of Mathematics and Cryptology
Military University of Technology, Warsaw, Poland
[2] Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland

SPACE-TIMES WITH INFINITESIMAL OPERATORS

Manuscript received: 13 July 2020
Manuscript accepted: 30 August 2020

Abstract: We present the concept of a differential manifold with infinitesimal operators and we investigate its geometric properties. We construct an algebra of real numbers with operators and using the Yoneda embedding we obtain generalized manifolds. This procedure is functorial, for every real manifold we obtain the associated manifold with infinitesimal operators. The aim of this paper is to propose the concept of space-time using methods of Synthetic Differential Geometry (SDG), which unifies space-time geometry with an algebra of operators.
Keywords: differential spaces, differential geometry, infinitesimals
Mathematics Subject Classification (2020): Primary 58A03, 58A40, 58A99

INTRODUCTION

In the paper we construct a generalized space-time which can be an arena of unifying general relativity (differential geometry of space-time [12]) and quantum mechanics (theory of operators on a Hilbert space [5]). Some generalizations of the notion of differential manifold as ringed spaces, called the differential spaces or the differentiable spaces, were introduced earlier by Sikorski [16, 17], Aronszajn [1] and Spallek [18], see also [10]. If M is a smooth manifold and $C^\infty(M)$ is the ring of smooth functions on it, we obtain the Sikorski differential space $(M, C^\infty(M))$ as a ringed space [13]. Differential spaces encode the structure of space in a ring of functions, and C^∞-rings of functions are a natural place for introducing infinitesimals, as it is done in synthetic differential geometry (see [3, 6, 7, 8, 9]). Let us notice (see [15]) that using the Yoneda embedding for two C^∞-rings A and B one can define a generalized space as the ringed space $(\mathrm{Hom}_{C^\infty}(A,B), \overline{A})$, where $\mathrm{Hom}_{C^\infty}(A,B)$ is the set of morphisms from A into B and the C^∞-algebra $\overline{A}$, called the differential strtucture of $\mathrm{Hom}_{C^\infty}(A,B)$, is defined in the following way:

For any $a \in A$ we define $\overline{a} : \mathrm{Hom}_{C^\infty}(A,B) \to B$ by

$$\overline{a}(\rho) = \rho(a)$$

and finally $\overline{A} = \{\overline{a} : a \in A\}$, which is obviously a C^∞-ring with the operation:

$$\omega(\overline{a}_1, \ldots, \overline{a}_n) = \overline{w(a_1, \ldots, a_n)}$$

for $n \in \mathbb{N}$, $\omega \in C^\infty(\mathbb{R}^n)$, $a_1, \ldots, a_n \in A$.

In Section 1, for any $\mathbb{R}$-algebra $\mathcal{U}$ we construct a C^∞-ring $R = \mathbb{R} \oplus \mathcal{U}\varepsilon$, with nilpotent ε such that $\varepsilon^2 = 0$. In fact, for future applications we consider $\mathcal{U}$ as an $\mathbb{R}$-algebra of operators on some Hilbert space.

In Section 2, for an arbitrary differential manifold M we construct the ringed space $(\overline{M}, C^\infty(\overline{M}))$, where $\overline{M} = \mathrm{Hom}_{C^\infty}(C^\infty(M), R)$ and $C^\infty(\overline{M}) := \overline{C^\infty(M))}$, which is called a manifold with infinitesimal operators.

In Section 3, we prove that C^∞-rings: $C^\infty(M)$ and $C^\infty(\overline{M})$ are isomorphic. In consequence we obtain a one-to-one correspondence between tensors on M and respective tensors on $\overline{M}$. The differential geometry on $\overline{M}$ can be interpreted as a copy of the differential geometry on M. $(\overline{M}, C^\infty(\overline{M}))$ is a richer space than $(M, C^\infty(M))$ but their differential geometries are „equivalent".

In Section 4, we present interesting examples of operators which illustrate our constructions. These operators and other concepts of the paper and of the work [15] will be applied to the theory of unification of relativity theory and quantum mechanics.

1. CARTESIAN SPACES WITH INFINITESIMAL OPERATORS

First, we recall the notion of the C^∞-ring, which plays an important part in our further considerations.

Definition 1. *A unital commutative $\mathbb{R}$-algebra A is a C^∞-*ring *if, given any $n, m \in \mathbb{N}$, $\omega \in C^\infty(\mathbb{R}^n)$ and $a_1, \ldots, a_n \in A$, the element $\omega(a_1, \ldots, a_n)$ is defined and the following conditions are satisfied:*

1. *for $\phi, \psi \in C^\infty(\mathbb{R}^2)$ such that $\phi(x_1, x_2) = x_1 \cdot x_2$, $\psi(x_1, x_2) = x_1 + x_2$, we have*

$$\phi(a_1, a_2) = a_1 \cdot a_2, \qquad \psi(a_1, a_2) = a_1 + a_2;$$

2. *for $\pi_i : \mathbb{R}^n \to \mathbb{R}$, $\pi_i(x_1, \ldots, x_n) = x_i$, $i = 1, \ldots, n$ we have*

$$\pi_i(a_1, \ldots, a_n) = a_i;$$

3. *for the constant function* $1 \in C^\infty(\mathbb{R}^n)$, *we have*

$$1(a_1,\dots,a_n) = 1_A;$$

4. *for* $\theta \in C^\infty(\mathbb{R}^n)$, $\omega_1,\dots,\omega_m \in C^\infty(\mathbb{R}^n)$ *we have*

$$(\theta\circ(\omega_1,\dots,\omega_m))(a_1,\dots,a_n) = \theta(\omega_1(a_1,\dots,a_n),\dots,\omega_m(a_1,\dots,a_n)).$$

Let A,B *be* C^∞*-rings. A homomorphism* $f : A \to B$ *of* $\mathbb{R}$*-algebras is called* C^∞-morphism *if, for any* $\omega \in C^\infty(\mathbb{R}^n)$, $n \in \mathbb{N}$, $a_1,\dots,a_n \in A$, *the following equality is satisfied:*

$$f(\omega(a_1,\dots,a_n)) = \omega(f(a_1),\dots,f(a_n)).$$

C^∞-rings as objects with C^∞-morphisms as morphisms form a category which will be denoted by C^∞.

Of course, $\mathbb{R}$ is a C^∞-ring with the following operation: for any $\omega\in C^\infty(\mathbb{R}^n)$, $x_1,\dots,x_n\in\mathbb{R}$, the element $\omega(x_1,\dots,x_n)$ is the value of ω for arguments $x_1,\dots,x_n$.

Now, let $\mathfrak{A}$ be an algebra of operators. We will construct an algebra of real numbers with infinitesimal operators denoted by $\mathbb{R}\oplus\mathfrak{A}\varepsilon$, satisfying the following conditions:

(i) every element has a form $x+\mathfrak{a}\varepsilon$, where $x\in\mathbb{R}$, $\mathfrak{a}\in\mathfrak{A}$ and $\varepsilon^2 = 0$,

(ii) $(x+\mathfrak{a}\varepsilon)+(y+\mathfrak{b}\varepsilon) = x+y+(\mathfrak{a}+\mathfrak{b})\varepsilon$,

(iii) $(x+\mathfrak{a}\varepsilon)\cdot(y+\mathfrak{b}\varepsilon) = x\cdot y+(x\mathfrak{b}+y\mathfrak{a})\varepsilon$.

Let us consider the pairs $(x,\mathfrak{a})$, where $x\in\mathbb{R}$ and $\mathfrak{a}\in\mathfrak{A}$. We define an *addition* and a *multiplication* by:

$$(x,\mathfrak{a})+(y,\mathfrak{b}) = (x+y,\mathfrak{a}+\mathfrak{b}),$$
$$(x,\mathfrak{a})\cdot(y,\mathfrak{b}) = (x\cdot y, x\mathfrak{b}+y\mathfrak{a})$$

for $x,y\in\mathbb{R}$ and $\mathfrak{a},\mathfrak{b}\in\mathfrak{A}$.

Let us put $x = (x,\mathfrak{o})$, $\mathfrak{a}\varepsilon = (0,\mathfrak{a})$, $\varepsilon = (0,\mathfrak{1})$, $1 = (1,\mathfrak{o})$, $0 = (0,\mathfrak{o})$, where $x\in\mathbb{R}$, $\mathfrak{a}\in\mathfrak{A}$, $\mathfrak{1}$ is the identity operator, $\mathfrak{o}$ is the zero operator in $\mathfrak{A}$. Every element $(x,\mathfrak{a})$ can be written in the following form:

$$(x,\mathfrak{a}) = (x,\mathfrak{o})+(0,\mathfrak{a}) = x\cdot(1,\mathfrak{o})+\mathfrak{a}\cdot(0,\mathfrak{1}) = x+\mathfrak{a}\varepsilon.$$

Lemma 2. $\mathbb{R}\oplus\mathfrak{A}\varepsilon$ *is a* C^∞*-ring with the operation*

$$\omega(x_1+\mathfrak{a}_1\varepsilon,\dots,x_n+\mathfrak{a}_n\varepsilon) = \omega(x_1,\dots,x_n)+\sum_{i=1}^{n}\omega'_{|i}(x_1,\dots,x_n)\mathfrak{a}_i\varepsilon$$

for any $\omega\in C^\infty(\mathbb{R}^n)$.

Proof. The verification of conditions 1–4 in Definition 1 is evident. □

In the sequel we will denote $\mathbb{R}\oplus\mathfrak{A}\varepsilon$ by R and will say that R is the real line with infinitesimal operators $\mathfrak{a}\varepsilon$, where $\mathfrak{a}\in\mathfrak{A}$. The Cartesian product $R^n = R\times\cdots\times R$ is called *n-dimensional space with infinitesimal operators*. An arbitrary point $(r_1,\ldots,r_n)\in R^n$ can be represented uniquely in the following form:

$$(r_1,\ldots,r_n) = (x_1+\mathfrak{a}_1\varepsilon,\ldots,x_n+\mathfrak{a}_n\varepsilon) = (x_1,\ldots,x_n)+(\mathfrak{a}_1,\ldots,\mathfrak{a}_n)\varepsilon,$$

where $(x_1,\ldots,x_n)\in\mathbb{R}^n$, $(\mathfrak{a}_1,\ldots,\mathfrak{a}_n)\in\mathfrak{A}^n$.

If we put $\mathfrak{d}_i := \mathfrak{a}_i\varepsilon$, then $\mathfrak{d}_i^2 = 0$ and we can write

$$(r_1,\ldots,r_n) = (x_1,\ldots,x_n)+(\mathfrak{d}_1,\ldots,\mathfrak{d}_n),$$

where $\mathfrak{d}_i\in\mathfrak{D} := \mathfrak{A}\varepsilon$. For any $f\in C^\infty(\mathbb{R}^n)$ we define $\bar f : R^n\to R$ by

$$\bar f(r_1,\ldots,r_n) = f(x_1,\ldots,x_n)+\sum_{i=1}^n f'_{|i}(x_1,\ldots,x_n)\mathfrak{a}_i\varepsilon$$

or equivalently

$$\bar f(x_1+\mathfrak{d}_1,\ldots,x_n+\mathfrak{d}_n) = f(x_1,\ldots,x_n)+\sum_{i=1}^n f'_{|i}(x_1,\ldots,x_n)\mathfrak{d}_i,$$

where $r_i = x_i+\mathfrak{d}_i$, $i=1,\ldots,n$.

Let us define $C^\infty(R^n) = \{\bar f : f\in C^\infty(\mathbb{R}^n)\}$. It is easy to see that $C^\infty(R^n)$ is a C^∞-ring with the operation:

$$\omega(\bar f_1,\ldots,\bar f_m) = \overline{\omega}\circ(\bar f_1,\ldots,\bar f_m)$$

for $f_i\in C^\infty(\mathbb{R}^n)$, $i=1,\ldots,m$, $\omega\in C^\infty(\mathbb{R}^n)$.

Definition 3. *A ringed space $(R^n, C^\infty(R^n))$ is called the* n-dimensional Cartesian space with infinitesimal operators.

One can prove the following lemma:

Lemma 4. *The mapping $H : C^\infty(\mathbb{R}^n)\to C^\infty(R^n)$ given by*

$$H(f) = \bar f \quad \text{for } f\in C^\infty(\mathbb{R}^n)$$

is an isomorphism of C^∞-rings.

2. MANIFOLDS WITH INFINITESIMAL OPERATORS

In this section, we will introduce the category of differential manifolds with infinitesimal operators using the Yoneda embedding [9]. Let $(M, C^\infty(M))$ be a differential manifold of

the dimension n. Let $\overline{M} := \mathrm{Hom}_{C^\infty}(C^\infty(M), \mathbb{R} \oplus \mathfrak{A}\varepsilon)$ be the set of all morphisms between C^∞-rings, $C^\infty(M)$ and $\mathbb{R} \oplus \mathfrak{A}\varepsilon$. Every morphism $\rho : C^\infty(M) \to \mathbb{R} \oplus \mathfrak{A}\varepsilon$ can be uniquely represented as a sum

$$\rho = \chi + v_\chi, \tag{1}$$

where $\chi : C^\infty(M) \to \mathbb{R}$ and $v_\chi : C^\infty(M) \to \mathfrak{A}\varepsilon$. Every morphism $\rho : C^\infty(M) \to \mathbb{R} \oplus \mathfrak{A}\varepsilon$ satisfies the following conditions:

$$\rho(f \cdot g) = \rho(f) \cdot \rho(g), \quad \rho(kf + g) = k\rho(f) + \rho(g) \tag{2}$$

for any $f, g \in C^\infty(M)$, $k \in \mathbb{R}$.

Lemma 5. *An arbitrary morphism $\rho \in \overline{M}$ can be uniquely represented as a sum:*

$$\rho = \mathrm{ev}_x + v_\chi, \tag{3}$$

where $\mathrm{ev}_x : C^\infty(M) \to \mathbb{R}$ is an evaluation of $C^\infty(M)$ at some point $x \in M$ and $v_\chi : C^\infty(M) \to \mathfrak{A}\varepsilon$ is a derivation at x with values in $\mathfrak{A}\varepsilon$.

Proof. It follows from (1) and (2) that

$$\chi(fg) + v_\chi(fg) = (\chi(f) + v_\chi(f)) \cdot (\chi(g) + v_\chi(g))$$

for any $f, g \in C^\infty(M)$. Therefore,

$$\chi(fg) + v_\chi(fg) = \chi(f) \cdot \chi(g) + v_\chi(f) \cdot \chi(g) + \chi(f) \cdot v_\chi(g)$$

for any $f, g \in C^\infty(M)$. Hence, $\chi(fg) = \chi(f) \cdot \chi(g)$ and $v_\chi(fg) = v_\chi(f) \cdot \chi(g) + \chi(f) \cdot v_\chi(g)$ for any $f, g \in C^\infty(M)$. In a similar way, we can prove that χ and v_χ are $\mathbb{R}$-linear mappings. Since $\chi : C^\infty(M) \to \mathbb{R}$ is a morphism of C^∞-rings, then χ is the evaluation of $C^\infty(M)$ at some point $x \in M$ (see [11],[2]). Thus, $\chi = \mathrm{ev}_x$ and $v_\chi : C^\infty(M) \to \mathfrak{A}\varepsilon$ satisfies the following equation:

$$v_\chi(fg) = v_\chi(f)g(x) + f(x)v_\chi(g) \quad \text{for any } f, g \in C^\infty(M). \qquad \square$$

Moreover, every morphism $\rho : C^\infty(M) \to \mathbb{R} \oplus \mathfrak{A}\varepsilon$ satisfies the following condition:

$$\rho(\omega(f_1, \ldots, f_n)) = \omega(\rho(f_1), \ldots, \rho(f_n)) \tag{4}$$

for any $f_1, \ldots, f_n \in C^\infty(M)$. Hence, from (1) we get

$$\chi(\omega(f_1, \ldots, f_n)) + v_\chi(\omega(f_1, \ldots, f_n)) = \omega\big(\chi(f_1) + v_\chi(f_1), \ldots, \chi(f_n) + v_\chi(f_n)\big)$$

for any $f_1, \ldots, f_n \in C^\infty(M)$. Thus, we obtain the following equality:

$$\chi(\omega(f_1, \ldots, f_n)) + v_\chi(\omega(f_1, \ldots, f_n)) = \omega(\chi(f_1), \ldots, \chi(f_n)) + \sum_{i=1}^{n} \omega'_{|i}(\chi(f_1), \ldots, \chi(f_n)) v_\chi(f_i)$$

for any $f_1,\dots,f_n \in C^\infty(M)$, or equivalently

$$\chi(\omega(f_1,\dots,f_n)) = \omega(\chi(f_1),\dots,\chi(f_n)) \quad \text{and}$$
$$v_\chi(\omega(f_1,\dots,f_n)) = \sum_{i=1}^{n} \omega'_{|i}(\chi(f_1),\dots,\chi(f_n))v_\chi(f_i)$$

for any $f_1,\dots,f_n \in C^\infty(M)$. The first equality means that χ is a morphism of C^∞-rings. In that case, $\chi = \mathrm{ev}_x$ for some $x \in M$. The second equality means that $v_\chi : C^\infty(M) \to \mathfrak{A}\varepsilon$ is a derivation at x and satisfies

$$v_\chi(\omega(f_1,\dots,f_n)) = \sum_{i=1}^{n} \frac{\partial \omega}{\partial x_i}(f_1(x),\dots,f_n(x))v_\chi(f_i)$$

for any $f_1,\dots,f_n \in C^\infty(M)$.

Definition 6. *Using the decomposition* (3) *in Lemma 4, we can define the projection* $\pi_M : \overline{M} \to M$ *by*

$$\pi_M(\rho) = x. \tag{5}$$

The mapping $\pi_M : \overline{M} \to M$ is a bundle of the derivations of the C^∞-ring $C^\infty(M)$ at points of M with values in $\mathfrak{A}\varepsilon$. Let us denote by $\mathrm{Der}_x(C^\infty(M),\mathfrak{A}\varepsilon)$ the set of all such derivations at x. Of course, the fiber $\pi_M^{-1}(x) = \mathrm{Der}_x(C^\infty(M),\mathfrak{A}\varepsilon)$. Let $\mathrm{top}M$ be the weakest topology in which all functions from $C^\infty(M)$ are continuous. We define the topology on $\overline{M}$ as the family of the sets $\{\overline{U} : U \in \mathrm{top}M\}$, where $\overline{U} = \mathrm{Hom}_{C^\infty}(C^\infty(U),\mathbb{R}\oplus\mathfrak{A}\varepsilon)$. It is the weakest topology in which the projection $\pi_M : \overline{M} \to M$ is continuous. Of course, $\overline{U} = \pi_M^{-1}(U)$ for $U \in \mathrm{top}M$.

For any $f \in C^\infty(M)$ we define $\bar{f} : \overline{M} \to R$ by

$$\bar{f}(\rho) = \rho(f) \quad \text{for } \rho \in \overline{M},$$

and $C^\infty(\overline{M}) := \{\bar{f} : f \in C^\infty(M)\}$. $C^\infty(\overline{M})$ is C^∞-ring with the operation

$$\omega(\bar{f}_1,\dots,\bar{f}_n) = \overline{\omega(f_1,\dots,f_n)}$$

for any $\omega \in C^\infty(\mathbb{R}^n)$, $f_1,\dots,f_n \in C^\infty(M)$, $n \in \mathbb{N}$.

Definition 7. *The ringed space* $(\overline{M},C^\infty(\overline{M}))$ *is called the* manifold with infinitesimal operators *associated to the differential manifold* $(M,C^\infty(M))$.

If $x : U \to V$ is a chart on M, U is open in M, V is open in $\mathbb{R}^n$, $x = (x_1,\dots,x_n)$, $x_i : U \to \mathbb{R}$ are coordinates of x, $i = 1,\dots,n$, then $\bar{x} : \overline{U} \to \overline{V}$ is a chart on $\overline{M}$ given by

$$\bar{x}(\rho) = (\rho(x_1),\dots,\rho(x_n)),$$

where $\overline{U} = \mathrm{Hom}_{C^\infty}(C^\infty(U),R)$, $\overline{V} = \bar{x}(\overline{U})$, $\overline{U}$ is open in $\overline{M}$, $\overline{V}$ is open in R^n.

We have the functor $M \mapsto \overline{M}$ from the category of differential manifolds to the category of manifolds with infinitesimal operators. If $F : M \to N$ is a smooth mapping, then the corresponding morphism $\overline{F} : \overline{M} \to \overline{N}$ is given by

$$\overline{F}(\rho) = \mathrm{ev}_{F(x)} + F_{*x} v_\chi \quad \text{for } \rho \in \overline{M}, \tag{6}$$

where $\rho = \mathrm{ev}_x + v_\chi$, $\chi = \mathrm{ev}_x$, ρ is uniquely presented as a sum, $F_{*x} v_\chi \in \mathrm{Der}_{F(x)}(C^\infty(N), \mathfrak{A}\varepsilon)$ is defined by the formula:

$$(F_{*x} v_\chi)(\beta) = v_\chi(\beta \circ F) \quad \text{for any } \beta \in C^\infty(N).$$

The following diagram:

$$\begin{array}{ccc} \overline{M} & \xrightarrow{\overline{F}} & \overline{N} \\ {\scriptstyle \pi_M}\downarrow & & \downarrow{\scriptstyle \pi_N} \\ M & \xrightarrow{F} & N \end{array}$$

commutes.

Example. *For $M = \mathbb{R}$ we have $\overline{\mathbb{R}} = \mathrm{Hom}_{C^\infty}(C^\infty(\mathbb{R}), \mathbb{R} \oplus \mathfrak{A}\varepsilon)$. Every element $\rho \in \overline{\mathbb{R}}$ can be presented as $\rho = \mathrm{ev}_x + \mathfrak{a}\varepsilon \frac{d}{dx}\big|_x$, where $x \in \mathbb{R}$ and $\mathfrak{a} \in \mathfrak{A}$ are unique for ρ. For any $f \in C^\infty(\mathbb{R})$ we define $\bar{f} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ by*

$$\bar{f}(\rho) = \rho(f) = f(x) + f'(x)\mathfrak{a}\varepsilon.$$

In Section 1 we constructed the real line R with infinitesimal operator $\mathfrak{A}\varepsilon$. The mapping $\overline{\mathbb{R}} \to R$, $\rho \mapsto x + \mathfrak{a}\varepsilon$ is a bijection and it is an isomorphism of the ringed spaces $(\overline{\mathbb{R}}, C^\infty(\overline{\mathbb{R}})) \to (R, C^\infty(R))$. We generalize this fact for the n-dimensional Cartesian space R^n.

Proposition 8. *The mapping $\Phi : \overline{\mathbb{R}^n} \to R^n$ given by*

$$\Phi(\rho) = (\rho(\pi_1), \ldots, \rho(\pi_n)),$$

where $\pi_i : \mathbb{R}^n \to \mathbb{R}$ are the projections, $i = 1, \ldots, n$, is an isomorphism of the ringed spaces $(\overline{\mathbb{R}^n}, C^\infty(\overline{\mathbb{R}^n}))$ and $(R^n, C^\infty(R^n))$.

Proof. Every ρ can be presented uniquely as $\rho = \mathrm{ev}_{\underline{x}} + \sum_{i=1}^n d_i \frac{\partial}{\partial x_i}\big|_x$, where $\underline{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $d_i \in \mathfrak{A}\varepsilon$ for $i = 1, \ldots, n$. It is easy to see that $\rho(\pi_i) = x_i + d_i$. Thus, the mapping Φ is given by the following formula $\Phi(\rho) = (x_1 + d_1, \ldots, x_n + d_n) = \underline{x} + \underline{d}$. It is evident that Φ is a bijection.

We will verify that $\Phi^* C^\infty(R^n) = C^\infty(\overline{\mathbb{R}^n})$. In fact, for any $f \in C^\infty(\mathbb{R}^n)$ we have $\bar{f} \in C^\infty(R^n)$ given by $\bar{f}(x_1 + d_1, \ldots, x_n + d_n) = f(x_1, \ldots, x_n) + \sum_{i=1}^n f'_{|i}(x_1, \ldots, x_n) d_i$. Now we consider $\Phi^* \bar{f} = \bar{f} \circ \Phi$. We will verify that $\bar{f} \circ \Phi \in C^\infty(\overline{\mathbb{R}^n})$. In fact,

$$\begin{aligned} (\bar{f} \circ \Phi)(\rho) &= \bar{f}(\Phi(\rho)) = \bar{f}(x_1 + d_1, \ldots, x_n + d_n) \\ &= f(x_1, \ldots, x_n) + \sum_{i=1}^n d_i \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = \rho(f) \end{aligned}$$

Therefore, $(\bar{f} \circ \Phi)(\rho) = \rho(f)$ for any $\rho \in \overline{\mathbb{R}^n}$. Thus, $\bar{f} \circ \bar{\Phi} \in C^\infty(\overline{\mathbb{R}^n})$. The further details of the proof are evident. □

3. DIFFERENTIAL GEOMETRY OF MANIFOLDS WITH INFINITESIMAL OPERATORS

Let $(M, C^\infty(M))$ be a differential manifold of dimension n, and let $(\overline{M}, C^\infty(\overline{M}))$ be the manifold with infinitesimal operators associated to M. Now, we prove the following proposition:

Proposition 9. *The mapping $J : C^\infty(M) \to C^\infty(\overline{M})$ given by*

$$J(f) = \bar{f} \quad \text{for } f \in C^\infty(M) \tag{7}$$

is an isomorphism of C^∞-rings.

Proof. Let $f, g \in C^\infty(M)$. One can see the following implication:

$$\bar{f} = \bar{g} \implies f = g.$$

Indeed, $\bar{f}(\rho) = f(\rho) + v_\chi(f)$ and $\bar{g}(\rho) = g(\rho) + v_\chi(g)$ for any $\rho \in \mathrm{Hom}_{C^\infty}(C^\infty(M), \mathbb{R} + \mathfrak{A}\varepsilon)$ with $\chi = \mathrm{ev}_p$, $f(p), g(p) \in \mathbb{R}$, $v_p(f), v_p(g) \in \mathfrak{A}\varepsilon$ and $p \in M$. Here, we have used the well-known fact that the only real-valued C^∞-morphism going from $C^\infty(M)$ is the evaluations (see [11]). Thus, $f = g$ and the mapping J is a bijection satisfying:

$$J(\omega(f_1, \ldots, f_n)) = \omega(J(f_1), \ldots, J(f_n))$$

for any $\omega \in C^\infty(\mathbb{R}^n)$, $f_1, \ldots, f_n \in C^\infty(M)$. Therefore, J is an isomorphism of C^∞-rings. □

Corollary 10. *The $C^\infty(M)$-module of derivations* $\mathrm{Der}(C^\infty(M))$ *is isomorphic to the $C^\infty(\overline{M})$-module of derivations* $\mathrm{Der}(C^\infty(\overline{M}))$.

Proof. For any $X \in \mathrm{Der}(C^\infty(M))$ we define $\overline{X} \in \mathrm{Der}(C^\infty(\overline{M}))$ by

$$\overline{X}(\bar{f}) = \overline{X(f)} \quad \text{for } f \in C^\infty(M).$$

It is easy to see the implication

$$\overline{X} = \overline{Y} \implies X = Y \quad \text{for any } X, Y \in \mathrm{Der}(C^\infty(M))$$

and $I(X) = J \circ X \circ J^{-1}$. Therefore, the mapping $I : \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(\overline{M}))$, given by

$$I(X) = \overline{X} \quad \text{for } X \in \mathrm{Der}(C^\infty(M)), \tag{8}$$

is an isomorphism of modules. □

The isomorphism J allows us to construct differential geometry on manifolds with infinitesimal operators.

Definition 11. *For any linear connection $\nabla : \mathrm{Der}(C^\infty(M)) \times \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(M))$ we define the linear connection $\overline{\nabla} : \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \to \mathrm{Der}(C^\infty(\overline{M}))$ by*

$$\overline{\nabla}_{\overline{X}} \overline{Y} = \overline{\nabla_X Y} \quad \text{for } X, Y \in \mathrm{Der}(C^\infty(M)).$$

In a similar manner, one can extend the usual definition of any tensor on M to a tensor on the manifold $\overline{M}$ with infinitesimal operators.

For any tensor $A : \mathrm{Der}(C^\infty(M)) \times \cdots \times \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(M))$ of the type $(1,n)$, we define the tensor $\overline{A} : \mathrm{Der}(C^\infty(\overline{M})) \times \cdots \times \mathrm{Der}(C^\infty(\overline{M})) \to \mathrm{Der}(C^\infty(\overline{M}))$ by

$$\overline{A}(\overline{X}_1,\ldots,\overline{X}_n) = \overline{A(X_1,\ldots,X_n)} \quad \text{for } X_1,\ldots,X_n \in \mathrm{Der}(C^\infty(\overline{M})). \tag{9}$$

Analogously, for any tensor $B : \mathrm{Der}(C^\infty(M)) \times \cdots \times \mathrm{Der}(C^\infty(M)) \to C^\infty(M)$ of the type $(0,n)$ we define the tensor $\overline{B} : \mathrm{Der}(C^\infty(\overline{M})) \times \cdots \times \mathrm{Der}(C^\infty(\overline{M})) \to C^\infty(\overline{M})$ by

$$\overline{B}(\overline{X}_1,\ldots,\overline{X}_n) = \overline{B(X_1,\ldots,X_n)} \quad \text{for } X_1,\ldots,X_n \in \mathrm{Der}(C^\infty(\overline{M})). \tag{10}$$

There is a one-to-one correspondence between the geometric structures (tensors) on $(M, C^\infty(M))$ and the respective geometric structures (tensors) on $(\overline{M}, C^\infty(\overline{M}))$. Differential geometry on $(M, C^\infty(M))$ can be lifted to $(\overline{M}, C^\infty(\overline{M}))$ and, conversely, differential geometry on $(\overline{M}, C^\infty(\overline{M}))$ can be projected onto $(M, C^\infty(M))$. The projection of geometric notion from $\overline{M}$ onto M can be organized using the mapping $\pi_M : \overline{M} \to M$ and $J^{-1} : C^\infty(\overline{M}) \to C^\infty(M)$ or $I^{-1} : \mathrm{Der}(C^\infty(\overline{M})) \to \mathrm{Der}(C^\infty(M))$. The mappings J^{-1} and I^{-1} are linear C^∞-isomorphisms. For any derivation $\mathbf{X} \in \mathrm{Der}(C^\infty(\overline{M}))$ there exists a unique derivation $X \in \mathrm{Der}(C^\infty(M))$, $X = I^{-1}(\mathbf{X})$, such that $\overline{X} = \mathbf{X}$. Let us denote the projection of $\mathbf{X}$ by $\pi_*\mathbf{X}$. Of course, $\pi_*\mathbf{X} = X$. It is easy to see, that $(\pi_*\mathbf{X})(f) = J^{-1}(\mathbf{X}(\bar{f}))$ for $f \in C^\infty(M)$.

For any tensor $\mathbf{A} : \mathrm{Der}(C^\infty(\overline{M})) \times \cdots \times \mathrm{Der}(C^\infty(\overline{M})) \to C^\infty(\overline{M})$ we define its projection $\pi_*\mathbf{A} = A$, $A : \mathrm{Der}(C^\infty(M)) \times \cdots \times \mathrm{Der}(C^\infty(M)) \to C^\infty(M)$ by

$$A(X_1,\ldots,X_n) = J^{-1}(\mathbf{A}(\overline{X}_1,\ldots,\overline{X}_n)) \quad \text{for } X_1,\ldots,X_n \in \mathrm{Der}(C^\infty(M)).$$

Analogously, we can define the projection $\pi_*\mathbf{A} = A$ of a tensor $\mathbf{A} : \mathrm{Der}(C^\infty(\overline{M})) \times \cdots \times \mathrm{Der}(C^\infty(\overline{M})) \to C^\infty(\overline{M})$ of the same type, given by

$$\pi_*\mathbf{A}(X_1,\ldots,X_n) = I^{-1}(\mathbf{A}(\overline{X}_1,\ldots,\overline{X}_n)) \tag{11}$$

for any $X_1,\ldots,X_n \in \mathrm{Der}(C^\infty(M))$.

For a linear connection $\boldsymbol{\nabla} : \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \to \mathrm{Der}(C^\infty(\overline{M}))$ on $\overline{M}$ we define its projection $\pi_*\boldsymbol{\nabla} : \mathrm{Der}(C^\infty(M)) \times \mathrm{Der}(C^\infty(M)) \to \mathrm{Der}(C^\infty(M))$ by

$$(\pi_*\boldsymbol{\nabla})(X,Y) = I^{-1}(\overline{\nabla}(\overline{X},\overline{Y})) \quad \text{for any } X,Y \in \mathrm{Der}(C^\infty(M)).$$

The Lie bracket of the ordered pair of derivations $\mathbf{X},\mathbf{Y} \in \mathrm{Der}(C^\infty(\overline{M}))$ is the derivation $[\mathbf{X},\mathbf{Y}] := \mathbf{X}\circ\mathbf{Y} - \mathbf{Y}\circ\mathbf{X}$. It is evident that

$$\pi_*[\mathbf{X},\mathbf{Y}] = [\pi_*\mathbf{X}, \pi_*\mathbf{Y}].$$

Analogously as on real manifolds we define the torsion tensor

$$\mathbf{T} : \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \to C^\infty(\overline{M})$$

and the curvature tensor

$$\mathbf{R} : \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \to \mathrm{Der}(C^\infty(\overline{M}))$$

by

$$\mathbf{T}(\mathbf{X},\mathbf{Y}) = \nabla_{\mathbf{X}}\mathbf{Y} - \nabla_{\mathbf{Y}}\mathbf{X} - [\mathbf{X},\mathbf{Y}]$$

and $\mathbf{R}(\mathbf{X},\mathbf{Y})\mathbf{Z} = \nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\mathbf{Z} - \nabla_{\mathbf{Y}}\nabla_{\mathbf{X}}\mathbf{Z} - \nabla_{[\mathbf{X},\mathbf{Y}]}\mathbf{Z}$ for any $\mathbf{X},\mathbf{Y},\mathbf{Z} \in \mathrm{Der}(C^\infty(M))$.

It is evident that the torsion T and the curvature R of the projection $\pi_*\mathbf{\nabla} = \nabla$ are the respective projections $T = \pi_*\mathbf{T}$, $R = \pi_*\mathbf{R}$. Analogously, if we consider the lift $\overline{\nabla}$ on $\overline{M}$ of a connection ∇ on M, the lifts of the torsion T and the curvature R of ∇ are the torsion $\mathbf{T}$ and the curvature $\mathbf{R}$ of $\mathbf{\nabla} := \overline{\nabla}$.

If $g : \mathrm{Der}(C^\infty(M)) \times \mathrm{Der}(C^\infty(M)) \to C^\infty(M)$ is a semi-Riemannian metric on M, we can consider the lift $\bar{g} : \mathrm{Der}(C^\infty(\overline{M})) \times \mathrm{Der}(C^\infty(\overline{M})) \to C^\infty(\overline{M})$ on $\overline{M}$. If ∇ is the Levi–Civita connection of g, then $\overline{\nabla}$ is the Levi–Civita connection of $\bar{g}$. The torsion and the curvature of $\overline{\nabla}$ are the lifts $\overline{T}$ and $\overline{R}$.

If (M,g) is a space-time, one can consider on M the Einstein equation

$$\mathrm{Ric} + \tfrac{1}{2}\Re g + \Lambda g = 8\pi\mathbb{T},$$

where Λ is the cosmological constant, Ric is the Ricci curvature (a symmetric $(0,2)$ tensor), $\Re \in C^\infty(M)$ is the scalar curvature and $\mathbb{T}$ is the energy-momentum tensor.

One can lift the Einstein equation on M to the manifold $\overline{M}$ with the infinitesimal operators:

$$\overline{\mathrm{Ric}} + \tfrac{1}{2}\overline{\Re}\bar{g} + \Lambda\bar{g} = 8\pi\overline{\mathbb{T}}.$$

We have obtained Einstein equation on $\overline{M}$. $(\overline{M},\bar{g})$ is called the *space-time with infinitesimal operators*.

4. POSITION AND MOMENTUM OPERATORS

Let us consider the real line R with the infinitesimal operators $R = \mathbb{R} \oplus \mathcal{U}\varepsilon$. Since the algebra $\mathfrak{A}$ contains the unit element $\mathbb{1}$, we can consider a C^∞-subalgebra $\mathcal{R}$ of R, composed of elements of the form $x + a\mathbb{1}\varepsilon$, $x \in \mathbb{R}$, $a \in \mathbb{R}$. So, $\mathcal{R} = \mathbb{R} \oplus \mathbb{R}\varepsilon$, and it is the ring of dual numbers. It is clear that $(\mathcal{R}^n, C^\infty(\mathcal{R}^n))$ is a subspace of n-dimensional Cartesian space with infinitesimal operators $(R^n, C^\infty(R^n))$. For any function $f \in C^\infty(\mathbb{R}^n)$ we have defined $\bar{f} : \mathcal{R}^n \to \mathcal{R}$ by

$$\bar{f}(r_1,\ldots,r_n) = f(x_1,\ldots,x_n) + \sum_{i=1}^{n} f'_{|i}(x_1,\ldots,x_n)a_i\varepsilon,$$

where $r_i = x_i + a_i\varepsilon$, $x_i \in \mathbb{R}$, $a_i \in \mathbb{R}$, $i = 1,\ldots,n$.

Now we consider interesting examples of operators, namely the operators of position $\overline{Q}_i : C^\infty(\mathcal{R}^n) \to C^\infty(\mathcal{R}^n)$, $i = 1,\ldots,n$, and the operators of momentum $\overline{P}_j : C^\infty(\mathcal{R}^n) \to C^\infty(\mathcal{R}^n)$, $j = 1,\ldots,n$. Let us recall that Euclidean operators of position and momentum in the space $C^\infty(\mathbb{R}^n)$ are defined by

$$\begin{aligned} Q_i &: C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n), \qquad (Q_i\psi)(x) = x_i\psi(x), \\ P_j &: C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n), \qquad (P_j\psi)(x) = -\psi'_{|j}(x) \end{aligned}$$

for $\psi \in C^\infty(\mathbb{R}^n)$, $x = (x_1,x_2,\ldots,x_n)$ (The last one is the real version of quantum mechanical operator of momentum.).

Now let us define

$$(\overline{Q}_i\overline{\psi})(r) = (\overline{Q_i\psi})(r), \qquad (\overline{P}_j\overline{\psi})(r) = (\overline{P_j\psi})(r)$$

for $\overline{\psi} \in C^\infty(\mathcal{R}^n)$, $r = (r_1, \ldots, r_n)$, $r_i = x_i + a_i\varepsilon$, $x_i, a_i \in \mathbb{R}$, $i, j = 1, \ldots, n$, or explicitly:

$$(\overline{Q}_i\overline{\psi})(r) = x_i\psi(x) + \psi(x)a_i\varepsilon + x_i\sum_{k=1}^{n}\psi'_{|k}(x)a_k\varepsilon,$$

$$(\overline{P}_j\overline{\psi})(r) = -\psi'_{|j} - \sum_{k=1}^{n}\psi''_{|jk}a_k\varepsilon.$$

Proposition 12. *The operators of position and momentum satisfy the following commutation relations:*

1° $[\overline{Q}_i, \overline{Q}_j] = 0$,

2° $[\overline{P}_i, \overline{P}_j] = 0$,

3° $[\overline{Q}_i, \overline{P}_j] = \delta_{ij}\mathrm{id}_{C^\infty(\mathcal{R}^n)}$

for $j = 1, \ldots, n$.

Proof. The first and second part are easily seen. We will check the last one. We have

$$\begin{aligned}(\overline{Q}_i\overline{P}_j\overline{\psi})(r) &= x_i(P_j\psi)(x) + (P_j\psi)(x)a_i\varepsilon + x_i\sum_{k=1}^{n}(P_j\psi)'_{|k}a_k\varepsilon\\ &= x_i(P_j\psi)(x) + (P_j\psi)(x)a_i\varepsilon + x_i\sum_{k=1}^{n}(-\psi''_{|jk})(x)a_k\varepsilon\\ &= -x_i\psi'_{|j}(x) - \psi'_{|j}(x)a_j\varepsilon + x_i\Big(\sum_{k=1}^{n}-\psi''_{|jk}(x)a_k\varepsilon\Big).\end{aligned}$$

On the other hand,

$$\begin{aligned}(\overline{P}_j\overline{Q}_i\overline{\psi})(r) &= -(x_i\psi)'_{|j}(x) - \sum_{k=1}^{n}(x_i\psi)''_{|jk}(x)a_k\varepsilon\\ &= \delta_{ij}\psi(x) - x_i\psi'_{|j}(x) - \sum_{k=1}^{n}\delta_{ij}\psi'_{|k}(x)a_k\varepsilon - \sum_{k=1}^{n}\delta_{ik}\psi'_{|j}(x)a_k\varepsilon - \sum_{k=1}^{n}x_i\psi''_{|jk}(x)a_k\varepsilon\\ &= -\delta_{ij}\psi(x) - x_i\psi'_{|j}(x) - \delta_{ij}\psi'_{|k}(x)a_k\varepsilon - \psi'_{|j}a_i\varepsilon - \sum_{k=1}^{n}x_i\psi''_{|jk}a_k\varepsilon.\end{aligned}$$

Computing the commutator of $\overline{Q}_i$ and $\overline{P}_j$ we obtain

$$[\overline{Q}_i, \overline{P}_j]\overline{\psi}(r) = \overline{Q}_i\overline{P}_j\overline{\psi}(r) - \overline{P}_j\overline{Q}_i\overline{\psi}(r) = \delta_{ij}\psi(x) + \delta_{ij}\sum_{k=1}^{n}\psi'_{|k}a_k\varepsilon = \delta_{ij}\overline{\psi}(r). \qquad \square$$

Thus, we obtain the commutation relations analogous to the classic Weyl–Heisenberg commutation relation known in quantum mechanics.

References

[1] N. Aronszajn, *Subcartesian and subriemannian spaces*, Notices. Amer. Math. Soc. 14 (1967), 111 pp.

[2] M. J. Cukrowski, Z. Pasternak-Winiarski and W. Sasin, *On real-valued homomorphisms in countably generated differential structures*, Demonstratio Math. 45 (2012), 665–676.

[3] E. J. Dubuc, *Sur les modèles de la géométrie différentielle synthétique*, Cah. Topol. Géom. Différ. 20 (1979), 231–279.

[4] M. Heller and W. Sasin, *Structured spaces and their application to relativistic physics*, J. Math. Phys. 36 (1995), 3644–3662.

[5] M. Heller, L. Pysiak and W. Sasin, *Noncommutative Unification of General Relativity and Quantum Mechanics*, J. Math. Phys. 46 (2005), 122501, 15 pp.

[6] A. Kock, *Synthetic differential geometry*, London Math. Soc. Lecture Note Ser. 51, Cambridge Univ. Press, Cambridge, 1981.

[7] A. Kock, *Synthetic geometry of manifolds*, Cambridge Tracts in Math. 180, Cambridge Univ. Press, Cambridge, 2010.

[8] R. Lavendhomme, *Basic concepts of synthetic differential geometry*, Kluwer, Dordrecht, 1996.

[9] I. Moerdijk and G. E. Reyes, *Smooth spaces versus continuous spaces in models for synthetic differential geometry*, J. Pure Appl. Algebra 32 (1984), 143–176.

[10] J. A. Navarro González, J. B. Sancho de Salas, *C^∞-differentiable spaces*, Lecture Notes in Math. 1824, Springer, Berlin, 2003.

[11] J. Nestruev, *Smooth manifolds and observables*, Grad. Texts in Math. 220, Springer, New York, 2003.

[12] B. O'Neill, *Semi-Riemannian geometry. With applications to relativity*, Pure Applied Math. 103, Academic Press, New York, 1983.

[13] R. S. Palais, *Real algebraic differential topology. Part I*, Math. Lecture Ser. 10. Publish or Perish, Inc., Wilmington, Del., 1981.

[14] L. Pysiak, *Time flow in a noncommutative regime*, Internat. J. Theoret. Phys. 46 (2007), 17–31.

[15] L. Pysiak, W. Sasin, M. Heller and T. Miller, *Functorial Differential Spaces and the Infinitesimal Structure of Space-Time*, Rep. Math. Phys. 85 (2020), 443–454.

[16] R. Sikorski, *Abstract covariant derivative*, Colloq. Math. 18 (1967), 251–272.

[17] R. Sikorski, *Differential modules*, Colloq. Math. 24 (1971), 45–79.

[18] K. Spallek, *Differenzierbare Räume*, Math. Ann. 180 (1969), 269–296.

Radosław Pytlak[1], **Damian Suski**[2]

[1] Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
[2] Institute of Automatic Control and Robotics,
Warsaw University of Technology, Warsaw, Poland

MINIMUM TIME CONTROL PROBLEM OF HYBRID SYSTEMS

Manuscript received: 15 September 2020
Manuscript accepted: 21 September 2020

Abstract: The aim of the paper is to provide sensitivity analysis for a minimum time control problem of hybrid systems. The analysis is established with the help of linearized equations and is also expressed by adjoint equations to systems equations. Both the linearized and the adjoint equations exhibit jumps at points in which a hybrid system changes a discrete state. On the basis of linearized equations a globally convergent algorithm is proposed. It is shown that any accumulation point of a sequence generated by the algorithm satisfies the weak maximum principle for minimum time control problem of hybrid systems.
Keywords: hybrid system, minimum time control problem, necessary optimality conditions
Mathematics Subject Classification (2020): 49J15, 49K15, 65K10, 34K34

1. INTRODUCTION

Hybrid systems are systems with mixed discrete-continuous dynamics ([22],[2]). The set of discrete states $\mathcal{Q}$ consists of a finite number of elements denoted by q. The admissible controls set $\mathcal{U}$ consists of control functions $u : I \to \Omega$ defined on a closed interval I with the values in $\Omega \in \mathbb{R}^m$. The continuous dynamics in each discrete state is described by ordinary differential equations (ODEs)

$$x' = f(x,u) \tag{1}$$

or more generally by differential-algebraic equations (DAEs)

$$0 = F(x',x,u), \tag{2}$$

where $x \in \mathbb{R}^n$, $f : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$, $F : \mathbb{R}^n \times \mathbb{R}^n \times \Omega \to \mathbb{R}^n$. The transitions between discrete states are triggered when the condition of the form $h(x) \leq 0$ stops to be satisfied, where $h : \mathbb{R}^n \to \mathbb{R}$. The functions $h(x)$ are called *guards* (or *switching functions*). In this paper the analysis is restricted to systems with autonomous transitions and without state jumps during transitions.

The optimal control problem with a hybrid system has been considered in many papers. The necessary optimality conditions for a class of hybrid systems without state jumps have been first formulated in [23]. In [3] the variational methods have been used to formulate adjoint equations for systems with state jumps. The Pontryagin maximum principle for hybrid systems with state jumps has been formulated for several classes of hybrid systems in [18], [15], [16], [17], [19], [20], [21], [7]. In papers [15], [16], [17], [19], [21] also algorithms based on the hybrid maximum principle are discussed. In [8] time optimal hybrid maximum principle is considered. Our paper does not consider optimal control problems in which switching times are decision variables ([14]).

In [10] an algorithm for optimal control problems with hybrid systems described by higher index DAEs has been introduced. Therein it is assumed that hybrid systems are described by discrete time state equations resulting from the discretization of system equations by an implicit Runge–Kutta method.

In none of these papers an optimal control problem with hybrid systems exhibiting sliding modes has been considered. In [11] some preliminary results on optimal control problems with sliding modes are given. The aim of this paper is to provide results on trajectory sensitivity analysis of hybrid systems in such a way that they could be used to construct algorithms for optimal control problems described by hybrid systems. The special attention is paid to minimum time control problems for which first order method is proposed. It is shown how the sensitivity analysis can be used to establish global convergence of the method and to derive necessary optimality conditions (in the form of the weak maximum principle) for these problems.

In order to introduce hybrid systems, consider a hybrid system with two discrete states collected in a set $\mathcal{Q} = \{1,2\}$ and assume that the transition from a discrete state $q = 1$ to $q = 2$ is triggered when $h(x) \leq 0$ stops to be satisfied, where $h : \mathbb{R}^n \to \mathbb{R}$. The transition from a discrete state $q = 2$ to $q = 1$ is triggered when $h(x) \geq 0 \Leftrightarrow -h(x) \leq 0$ stops to be satisfied. The border surface

$$\Sigma = \{x \in \mathbb{R}^n : h(x) = 0\} \tag{3}$$

is called the *switching surface*.

If the hybrid system starts its evolution from a discrete state $q = 1$ the continuous state evolves according to an equation

$$x' = f_1(x,u).$$

At a transition time t_t the continuous state trajectory reaches the switching surface so the following holds

$$h(x(t_t)) = 0. \tag{4}$$

The first order condition which guarantees that the continuous state trajectory will cross the switching surface Σ is ([6])

$$h_x(x(t_t))f_1(x(t_t),u(t_t)) > 0 \tag{5}$$

where $h_x^T(x)$ is the normal vector to Σ at x. If at a transition time

$$h_x(x(t_t))f_2(x(t_t),u(t_t)) > 0 \tag{6}$$

then the discrete state changes from $q = 1$ to $q = 2$ and the continuous state continues the evolution according to the equation $x' = f_2(x,u)$. If at a transition time

$$h_x(x(t_t))f_2(x(t_t),u(t_t)) < 0 \tag{7}$$

then both vector fields $f_1(x,u)$ and $f_2(x,u)$ point towards the surface Σ and the *sliding motion* phenomenon occurs ([4]). We assume that for any control u the hybrid system does not exhibit sliding motion. It means that at a switching time t_t either (5) and (6) hold, or

$$h_x(x(t_t))f_2(x(t_t),u(t_t)) < 0 \tag{8}$$
$$h_x(x(t_t))f_1(x(t_t),u(t_t)) < 0 \tag{9}$$

are satisfied (in this case the discrete state changes from $q = 2$ to $q = 1$). The paper [12] discusses the general case of a hybrid system which can have sliding motions.

For the simplicity of presentation we discuss optimal control problems with hybrid systems which can only have two discrete states.

2. TRAJECTORY SENSITIVITY ANALYSIS

Taking into account the considerations and definitions presented in the previous section, the optimal control problem of interest—($\mathbf{P}'$), can be defined as follows:

$$\min_{u,t_f} \phi(x(t_f),t_f) \tag{10}$$

subject to the constraints

$$\begin{array}{ll} x' = f_1(x,u) & \text{if } q = 1 \\ x' = f_2(x,u) & \text{if } q = 2 \end{array} \tag{11}$$

and the terminal constraints

$$g_i^1(x(t_f)) = 0 \ \forall i \in E \tag{12}$$
$$g_j^2(x(t_f)) \leq 0 \ \forall j \in I. \tag{13}$$

It is assumed that the initial state $x(0) = x_0$ is fixed and E, I are finite sets of indices. Notice that when $\phi(x(t_f), t_f) = t_f$ we have a minimum time control problem.

Before stating the set of admissible controls we do time transformation in order to 'separate' final time t_f, as a decision variable, from control variables u defined on the normalized horizon $[0,1]$. We can achieve that by taking the time transformation

$$[0, t_f] \ni t \to \tau \in [0,1] : \ \tau = \frac{t}{t_f}. \tag{14}$$

Since we have $d\tau = \frac{dt}{t_f}$ and the control problem $(\mathbf{P}')$ becomes

$$\min_{u, t_f} \phi(x(1), t_f) \tag{15}$$

subject to the constraints

$$\begin{array}{ll} x' = t_f f_1(x,u) & \text{if } q = 1 \\ x' = t_f f_2(x,u) & \text{if } q = 2 \end{array} \tag{16}$$

and the terminal constraints

$$g_i^1(x(1)) = 0 \ \forall i \in E \tag{17}$$

$$g_j^2(x(1)) \leq 0 \ \forall j \in I. \tag{18}$$

For this formulation of our optimal control problem we can introduce the set of admissible controls. We assume that u belongs to the set

$$\mathcal{U} = \{u \in \mathcal{L}_m^1[0,1] : \ u(t) \in \Omega, \text{a. e. on } [0,1] = T\} \tag{19}$$

where Ω is a closed convex set in $\mathbb{R}^m$. Furthermore, since the minimum time control problem is considered, the following is postulated

$$t_f \in [t_f^{\min}, t_f^{\max}] = \mathcal{T}_f. \tag{20}$$

The above optimal control problem (15)–(20) we call the problem $(\mathbf{P})$.

The considered control problem can be expressed as an optimization problem over the set of control functions and the set of parameters in $\mathbb{R}$ with the aid of the functions $\bar{F}_0 : \mathcal{U} \times \mathcal{T}_f \to \mathbb{R}$, $\bar{g}_i^1 : \mathcal{U} \times \mathcal{T}_f \to \mathbb{R}$ for $i \in E$, $\bar{g}_j^2 : \mathcal{U} \times \mathcal{T}_f \to \mathbb{R}$ for $j \in I$:

$$\begin{aligned} \bar{F}_0(u, t_f) &= \phi(x^{u,t_f}(1), t_f) \\ \bar{g}_i^1(u, t_f) &= g_i^1(x^{u,t_f}(1)) \ \forall i \in E \\ \bar{g}_j^2(u, t_f) &= g_j^2(x^{u,t_f}(1)) \ \forall j \in I, \end{aligned}$$

provided that x is the unique function of u and t_f, so one can write x^{u,t_f}.

The reformulated problem is

$$\min_{u \in \mathcal{U}, \ t_f \in \mathcal{T}_f} \bar{F}_0(u, t_f) \tag{21}$$

subject to

$$\begin{aligned} \bar{g}_i^1(u,t_f) &= 0 \ \forall i \in E \quad (22)\\ \bar{g}_j^2(u,t_f) &\leq 0 \ \forall j \in I. \quad (23) \end{aligned}$$

The parameter t_f can be treated as a constant function: $u_{m+1}(t) \equiv t_f$ on $[0,1]$ so we can define the extended admissible set

$$\mathcal{U}^e = \left\{ u_e = (u, u_{m+1}) \in \mathcal{L}^1_{m+1}[0,1] : \ u(t) \in \Omega \text{ a. e. on } [0,1], \right. \\ \left. u_{m+1}(t) \in \mathcal{T}_f, \ u_{m+1} \equiv const \right\}. \quad (24)$$

The following notation will be needed to formulate some relations (the presentation is for function u): when u is such that its limits, at time t_t, and stated below exist then

$$u(t_t^-) = \lim_{t \to t_t, t < t_t} u(t), \ \ u(t_t^+) = \lim_{t \to t_t, t > t_t} u(t).$$

For example, when the hybrid system changes its discrete state from $q = 1$ to $q = 2$, the transition conditions can be stated as

$$\begin{aligned} h_x(x(t_t^-)) f_1(x(t_t^-), u(t_t^-)) &> 0 \quad (25)\\ h_x(x(t_t^+)) f_2(x(t_t^+), u(t_t^+)) &> 0 \quad (26) \end{aligned}$$

The notation, as applied in (25)–(26), simplifies presentation of some papers results.

After introducing the extended control u_e we can reformulated the equations for the hybrid system in its discrete states

$$\begin{array}{ll} x' = t_f f_1(x,u) = f_1^e(x,u_e) & \text{if } q = 1 \\ x' = t_f f_2(x,u) = f_2^e(x,u_e) & \text{if } q = 2 \end{array}. \quad (27)$$

This representation of a continuous behaviour of a hybrid system in a discrete state facilitates the derivation of sensitivity results. Furthermore, it enables us to use control functions defined on the fixed interval T as admissible controls (and for that reason this time transformation is used in computational methods aimed at solving minimum time control problems–[9]). However, final sensitivity analysis, adjoint equations and necessary optimality conditions will be presented by referring directly to controls u and the time parameter t_f.

Our approach to sensitivity of the considered problem heavily relies on the results stated in [9] (see also [13]). Therein, the sensitivity analysis is based on the linearized equations to system equations.

Suppose that our system of interest is as follows

$$x' = f^e(x, u_e), \quad (28)$$

where $x(0) = x_0$ (fixed) and u_e and f_e are defined as in (24), (27), with the help of function $f^e : \mathbb{R}^n \times \mathbb{R}^{m+1} \to \mathbb{R}^n$, the time horizon is T.

If we denote by x^{u_e} a solution to the equations for a given $u_e \in \mathcal{U}^e$ and by $x^{u_e+d_e}$ a solution for a perturbation d_e such that $u_e + d_e \in \mathcal{U}^e$ then $x^{u_e+d_e}$ can be approximated by a solution y^{u_e,d_e} to the linearized equations

$$\begin{aligned} y' &= (f^e)_x(x,u_e)y + (f^e)_{u_e}(x,u_e)d_e && (29)\\ y(0) &= 0. && (30)\end{aligned}$$

In particular, the following hold

$$\begin{aligned} \|x^{u_e}\|_{\mathcal{L}^\infty} &\le c_1 && (31)\\ \|x^{u_e+d_e} - x^{u_e}\|_{\mathcal{L}^\infty} &\le c_2\|d_e\|_{\mathcal{L}^2} && (32)\\ \|y^{u_e,d_e}\|_{\mathcal{L}^\infty} &\le c_3\|d_e\|_{\mathcal{L}^1} && (33)\end{aligned}$$

for some positive constants c_1, c_2, c_3. Furthermore, there exists a function $o : [0,\infty) \to [0,\infty)$ such that $\lim_{s\to 0^+} o(s)/s = 0$ and

$$\|x^{u_e+d_e} - \left(x^{u_e} + y^{u_e,d_e}\right)\|_{\mathcal{L}^\infty} \le o(\|d_e\|_{\mathcal{L}^\infty}) \quad (34)$$

(Propositions 1.1, 1.3 in [9]).

The linearized equations (29) can be expressed in terms of f, u and t_f:

$$y' = t_f f_x(x,u)y + t_f f_u(x,u)d + f(x,u)d_t, \quad (35)$$

where $d_e = (d,d_t)$, $u+d \in \mathcal{U}$, $t_f + d_t \in \mathcal{T}_f$.

As shown in [9] the relations (31)–(34) hold provided that the assumption **(H1)** is satisfied:

(**H1**) Ω is convex and a compact set. $\mathcal{T}_f = [t_f^{\min}, t_f^{\max}]$ is such that $0 < t_f^{\min} < t_f^{\max} < \infty$. $f(\cdot,\cdot)$ is differentiable, f, f_x and f_u are continuous and there exists $K < \infty$ such that

$$\|f_x(x,u)\| \le K \text{ for all } (x,u) \in \mathbb{R}^n \times \Omega. \quad (36)$$

In order to simplify the notation, for a given u_e, x is written instead of x^{u_e}, x^{d_e} instead of $x^{u_e+d_e}$ and y^{d_e} instead of y^{u_e,d_e}.

Relations (31)–(34) can be used to provide first order approximations to the functionals in the problem **(P)**. Indeed, one can show (under certain assumptions specified later) that the expressions

$$\begin{aligned} \langle \nabla \bar{F}_0(u_e), d_e\rangle &= \phi_x(x(1),t_f)y^{d_e}(1) + \phi_{t_f}(x(1),t_f)d_t && (37)\\ \langle \nabla \bar{g}_i^1(u_e), d_e\rangle &= \left(g_i^1\right)_x(x(1))y^{d_e}(1),\ i \in E && (38)\\ \langle \nabla \bar{g}_j^2(u_e), d_e\rangle &= \left(g_j^2\right)_x(x(1))y^{d_e}(1),\ j \in I. && (39)\end{aligned}$$

estimate $\bar{F}_0(u_e+d_e) - \bar{F}_0(u_e)$, $\bar{g}_i^1(u_e+d_e) - \bar{g}_i^1(u_e)$, $i \in E$, $\bar{g}_j^2(u_e+d_e) - \bar{g}_j^2(u_e)$, $j \in I$ with accuracy $o(\|d_e\|_{\mathcal{L}^\infty})$ where o is a function such that $o : [0,\infty) \to [0,\infty)$ and $\lim_{s\to 0^+} o(s)/s = 0$.

We will use these estimates to construct a globally convergent algorithm for solving the problem **(P)**.

The estimates can be applied both to system equations in the discrete state $q = 1$ and in the discrete state $q = 2$ provided that the hypothesis **(H1)** applies to the function f_1 and to the function f_2 and that switching times t_t are fixed.

However, the switching times are determined by controls u and the parameter t_f since switching times are the results of intersecting state trajectory x^{u_e} with the switching surface. Switching times are thus functions of extended controls u_e. The consequence of that is that in order to establish linearized equations for the hybrid system on the whole horizon T we have to evaluate differentials of switching times and take them into account while deriving solutions to linearized equations considered on the whole horizon.

We assume that switching times occurring at our hybrid system are of the type when the system changes its discrete state from $q = 1$ to $q = 2$, or from $q = 2$ to $q = 1$. In other words the hybrid system does not exhibit sliding motion. It means that at switching times either relations (5)–(6), or (8)–(9) are satisfied. However, in order to provide the sensitivity analysis of 'global nature' (needed for the convergence analysis of an algorithm) we have these relations to be fulfilled in neighourhoods of switching times—these conditions are stated as the hypothesis **(H2)**.

In order to analyse the changes of switching times due to the perturbations of u_e suppose that for the extended control u_e the switching time t_t is evaluated according to the equation

$$h(x(t_t^-)) = 0.$$

When the control u_e is perturbed by d_e then a new switching time, denoted by $t_t^{d_e}$, will satisfy

$$h(x^{d_e}(t_t^{d_e-})) = 0,$$

and, in general case, $t_t^{d_e} \neq t_t$.

After the transition the state function x^d evolves on an interval $[t_t^{d_e}, 1]$ according to the equations

$$\left(x^{d_e}\right)' \;=\; f_2^e(x^{d_e}, u_e + d_e).$$

We will show that there exists the linear operator dt_t which assigns to each d_e a real number $dt_t^{d_e}$ such that the following condition (see *Theorem 1*)

$$t_t^{d_e} - t_t = dt_t^{d_e} + o(\|d_e\|_{\mathcal{L}^\infty}) \tag{40}$$

holds for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$. Here, $o : (0,\infty) \to (0,\infty)$ and $\lim_{s\to 0} s^{-1} o(s) = 0$. Moreover, $dt_t^{d_e}$ is given by the formula

$$dt_t^{d_e} = -\frac{h_x(x(t_t^-))y^{d_e}(t_t^-)}{h_x(x(t_t^-))f_1^e(x(t_t^-), u_e(t_t^-))}. \tag{41}$$

Then there exists the operator dx_- which assigns to each d_e a vector $dx_-^{d_e}$ according to the formula

$$dx_-^{d_e} = y^{d_e}(t_t^-) + f_1^e(x(t_t^-), u_e(t_t^-))dt_t^{d_e}. \tag{42}$$

According to *Theorem 1*, operator dx_- satisfies the condition

$$\|x^{d_e}(t_t^{d_e-}) - x(t_t^-) - dx_-^{d_e}\| \leq o(\|d_e\|_{\mathcal{L}^\infty}), \tag{43}$$

for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$, where $o : (0,\infty) \to (0,\infty)$ and $\lim_{s\to 0} s^{-1}o(s) = 0$.

Furthermore, from *Theorem 1* there exists the operator dx_+ which assigns to each d_e a vector $dx_+^{d_e}$ according to the formula

$$dx_+^{d_e} = y^{d_e}(t_t^+) + f_2^e(x(t_t^+), u_e(t_t^+))dt_t^d, \tag{44}$$

which satisfies the condition

$$\|x^d(t_t^{d+}) - x(t_t^+) - dx_+^d\| \leq o(\|d\|_{\mathcal{L}^\infty}) \tag{45}$$

for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$, where $o : (0,\infty) \to (0,\infty)$ and $\lim_{s\to 0} s^{-1}o(s) = 0$.

Since dx_-^d and dx_+^d are linear operators (y^{d_e} is linear and due to (41) $dt_t^{d_e}$ is also linear) and (43), (45) are satisfied, they are differential. This implies that we must have $dx_-^d = dx_+^d$ and as a consequence of that y^{d_e} can exhibit a jump at t_t:

$$y^{d_e}(t_t^+) \;=\; y^{d_e}(t_t^-) + \left[f_1^e(x(t_t^-), u_e(t_t^-)) - f_2^e(x(t_t^+), u_e(t_t^+))\right] dt_t^{d_e}, \tag{46}$$

provided that $dt_t^{d_e} \neq 0$.

The analysis above has been possible due to *Theorem 1* which establishes results concerning differentials associated with changes of switching times for a particular type of hybrid systems in which sliding motion does not occur. The proof of the theorem can be carried out in a similar way as the proof of Theorem 3.1 in [12].

Theorem 1 requires several assumptions (similar to assumptions (H2) and (H3) in [12]). The meaning of the assumption **(H2)** has already been discussed, the other assumption, **(H3)**, is needed since we explore behaviour of the hybrid system 'just before' (and 'just after') a discrete state switching and the system is controlled by u fulfilling mild restrictions, i.e., $u \in \mathcal{U}$.

(**H2**) function $h(\cdot)$ is differentiable and there exist $0 < L_1 < +\infty$ and $0 < L_2 < +\infty$ such that

$$\|h_x(\hat{x}) - h_x(x)\| \leq L_1\|\hat{x} - x\| \tag{47}$$

$$|h_x(\hat{x})f_i(\hat{x},\hat{u}) - h_x(x)f_i(x,u)| \leq L_2\|(\hat{x},\hat{u}) - (x,u)\| \tag{48}$$

for all (x,u), $(\hat{x},\hat{u})$ in $\mathbb{R}^n \times \Omega$.

Furthermore, there exists $\varepsilon > 0$ and $0 < L_3 < +\infty$ such that for all switching points t_t (each switching time corresponds to some $u_e \in \mathcal{U}^e$) and for all their perturbations $t_t^{d_e}$

triggered by perturbations d_e such that $u_e + d_e \in \mathcal{U}^e$, $\|d_e\|_{\mathcal{L}^2} \leq \varepsilon$, and for all $\theta \in [0,1]$ we have

$$h_x(x(\tau^{\theta,d_e}(t_t)))f_i(x(\tau^{\theta,d_e}(t_t)),u(\tau^{\theta,d_e}(t_t))) \geq L_3, \tag{49}$$

or

$$h_x(x(\tau^{\theta,d_e}(t_t)))f_i(x(\tau^{\theta,d_e}(t_t)),u(\tau^{\theta,d_e}(t_t))) \leq -L_3, \tag{50}$$

for $i = 1,2$, where

$$\tau^{\theta,d_e}(t_t) \;=\; t_t + \theta(t_t^{d_e} - t_t). \tag{51}$$

(**H3**) For any $u \in \mathcal{U}$ and any switching point t_t the following limits exist

$$\lim_{t \to t_t, t<t_t} u(t), \;\; \lim_{t \to t_t, t>t_t} u(t) \tag{52}$$

(and are denoted by $u(t_t^-)$ and $u(t_t^+)$ respectively).

Theorem 1. *Suppose that at the first considered discrete state the system evolution is given by*

$$x' \;=\; f_1^e(x,u_e) \tag{53}$$

and at the other by

$$x' \;=\; f_2^e(x,u_e). \tag{54}$$

We assume that functions defining systems evolution in the both discrete states satisfy **(H1)***, and if the system changes state from the first state to the other at the switching time t_t satisfying*

$$h(x(t_t)) = 0, \tag{55}$$

and the hypothesis **(H2)** *holds for equations (53)–(54) then*

$$h_x(x(t_t^-))y^{d_e}(t_t^-) + h_x(x(t_t^-))f_1^e(x(t_t^-),u_e(t_t^-))(t_t^{d_e} - t_t) + o(\|d\|_{\mathcal{L}^\infty}) = 0,$$

for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$ where o is such that $\lim_{s\to 0}|o(s)|/s = 0$, and

$$dt_t^{d_e} = -\left[h_x(x(t_t^-))y^{d_e}(t_t^-)\right] / \left[h_x(x(t_t^-))f_1^e(x(t_t^-),u_e(t_t^-))\right]. \tag{56}$$

Furthermore,

(i)

$$\left\| x_1^{d_e}(t_t^{d_e-}) - x_1(t_t^-) - y_1^{d_e}(t_t^-) - f_1^e(x(t_t^-),u_e(t_t^-))dt_t^{d_e} \right\| \leq o(\|d_e\|_{\mathcal{L}^\infty}), \tag{57}$$

for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$ where x_1, $x_1^{d_e}$ are solutions to the equation (53) and $y_1^{d_e}$ are solutions to the linearized equations associated with equations (53), here o is such that $\lim_{s\to 0}|o(s)|/s = 0$;

(ii)

$$\left\| x_2^{d_e}(t_t^{d_e+}) - x_2(t_t^+) - y_2^{d_e}(t_t^+) - f_2^e(x(t_t^+), u_e(t_t^+))dt_t^{d_e} \right\| \leq o(\|d_e\|_{\mathcal{L}^\infty}), \qquad (58)$$

for all u_e and d_e such that $u_e + d_e \in \mathcal{U}^e$ where x_2, $x_2^{d_e}$ are solutions to the equations (54) and $y_2^{d_e}$ are solutions to the linearized equations associated with equations (54), here o is such that $\lim_{s\to 0} |o(s)|/s = 0$.

The fact that solutions to linearized equations of hybrid systems can exhibit jumps (as stated in (46)) causes that relations (31)–(34) do not apply to hybrid systems. Furthermore, in the case of the considered hybrid system a trajectory x generated by a control $u_e \in \mathcal{U}^e$ will consist of pieces of trajectories of the system being in discrete state $q = 1$ (on the time intervals A_i^1, $i \in I^1$) and pieces of trajectories of the system being in the discrete state $q = 2$ (on the time intervals A_i^2, $i \in I^2$). We have $\cup_{i\in I^i} A_i^1 \cup_{i\in I^2} A_i^2 = [0,1]$.

The first issue is resolved by redefining our meaning of solutions to linearized equations—we substitute y^{d_e} by $y_{rc}^{d_e}$

$$y_{rc}^{d_e}(t) = \begin{cases} y^{d_e}(t), & \text{if } t \in [0,1],\ t \neq t_t \\ y^{d_e}(t_t^+), & \text{if } t = t_t \end{cases} .$$

The second issue, which concerns the varying sizes of subintervals A_i^1 and A_i^2 does not restrain us from using (31)–(34) in relation to hybrid systems as was shown in the proof of Theorem 3.2 presented in [12]. *Theorem 2* is Theorem 3.2 adopted to the hybrid systems considered in this paper. *Theorem 2* requires the additional assumption which postulates the finite number of switching times for any admissible control u_e.

Let $N_t(u_e)$ be the number of switching times triggered by a control $u_e \in \mathcal{U}^e$ then the hypothesis is as follows.

(H4) There exists a nonnegative integer number $I_t < +\infty$ such that $N_t(u_e) \leq I_t$, $\forall u_e \in \mathcal{U}^e$.

Theorem 2. *Suppose that x is the trajectory generated by $u_e \in \mathcal{U}^e$, x^{d_e} the trajectory generated by $u_e + d_e$ and y^{d_e} is the solution to the linearized equations induced by the perturbation d_e of u_e. Then, if* **(H1)**, **(H2)**, **(H3)** *and* **(H4)** *are satisfied, there exist positive constants c_1, c_2, c_3 and a function $o : ([0,\infty] \to (0,\infty)$ such that* $\lim_{s\to 0^+} o(s)/s = 0$ *for which the following hold*

$$\begin{aligned}
\|x\|_{\mathcal{L}^\infty} &\leq c_1 && (59)\\
\|x^{d_e} - x\|_{\mathcal{L}^\infty} &\leq c_2\|d_e\|_{\mathcal{L}^\infty} && (60)\\
\|y_{rc}^{d_e}\|_{\mathcal{L}^\infty} &\leq c_3\|d_e\|_{\mathcal{L}^1} && (61)\\
\|x^{d_e} - \left(x + y_{rc}^{d_e}\right)\|_{\mathcal{L}^\infty} &\leq o\left(\|d_e\|_{\mathcal{L}^\infty}\right), && (62)\\
\left\| x^{d_e}(t_t^{d_e-}) - x(t_t^-) - dx_-^{d_e} \right\| &\leq o(\|d_e\|_{\mathcal{L}^\infty}) && (63)\\
\left\| x^{d_e}(t_t^{d_e+}) - x(t_t^+) - dx_+^{d_e} \right\| &\leq o(\|d_e\|_{\mathcal{L}^\infty}) && (64)
\end{aligned}$$

for any $u_e \in \mathcal{U}^e$ and d_e such that $u_e + d_e \in \mathcal{U}^e$, and any switching point t_t.

Relations (59)–(64) enable us to: derive adjoint equations for hybrid systems; construct globally convergent algorithms for the problem **(P)**; state necessary optimality conditions for the considered optimal control problem. The benefits of relations (59)–(64) are illustrated by considering a first order method for solving the problem **(P)**. The method which will be discussed for the rest of the paper is analyzed in details in [9] therefore we only focus on these parts of the analysis which need attention when optimal control problems with hybrid systems are concerned.

3. GLOBALLY CONVERGENT ALGORITHM

The method we propose for solving the problem (**P**) is based on an exact penalty function. By using an exact penalty function approach, instead of solving the problem (**P**), we solve the problem ($\mathbf{P_c}$)

$$\min_{u_e \in \mathcal{U}^e} \bar{F}_c^e(u_e) \tag{65}$$

in which the exact penalty function $\bar{F}_c^e(u_e)$ is defined as follows

$$\bar{F}_c^e(u_e) = \bar{F}_0^e(u_e) + c \max\left[0, \max_{i \in E} \left|\bar{g}_i^1(u_e)\right|, \max_{j \in I} \bar{g}_j^2(u_e)\right] \tag{66}$$

For fixed c and u_e the direction finding subproblem, $\mathbf{P_c}(\mathbf{u_e})$, for the problem ($\mathbf{P_c}$) is:

$$\min_{d_e \in \mathcal{D}_{u_e}, \beta \in \mathbb{R}} \left[\langle \nabla \bar{F}_0^e(u_e), d_e \rangle + c\beta + 1/2 \|d_e\|_{\mathcal{L}^2}^2\right]$$

subject to

$$\begin{aligned} \left|\bar{g}_i^1(u_e) + \left\langle \nabla \bar{g}_i^1(u_e), d_e \right\rangle\right| &\leq \beta \quad \forall i \in E \\ \bar{g}_j^2(u_e) + \left\langle \nabla \bar{g}_j^2(u_e), d_e \right\rangle &\leq \beta \quad \forall j \in I. \end{aligned}$$

Here,

$$\mathcal{D}_{u_e} \left\{ d_e \in \mathcal{L}_{m+1}^2[T] : \; d_e \in \mathcal{U}^e - u_e \right\}.$$

The subproblem can be reformulated as an optimization problem with the objective function which is strictly convex. The problem therefore has the unique solution $(\bar{d}_e, \bar{\beta})$. Since this solution depends on c and u_e, we may define *descent function* $\sigma_c(u_e)$ and *penalty test function* $t_c(u_e)$, to be used to test optimality of a control u and to adjust c, respectively, as

$$\sigma_c(u_e) = \left\langle \nabla \bar{F}_0^e(u_e), \bar{d}_e \right\rangle + c\left[\bar{\beta} - M(u_e)\right] \tag{67}$$

and

$$t_c(u_e) = \sigma_c(u_e) + M(u_e)/c \tag{68}$$

for given $c > 0$ and $u_e \in \mathcal{U}^e$. Here,

$$M(u_e) = \max\left[0, \max_{i\in E}\left|\bar{g}_i^1(u_e)\right|, \max_{j\in I}\bar{g}_j^2(u_e)\right],$$

Our algorithm is as follows.

Algorithm. Fix parameters: $\gamma,\ \eta \in (0,1)$, $c_0 > 0$, $\varkappa > 1$.

1. Choose the initial control $u_e^0 \in \mathcal{U}^e$. Set $k = 0$, $c^{-1} = c^0$.
2. Let c^k be the smallest number chosen from $\{c^{k-1}, \varkappa c^{k-1}, \varkappa^2 c^{k-1}, \ldots\}$ such that the solution (d_e^k, β^k) to the direction finding subproblem $\mathbf{P_{c^k}}(\mathbf{u_e^k})$ satisfies

$$t_{c^k}(u_e^k) \le 0. \tag{69}$$

   If $\sigma_{c^k}(u_e^k) = 0$ then STOP.
3. Let α^k be the largest number chosen from the set $\{1, \eta, \eta^2, \ldots,\}$ such that

$$u_e^{k+1} = u_e^k + \alpha^k d_e^k$$

   satisfies the relation

$$\bar{F}_{c^k}(u_e^{k+1}) - \bar{F}_{c^k}(u_e^{)} \quad \le \quad \gamma\alpha_k\sigma_{c^k}(u_e^k). \tag{70}$$

   Increase k by one. Go to Step 2.

In order to establish convergence of *Algorithm* we need to introduce two additional hypotheses. The first one concerns the functions defining the objective and constraints:

(H5) ϕ, g_i^1, $i \in E$, g_j^2, $j \in I$ are continuously differentiable functions.

The second one is related to a constraint qualification. To this end we first introduce the set

$$\mathcal{D}^e = \left\{d_e \in \mathcal{L}_{m+1}^2[T] :\ \text{there } \exists\ u_e \in \mathcal{U}^e \text{ such that } u_e + d_e \in \mathcal{U}^e\right\}$$

and the set

$$\mathcal{F}^e(u_e) = \left\{d_e \in \mathcal{D}^e :\ \max_{j\in I}\left\langle \nabla \bar{g}_j^2(u_e), d_e\right\rangle < 0\right\}.$$

Then the constraint qualification condition takes the form

(CQ) for each $u_e \in \mathcal{U}^e$, $\mathcal{F}^e(u_e) \neq \emptyset$, and in the case $E \neq \emptyset$ we have

$$0 \in \text{interior}\mathcal{E}(u_e) \tag{71}$$

where

$$\mathcal{E}(u_e) = \left\{ \left\{ \langle \nabla \bar{g}_i^1(u_e), d_e \rangle \right\}_{i \in E} \in \mathbb{R}^{|E|} : \ d_e \in \mathcal{F}^e(u_e) \right\}.$$

Under stated assumptions *Algorithm* is globally convergent in the following sense.

Theorem 3. *Assume that data for* **(P)** *satisfies hypotheses* (**H1**), (**H2**), (**H3**), (**H4**), (**H5**) *and* (**CQ**). *Let* $\{u_e^k\}$ *be a sequence of controls generated by Algorithm and let* $\{c^k\}$ *be a sequence of the corresponding penalty parameters. Then*

i) $\{c^k\}$ *is a bounded sequence*

ii)

$$\lim_{k \to \infty} \sigma_{c^k}(u_e^k) = 0, \ \lim_{k \to \infty} M(u_e^k) = 0. \tag{72}$$

iii) Let $\bar{u}_e$ *be a* $\mathcal{L}^\infty$ *limit point of the sequence* $\{u_e^k\}$ *and* $\bar{x}$ *the trajectory corresponding to* $\bar{u}_e$, *then the necessary optimality conditions hold:*

(NC) *:*

$$0 \leq \min_{d_e \in \mathcal{D}_{\bar{u}_e}} \left[\phi_x(\bar{x}(1), t_f) y^{d_e}(1) + \phi_{t_f}(\bar{x}(1), t_f) d_t \right] \tag{73}$$

subject to the constraints

$$g_i^1(\bar{x}(1)) + \left(g_i^1\right)_x (\bar{x}(1)) y^{d_e}(1) \ = \ 0, \ i \in E \tag{74}$$

$$g_j^2(\bar{x}(1)) + \left(g_j^2\right)_x (\bar{x}(1)) y^{d_e}(1) \ \leq \ 0, \ j \in I_{0,\bar{u}_e} \tag{75}$$

together with $g_i^1(\bar{x}(1) = 0$, $i \in E$, $g_j^2(\bar{x}(1)) \leq 0$, $j \in I$. *Here,*

$$I_{\varepsilon, u_e} = \left\{ j \in I : \ \bar{g}_j^2(u_e) \geq \max_{j \in I} \bar{g}_j^2(u_e) - \varepsilon \right\}.$$

Proof. (*sketch*) Notice that the descent function $\sigma_{c^k}(u_e^k)$ is nonpositive valued at each iteration. Indeed, we have

$$\left\langle \nabla \bar{F}_0^e(u_e^k), d_e \right\rangle + c^k \beta + 1/2 \|d_e\|_{\mathcal{L}^2}^2 \leq c^k M(u^k),$$

which holds because $0 \in \mathcal{U}^e - u_e^k$. This implies that

$$\left\langle \nabla \bar{F}_0^e(u_e^k), d_e \right\rangle + c^k \left[\beta - M(u_e^k) \right] \leq -1/2 \|d_e\|_{\mathcal{L}^2}^2 \leq 0. \tag{76}$$

Algorithm generates a sequence of controls $\{u_e^k\}$ and the corresponding sequence of penalty parameters $\{c^k\}$ such that $\{c^k\}$ is bounded and any accumulation point of $\{u_e^k\}$ satisfies optimality conditions in the form of the weak maximum principle for the problem (**P**), i.e. $\sigma_{\bar{c}}(\bar{u}_e) = 0$ for the limit point $\bar{u}_e$ and the limit point of the sequence $\{c^k\}$. But $\sigma_{\bar{c}}(\bar{u}_e) = 0$ implies that $M(\bar{u}_e) = 0$ due to the definition (68) and since (69) holds.

The proof of the theorem follows the scheme of the proof of Theorem 5.1 in [9]. It is heavily based on sensitivity results stated in Propositions 1.1 and 1.3 (in the case of Theorem 5.1) and stated in *Theorem 2* (in the case of the considered theorem). The differences in these sensitivity results are not significant as far as the convergence of *Algorithm* is concerned (*Algorithm* can be applied to optimal control problems with dynamics: $x' = f(x,u)$ and it will be globally convergent according to Theorem 5.1.).

The proof of Theorem 5.1 is carried out in two steps. In the first step it is shown that under **(H1)** and **(CQ)** for any $u \in \mathcal{U}$ there exists a finite $\hat{c} > 0$ such that for $c \geq \hat{c}$ $t_c(u) \leq 0$ is satisfied. In the second step it is demonstrated that under **(H1)** and **(H5)** for any $u \in \mathcal{U}$ and $c > 0$ such that $\sigma_c(u) \leq 0$ there exists $\hat{c} > 0$ for which (70) holds for any $\alpha \in (0, \hat{\alpha})$, provided that $\sigma_c(u) < 0$. The analysis carried in these two steps provides also the justification for $\sigma_c(\bar{u}_e) = 0$, $M(\bar{u}_e) = 0$ to be used as necessary optimality conditions. □

4. ADJOINT EQUATIONS

As it is shown in [9] the implementation of an algorithm for solving the considered optimal control problem requires the evaluation of the scalar products: $\langle \nabla \bar{F}_0^e(u_e), d_e \rangle$, $\langle \nabla \bar{g}_i^1(u_e), d_e \rangle$, $i \in E$, $\langle \nabla \bar{g}_j^2(u_e), d_e \rangle$, $j \in I$.

The system can change its discrete state several times on the interval $[0,1]$. For the simplicity of presentation it is assumed that in the time interval $[0,t_t]$ the system evolves according to the equation $x' = f_1^e(x,u_e)$. At a transition time t_t the continuous state trajectory crosses the switching surface and then is determined by the equation $x' = f_2^e(x,u_e)$ up to a final time 1.

Proposition 4. *Assume that* **(H1)**, **(H2)** *and* **(H3)** *are satisfied. Suppose that the system evolves on the time interval* $[0,t_t]$ *according to the equation* $x' = f_1^e(x,u_e)$. *At a transition time* t_t *the continuous state trajectory crosses the switching surface and then is determined by the equation* $x' = f_2^e(x,u_e)$ *up to a final time* 1. *If* ϕ *is continuously differentiable with respect to its arguments then*

$$\begin{aligned}\langle \nabla \bar{F}_0^e(u_e), d_e \rangle = \big[& \phi_{t_f}(x(1), t_f) \\ & - \int_0^{t_t} \lambda_1^T(t) f_1(x(t), u(t)) d(t) dt - \int_{t_t}^1 \lambda_2^T(t) f_2(x(t), u(t)) d(t) dt \Big] d_t \\ & - \int_0^{t_t} \lambda_1^T(t) (f_1^e)_u(x(t), u_e(t)) d(t) dt - \int_{t_t}^1 \lambda_2^T(t) (f_2^e)_u(x(t), u_e(t)) d(t) dt. \end{aligned} \tag{77}$$

where λ_1, λ_2 are solutions to the adjoint equations:

$$(\lambda_1^T)'(t) = -\lambda_1^T(t)(f_1^e)_x(x(t),u_e(t)),\ t\in[0,t_t] \tag{78}$$

and

$$(\lambda_2^T)'(t) = -\lambda_2^T(t)(f_2^e)_x(x(t),u_e(t)),\ t\in[t_t,1). \tag{79}$$

with the terminal condition

$$\lambda_2(1) = -\phi_x^T(x(1),t_f). \tag{80}$$

and the jump conditions

$$\pi h_x(x(t_t))^T + \lambda_1(t_t) - \lambda_2(t_t) = 0 \tag{81}$$

$$\lambda_2(t_t)^T f_2^e(x(t_t^+),u_e(t_t^+)) - \lambda_1(t_t)^T f_1^e(x(t_t^-),u_e(t_t^-)) = 0. \tag{82}$$

where π is the number which under **(H2)** *can be evaluated from (81)–(82).*

Proof. To derive the adjoint equations the following augmented functional is constructed

$$\begin{aligned}\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi) = {}& \phi(x(1),t_f) + \pi h(x(t_t)) \\ & + \int_0^{t_t}\left[\lambda_1^T(t)\left(x'(t) - f_1^e(x(t),u_e(t))\right)\right]dt \\ & + \int_{t_t}^1 \lambda_2^T(t)\left(x'(t) - f_2^e(x(t),u_e(t))\right)dt.\end{aligned}$$

One can evaluate the variation of the augmented functional

$$\begin{aligned}d\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi) = {}& \phi_x(x(1),t_f)dx(1) + \phi_{t_f}(x(1),t_f)d_t \\ & + \pi h_x(x(t_t))dx(t_t) + \lambda_1^T(t_t)\left(x'(t_t^-) - f_1^e(x(t_t^-),u_e(t_t^-))\right)dt_t \\ & + d\left[\int_0^{t_t}\left[\lambda_1^T(t)\left(x'(t) - f_1^e(x(t),u_e(t))\right)\right]dt\right] - \lambda_2^T(t_t)\left(x'(t_t^+)\right. \\ & \left. - f_2^e(x(t_t^+),u_e(t_t^+))\right)dt_t + d\left[\int_{t_t}^1 \lambda_2^T(t)\left(x'(t) - f_2^e(x(t),u_e(t))\right)dt\right].\end{aligned}$$

By taking into account the fact that $dx(1) = y^{d_e}(1)$ and by integrating by parts the formulas $\int\lambda(t)x(t)dt$ one can obtain

$$\begin{aligned}d\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi) = {}& \phi_x(x(1),t_f)y^{d_e}(1) \\ & + \phi_{t_f}(x(1),t_f)d_t + \pi h_x(x(t_t))dx(t_t) \\ & + \lambda_1^T(t_t)\left(x'(t_t^-) - f_1^e(x(t_t^-),u_e(t_t^-))\right)dt_t \\ & + d\left[\left[\lambda_1^T(t)x(t)\right]_0^{t_t}\right] - d\left[\int_0^{t_t}\left((\lambda_1^T)'(t)x(t)\right.\right. \\ & \left.\left. + \lambda_1^T(t)\left(f_1^e(x(t),u_e(t))\right)\right)dt\right] - \lambda_2^T(t_t)\left(x'(t_t^+) - f_2^e(x(t_t^+),u_e(t_t^+))\right)dt_t \\ & + d\left[\left[\lambda_2^T(t)x(t)\right]_{t_t}^1\right] - d\left[\int_{t_t}^1\left((\lambda_2^T)'(t)x(t) + \lambda_2^T(t)f_2^e(x(t),u_e(t))\right)dt\right]\end{aligned}$$

Expanding further the variations and taking into account the initial conditions of the linearized equations the following equation is obtained

$$\begin{aligned}
d\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi) &= \phi_{t_f}(x(1),t_f)d_t + \phi_x(x(1))y^{d_e}(1)\\
&+ \pi h_x(x(t_t))dx(t_t)\\
&+ \lambda_1^T(t_t)x'(t_t^-)dt_t - \lambda_1^T(t_t)f_1^e(x(t_t^-),u_e(t_t^-))dt_t\\
&+ \lambda_1^T(t_t)y^{d_e}(t_t^-) - \int_0^{t_t}\Big((\lambda_1^T)'(t)y^{d_e}(t)\\
&+ \lambda_1^T(t)(f_1^e)_x(x(t),u_e(t))y^{d_e}(t) + \lambda_1^T(t)(f_1^e)_{u_e}(x(t),u_e(t))d_e(t)\Big)dt\\
&- \lambda_2^T(t_t)x'(t_t^+)dt_t + \lambda_2^T(t_t^+)f_2^e(x(t_t^+),u_e(t_t^+))dt_t + \lambda_2^T(1)y^{d_e}(1) - \lambda_2^T(t_t)y^{d_e}(t_t^+)\\
&- \int_{t_t}^1\Big((\lambda_2^T)'(t)y^{d_e}(t) + \lambda_2^T(t)(f_2^e)_x(x(t),u_e(t))y^{d_e}(t)\\
&+ \lambda_2^T(t)(f_2^e)_{u_e}(x(t),u_e(t))d_e(t)\Big)dt.
\end{aligned}$$

Now the formula for the differential $dx(t_t)$ is utilized and rearrangement of the components with respect to differentials $dx(t_t)$, dt_t and variations $y^{d_e}(1)$, $y^{d_e}(t)$, $d(t)$, d_t leads to

$$\begin{aligned}
d\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi) &= \phi_{t_f}(x(1),t_f)d_t + \big(\phi_x(x(1),t_f) + \lambda_2^T(1)\big)y^{d_e}(1)\\
&+ \big(\pi h_x(x(t_t)) + \lambda_1^T(t_t) - \lambda_2^T(t_t)\big)dx(t_t)\\
&+ \big(\lambda_2^T(t_t)f_2^e(x(t_t^+),u_e(t_t^+)) - \lambda_1^T(t_t)f_1^e(x(t_t^-),u_e(t_t^-))\big)dt_t\\
&- \int_0^{t_t}\Big(\big((\lambda_1)'(t) + \lambda_1^T(t)(f_1^e)_x(x(t),u_e(t))\big)y^{d_e}(t)\\
&+\lambda_1^T(t)(f_1^e)_u(x(t),u_e(t))d(t) + \lambda_1^T(t)f_1(x(t),u(t))d_t\Big)dt\\
&- \int_{t_t}^1\Big(\big((\lambda_2^T)'(t) + \lambda_2^T(t)(f_2^e)_x(x(t),u_e(t))\big)y^{d_e}(t)\\
&+\lambda_2^T(t)(f_2^e)_u(x(t),u_e(t))d(t) + \lambda_2^T(t)f_2(x(t),u(t))d_t\Big)dt.
\end{aligned}$$

Now conditions for adjoint equations are stated in such a way that the expressions with differentials $dx(t_t)$, dt_t and variations $y^{d_e}(1)$, $y^{d_e}(t)$ disappear, so eventually only the coefficients with variations $d(t)$, d_t remain.

To this end the following components have to be equal to zero

$$\begin{aligned}
&\big((\lambda_1^T)'(t) + \lambda_1^T(t)(f_1^e)_x(x(t),u_e(t))\big)y^{d_e}(t),\ t\in[0,t_t]\\
&\big((\lambda_2^T)'(t) + \lambda_2^T(t)(f_2^e)_x(x(t),u_e(t))\big)y^{d_e}(t),\ t\in[t_t,1].
\end{aligned}$$

This can be achieved by assuming that λ_1, λ_2 are solutions to the equations (78)–(79) together with the transversality condition (80).

When we zero components related to $dx(t_t)$ and dt_t we come to the equations (81)–(82). Under the assumption **(H2)** these equations can be solved with respect to π and $\lambda_1(t_t)$ to get

$$\begin{aligned}
\pi &= \frac{\lambda_2(t_t)^T\left(f_1^e(x(t_t^-),u_e(t_t^-)) - f_2^e(x(t_t^+),u_e(t_t^+))\right)}{h_x(x(t_t)f_1^e(x(t_t^-),u_e(t_t^-))}\\
\lambda_1(t_t) &= \lambda_2(t_t) - \pi h_x(x(t_t))^T.
\end{aligned}$$

Having solutions to the adjoint equations for λ_1 and λ_2 the first variation of a cost function $\phi(x(1),t_f)$, with respect to a control function variation d and d_t, we arrive at the thesis if we notice that $\left\langle \nabla \bar{F}_0^e(u_e), d_e \right\rangle = d\Phi(x,u,t_f,\lambda_1,\lambda_2,\pi)$. □

5. THE WEAK MAXIMUM PRINCIPLE

On the basis of the defined adjoint equations one can formulate the weak maximum principle for the considered problem. Suppose that $\bar{u}_e = (\bar{u},\bar{t}_f)$ is the problem solution. The weak maximum principle for the problem **(P)** can take a quite complicated form, depending on the number of switching points triggered by the optimal control $\bar{u}_e$. In order to exemplify the possible conditions stated by the weak maximum principle, we assume that there is only one switching point and at this point the system changes its discrete state from the state $q = 1$ to the state $q = 2$—we call this case as *Case 1–2*. For this case the necessary optimality conditions (**NC**$^{\mathbf{12}}$) will shape as follows (they are expressed in terms of the original problem formulation, so $\bar{t}_t$ is the switching time evaluated on the time interval $[0,\bar{t}_f]$).

(**NC**$^{\mathbf{12}}$): There exist: nonnegative numbers α_j^2, $j \in I$, numbers α_i^1, $i \in E$ such that $\sum_{i\in E} \left|\alpha_i^1\right| + \sum_{j\in I} \alpha_j^2 \neq 0$; number π; absolutely continuous function λ_1, λ_2 such that the following conditions hold:

(i) terminal conditions

$$\lambda_2(\bar{t}_f) = \phi_x^T(\bar{x}(\bar{t}_f),\bar{t}_f) + \sum_{i\in E} \alpha_i^1 \left(g_i^1\right)_x^T (\bar{x}(\bar{t}_f)) + \sum_{j\in I} \alpha_j^2 \left(g_j^2\right)_x^T (\bar{x}(\bar{t}_f))$$

(ii) adjoint equations

a.e. on $[\bar{t}_t,\bar{t}_f)$

$$\lambda_2' = -(f_2)_x^T(\bar{x},\bar{u})\lambda_2;$$

a.e. on $[0,\bar{t}_t]$

$$\lambda_1' = -(f_1)_x^T(\bar{x},\bar{u})\lambda_1$$

(iii) jump conditions

$$\begin{aligned} \pi h_x(\bar{x}(\bar{t}_t))^T + \lambda_1(\bar{t}_t) - \lambda_2(\bar{t}_t) &= 0 \\ \lambda_2(\bar{t}_t)^T f_2(\bar{x}(\bar{t}_t^+),\bar{u}(\bar{t}_t^+)) - \lambda_1(\bar{t}_t)^T f_1(\bar{x}(\bar{t}_t^-),\bar{u}(\bar{t}_t^-)) &= 0. \end{aligned}$$

from which terminal conditions for λ_1 at point $\bar{t}_t$ can be evaluated;

(iv) the weak maximum principle

a.e. on $[\bar{t}_t,\bar{t}_f)$

$$\lambda_2^T(t)\,(f_2)_u(\bar{x}(t),\bar{u}(t))u \leq \lambda_2^T(t)\,(f_2)_u(\bar{x}(t),\bar{u}(t))\bar{u}(t)$$

a.e. on $[0,\bar{t}_t)$

$$\lambda_1^T(t)\,(f_1)_u(\bar{x}(t),\bar{u}(t))u \le \lambda_1^T(t)\,(f_1)_u(\bar{x}(t),\bar{u}(t))\bar{u}(t)$$

for all $u \in \Omega$;

$$\begin{aligned}-\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f)t_f + \frac{t_f}{\bar{t}_f}\left(\int_0^{\bar{t}_t,} \lambda_1(t)f_1(\bar{x}(t),\bar{u}(t))dt + \int_{\bar{t}_t}^{t_f} \lambda_2(t)f_2(\bar{x}(t),\bar{u}(t))dt\right)\\ \le -\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{t}_f + \int_0^{\bar{t}_t} \lambda_1(t)f_1(\bar{x}(t),\bar{u}(t))dt + \int_{\bar{t}_t}^{\bar{t}_f} \lambda_2(t)f_2(\bar{x}(t),\bar{u}(t))dt\end{aligned}$$

for all $t_f \in \mathcal{T}_f$;
(v) complementarity conditions

$$\alpha_j^2 = 0, \;\text{ if }\; j \notin I_{0,\bar{u}_e}.$$

Having adjoint equations, assuming the constraint qualification **(CQ)** and taking into account *Theorem 2* one can derive necessary optimality conditions for the problem **(P)** and its case *Case 1–2*. The necessary optimality conditions (**NC**$^{\mathbf{12}}$) are stated in the form of the weak maximum principle.

The maximum condition stated for the final time variable will agree with the well–known condition for minimum time control problems if a hybrid system stays in one discrete state on the whole horizon $[0,\bar{t}_f]$.

Suppose that the discrete states switching does not occur, so for the optimal pair $(\bar{u},\bar{t}_f)$ the continuous state is decsribed by the equation $x' = f_1(x,u)$. Then, according (for example) to [5] the Hamiltonian $H(\bar{x},\bar{u},\lambda_1) = \lambda_1^T f_1(\bar{x},\bar{u})$ is a constant function of time and we can take $H(\bar{x}(\bar{t}_f),\bar{u}(\bar{t}_f),\lambda_1(\bar{t}_f))$ as this constant value. Then we have

$$\begin{aligned}-\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f)t_f + \frac{t_f}{\bar{t}_f}H(\bar{x}(\bar{t}_f),\bar{u}(\bar{t}_f),\lambda_1(\bar{t}_f))\bar{t}_f\\ \le -\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f)\bar{t}_f + H(\bar{x}(\bar{t}_f),\bar{u}(\bar{t}_f),\lambda_1(\bar{t}_f))\bar{t}_f\end{aligned}$$

from which the standard maximum condition for the minimum time control problem follows

$$\begin{aligned}\left(-\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f) + H(\bar{x}(\bar{t}_f),\bar{u}(\bar{t}_f),\lambda_1(\bar{t}_f))\right)t_f\\ \le \left(-\phi_{t_f}(\bar{x}(\bar{t}_f),\bar{t}_f) + H(\bar{x}(\bar{t}_f),\bar{u}(\bar{t}_f),\lambda_1(\bar{t}_f))\right)\bar{t}_f\end{aligned}$$

(see, for example, [5]).

Theorem 5. *Assume that the hypotheses* (**H1**), (**H2**), (**H3**), (**H4**), (**H5**), (**CQ**) *for the problem* **(P)** *are satisfied. If* $(\bar{x},\bar{u},\bar{t}_f)$ *is a solution to the problem* **(P)** *and Case 1–2 holds, then the necessary optimality conditions* (**NC**$^{\mathbf{12}}$) *are satisfied.*

Proof. As shown in the proof of Theorem 5.1 in [13] the conclusions of the proof of *Theorem 3* can be expressed by

$$\min_{d_e\in\mathcal{D}_{\bar{u}_e}}\max_{\gamma\in\mathcal{K}}\Psi(d_e,\gamma)=0$$

where

$$\mathcal{K}=\Big\{\gamma=\big(\alpha_0,\{\alpha_i^1\}_{i\in E},\{\alpha_j^2\}_{j\in I}\big)\in\mathbb{R}^{1+|E|+|I|}:\ \alpha_0\geq 0,\ \alpha_j^2\geq 0,\ j\in I,$$
$$\alpha_0+\sum_{i\in E}\big|\alpha_i^1\big|+\sum_{j\in I}\alpha_j^2=1,\ \alpha_j^2=0\ \text{ if } j\notin I_{0,\bar{u}_e}\Big\}$$

and

$$\Psi(d_e,\gamma):=\alpha_0\left\langle\nabla\bar{F}_0^e(\bar{u}),d_e\right\rangle+c\left(\sum_{i\in E}\alpha_i^1\left\langle\nabla\bar{g}_i^1(\bar{u}_e),d_e\right\rangle+\sum_{j\in I_{0,\bar{u}_e}}\alpha_j^2\left\langle\nabla\bar{g}_j^2(\bar{u}_e),d_e\right\rangle\right).$$

$\Psi(\cdot,\gamma)$ is a linear function on $\mathcal{L}_{m+1}^2[T]$ of which $\mathcal{D}_{\bar{u}_e}$ is a convex subset. $\Psi(d,\cdot)$ is a bounded linear map and $\mathcal{K}$ is a compact convex set with respect to the product topology of $\mathbb{R}^{1+|E|+|I|}$. It follows from the minimax theorem ([1]) that there exists some nonzero $\bar{\gamma}\in\mathcal{K}$ such that

$$\min_{d_e\in\mathcal{D}_{\bar{u}_e}}\max_{\gamma\in\mathcal{K}}\Psi(d_e,\gamma)=\min_{d_e\in\mathcal{D}_{\bar{u}_e}}\Psi(d_e,\bar{\gamma})=0,\tag{83}$$

with $\bar{\gamma}=(\bar{\alpha}_0,\{\bar{\alpha}_i^1\}_{i\in E},\{\bar{\alpha}_j^2\}_{j\in I})$.

Since the constraint qualification (**CQ**) holds we can show that $\bar{\alpha}_0\neq 0$.

The adjoint equations have been derived for the functional $\bar{F}_0^e(u)$, however similar analysis could be carried out for the functional

$$H(u)=\bar{F}_0^e(u_e)+c\left(\sum_{i\in E}\bar{\alpha}_i^1\bar{g}_i^1(u_e)+\sum_{j\in I}\bar{\alpha}_j^2\bar{g}_j^2(\bar{u}_e)\right),$$

and then we take $\alpha_i^1=c\bar{\alpha}_i^1$, $i\in E$, $\alpha_j^2=c\bar{\alpha}_j^2$, $j\in I$ (notice that $c>0$).

Since the obtained conclusion is for the problem **(P)** we need to apply the time transformation $[0,1]\ni\tau\to t\in[0,\bar{t}_f]:t=\bar{t}_f\tau$ to arrive at the conditions (**NC**$^{\mathbf{12}}$). □

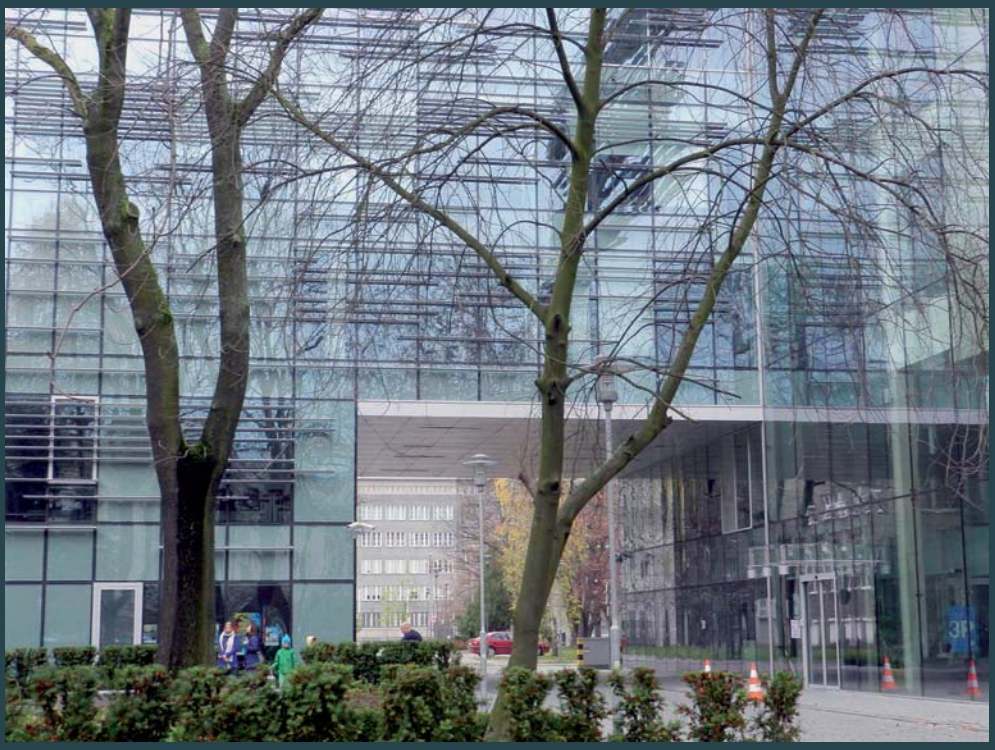
References

[1] J. Aubin and I. Ekeland. *Applied Nonlinear Analysis*. Wiley-Interscience, New York, 1984.

[2] M. S. Branicky, V. S. Borkar, and S. K. Mitter. A unified framework for hybrid control: model and optimal control theory. *IEEE Trans. on Autom. Control*, 43:31–65, 1998.

[3] A. Bryson and Y. Ho. *Applied Optimal Control*. Hemisphere, New York, 1975.

[4] L. Dieci and L. Lopez. Sliding motion in Filippov differential systems: Theoretical results and a computational approach. *SIAM J. Numer. Anal.*, 47:2023–2051, 2009.
[5] D. Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2012.
[6] J. Lygeros, K. H. Johansson, S. S. Sastry, and M. Egerstedt. On the existence of executions of hybrid automata. In *Proceed. of the 38th IEEE CDC, Phoenix, Arizona*, pages 2249–2254, 1999.
[7] A. Pakniyat and P. Caines. The hybrid minimum principle in the presence of switching costs. In *Proceed. of the 38th IEEE CDC, December 10-13, Florence, Italy*, pages 3831–3836, 2013.
[8] A. Pakniyat and P. Caines. Time optimal hybrid minimum principle and the gearchanging problem fo relectric vehicles. In *Proceed. of the 5th IFAC Conference on Analysis and Design of Hybrid Systems October 14-16, 2015. Georgia Tech, Atlanta, USA*, pages 187–192, 2015.
[9] R. Pytlak. *Numerical Methods for Optimal Control Problems with State Constraints. Lecture Notes in Mathematics 1707*. Springer–Verlag, Berlin, Heidelberg, 1999.
[10] R. Pytlak and D. Suski. On solving hybrid optimal control problems with higher index DAEs. *Optimization Methods & Software*, 32:940–962, 2017.
[11] R. Pytlak and D. Suski. Optimal control of hybrid systems with sliding modes. In *Springer Proceedings in Mathematics and Statistics, Vol. 248*, pages 283–293, 2017.
[12] R. Pytlak and D. Suski. Trajectory sensitivity analysis of hybrid systems with sliding motion. *submitted for publication*, 2020.
[13] R. Pytlak and R. Vinter. A feasible directions algorithm for optimal control problems with state and control constraints: convergence analysis. *SIAm J. on Control and Optimization*, 36:1999–2019, 1998.
[14] C. Seatzu, D. Corona, A. Giua, and A. Bemporad. Optimal control of continuous time switched affine systems. *IEEE Transactions on AC*, 51:726–741, 2006.
[15] M. S. Shaikh. Optimal control of hybrid systems: theory and algorithms, Ph.D. dissertation, McGill University, Montreal, 2004.
[16] M. S. Shaikh and P. Caines. On the hybrid optimal control problem: Theory and algorithms. *IEEE Transactions on AC*, 52:1587–1603, 2007.
[17] M. S. Shaikh and P. Caines. Correction to: On the hybrid optimal control problem: Theory and algorithms. *IEEE Transactions on AC*, 54:1, 2009.
[18] H. Sussmann. A maximum principle for hybrid optimal control problems. In *Proceed. of the 38th IEEE CDC, Phoenix, Arizona*, pages 425–430, 1999.
[19] F. Taringoo and P. Caines. Gradient-geodesic HMP algorithms for the optimization of hybrid systems based on the geometry of switching manifolds. In *Proceed. of the 38th IEEE CDC, December 15-17, Atlanta, GA*, pages 1534–1539, 2010.
[20] F. Taringoo and P. Caines. On the extension of the hybrid minimum principle to Riemannian manifolds. In *Proceed. of the 38th IEEE CDC, December 12-15, Orlando, Fl*, pages 3301–3306, 2011.
[21] F. Taringoo and P. Caines. On the optimal control of hybrid systems on Lie groups and the exponential gradient HMP algorithm. In *Proceed. of the 38th IEEE CDC, December 10-13, Florence, Italy*, pages 2653–2658, 2013.
[22] A. van der Schaft and H. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Springer–Verlag, London, 2000.
[23] H. Witsenhausen. A class of hybrid-state continuous-time dynamic systems. *IEEE Transactions on AC*, 11:161–167, 1966.



ISBN 978-83-8156-156-3