ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ
„СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА
И ИНФОРМАТИКА

КНИГА 1 – МАТЕМАТИКА И МЕХАНИКА
КНИГА 2 – ПРИЛОЖНА МАТЕМАТИКА И
ИНФОРМАТИКА

Том 91

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA
„ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES
ET INFORMATIQUE

LIVRE 1 – MATHEMATIQUES ET MECANIQUE
LIVRE 2 – MATHEMATIQUES APPLIQUEE ET
INFORMATIQUE

Tome 91

СОФИЯ • 1999

УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“

SOFIA • 1999

ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Книга 1 — МАТЕМАТИКА И МЕХАНИКА
Книга 2 — ПРИЛОЖНА МАТЕМАТИКА
И ИНФОРМАТИКА

Том 91
1997

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA "ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE

Livre 1 — MATHEMATIQUES ET MECANIQUE
Livre 2 — MATHEMATIQUES APPLIQUEE ET INFORMATIQUE

Tome 91
1997

СОФИЯ • 1999 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
PRESSES UNIVERSITAIRES "ST. KLIMENT OHRIDSKI"

Annuaire de l' Université de Sofia "St. Kliment Ohridski"
Faculté de Mathématiques et Informatique

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

Address for correspondence:

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 Blvd J. Bourchier, P.O. Box 48
BG-1164 Sofia, Bulgaria

Fax xx(359 2) 687 180
electronic mail:
`annuaire@fmi.uni-sofia.bg`

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

ISSN 0205–0808

Professor Yaroslav Tagamlitzki

On December 20, 1997 the Faculty of Mathematics and Informatics organized a special Scientific Session devoted to the 80th birthday of Prof. Yaroslav Tagamlitzki (1917–1983). In this annual the scientific program of the Session is included, together with the full texts of some of the lectures presented there.

Professor Tagamlitzki was a brilliant and creative mathematician, well-known in the mathematical community, the teacher of several generations of Bulgarian mathematicians.

With this issue the Editorial Board pays a small tribute to the memory of Professor Tagamlitzki.

The Editorial Board

CONTENTS

Book 1

MATHEMATICS AND MECHANICS

Book 2

APPLIED MATHEMATICS AND INFORMATICS

ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Книга 1 — МАТЕМАТИКА И МЕХАНИКА

Том 91
1997

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA "ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE

Livre 1 — MATHEMATIQUES ET MECANIQUE

Tome 91
1997

СОФИЯ • 1999 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
PRESSES UNIVERSITAIRES "ST. KLIMENT OHRIDSKI"

НАУЧНА СЕСИЯ
В ПАМЕТ НА ПРОФ. ЯРОСЛАВ ТАГАМЛИЦКИ

София, 20 декември 1997 г.

НАУЧНА ПРОГРАМА

Предиобедно пленарно заседание

ФМИ, аудитория 229

9.15–10.00 ч. В. Чакалов, Д. Скордев — Животът и делото на Ярослав Тагамлицки

Секционни заседания

Секция А

ФМИ, аудитория 326

председател: В. Чакалов

10.15–10.35 ч. Т. Тонков — Върху една процедура с приложения в анализа и теорията на числата

10.35–10.55 ч. Н. Зяпков, И. Михайлов — Съпътстващи задачи в теорията за вложимост на полета

10.55–11.15 ч. С. Марков — Алгебрични свойства на изпъкнали тела

председател: Р. Леви

11.30–11.50 ч. Ж. Желязков, Ц. Игнатов, А. Маринов — Върху една задача от теория на риска

11.50–12.10 ч. В. Чакалов — Интерполация и апроксимация на аналитични функции

12.10–12.30 ч. Н. Буюклиев — К-теория на C^*-алгебрата на операторите на Винер-Хопф за полиедрален конус в R^n

Секция Б

ФМИ, аудитория 229

председател: Д. Скордев

10.15–10.35 ч. Н. Хаджииванов — Усилване на една теорема на Ердьош-Клейтмън

10.35–10.55 ч. Е. Недялков, Н. Ненов — Всеки 11-върхов граф без 4-клики има свободно от 3-клики 2-разлагане на върховете

10.55–11.15 ч. С. Михов — Директно построяване на ацикличен автомат по даден списък

председател: Н. Хаджииванов

11.30–11.50 ч. И. Сосков — Равномерни операции

11.50–12.10 ч. Д. Скордев — Математико-логически подход към някои въпроси, свързани с теоремата на Я. Тагамлицки за отделимост

12.10–12.30 ч. Д. Вакарелов — Едно приложение в логиката на теоремата на Я. Тагамлицки за отделимост

Секция В
ФМИ, аудитория 229

председател: Т. Генчев

14.30–14.50 ч. П. Попиванов — Локална неразрешимост на линейни и нелинейни частни диференциални уравнения

14.50–15.10 ч. Е. Хорозов — Нови класове от ортогонални полиноми, възникващи като собствени функции на линейни диференциални оператори

15.10–15.30 ч. Н. Попиванов — Нелокална регуляризация за някои некоректни задачи за частни диференциални уравнения

председател: Е. Хорозов

15.45–16.05 ч. Ц. Дончев, Й. Славов — Функционално-диференциални включвания с обобщени решения и сингулярен параметър

16.05–16.25 ч. Л. Каранджулов — Асимптотично решение на определен клас сингулярно смутени линейни гранични задачи от обикновени диференциални уравнения

16.25–16.45 ч. А. Живков — Пумпал на Лагранж

16.45–17.05 ч. Ж. Атанасов — Негладки функционали на Ляпунов и инвариантност за диференциални уравнения със закъснение

Секция Г
ФМИ, аудитория 326

председател: Е. Христов

15.45–16.05 ч. И. Димовски — Циклични елементи на оператори за интегриране

16.05–16.25 ч. Х.-П. Блатт, В. Андриевский, Р. Ковачева — Апроксимация върху неограничени интервали

16.25–16.45 ч. А. Томова — Приложение на теорията на множествата на Жюлия към модифицираните методи на Нютон-Рафсон и Чебишев

Следобедно пленарно заседание
ФМИ, аудитория 229

17.15 ч. — ... Споделяне на спомени за проф. Ярослав Тагамлицки.

Закриване на научната сесия.

Слово, произнесено от проф. Т. Генчев при откриването на юбилейната научна конференция по случай осемдесетата годишнина от рождението на проф. Тагамлицки

Уважаеми колеги, скъпи гости,

Организационният комитет ме натовари с приятната и почетна задача да открия научната сесия в памет на бележития български математик, незабравимия учител на всички нас, професор Ярослав Тагамлицки.

Русин по националност, прекарал целия си съзнателен живот в България, Тагамлицки израства и се развива като български математик и български патриот. Той се гордееше, че е професор именно в Софийския университет и твърдо вярваше, че нашият факултет е научен център, в който са правени и се правят изследвания от европейски мащаб. Оставяйки ни непреходни резултати в класическия и функционалния анализ, Тагамлицки подкрепи своето верую с едно забележително творческо дело. Той смяташе, че българската математика е достатъчно зряла, за да може сама да преценява своите и чуждите постижения, и видимо се дразнеше, когато при обсъждането на научни въпроси вместо доводи по същество се привеждаха цитати от чужди авторитети. Неговото изострено чувство за ценностите в науката уравновесяваше вродената му доброжелателност и го предпазваше от прибързани похвали. Тагамлицки беше взискателен, но неговата взискателност не беше едностранчива; той беше строг съдник не само на чуждото, но и на своето творчество. В страна с неукрепнали научни традиции, при това в процес на екстензивно развитие като следвоенна България, неговата взискателност, която не беше нищо друго освен уважение към науката, често биваше неправилно разбирана и му носеше немалко огорчения. Въпреки това той изпълняваше достойно мисията на културтрегер, с която се беше нагърбил.

За да разберем истинското място на Тагамлицки в историята на нашия факултет, трябва да се върнем назад в началото на петдесетте години. По онова време, от второто поколение университетски професори, само Обрешков беше още в творческа възраст, но не ръководеше научен семинар и поради особеностите на своя характер не правеше усилия да създаде научна школа. От друга страна, като следствие от войната старите научни контакти със Запада (имам предвид Германия и Франция) бяха прекъснати, а новите, преди всичко с МГУ, още не бяха създадени. Именно тогава, рано достигнал творческа зрялост, Тагамлицки създаде семинар, който ръководеше с възрожденска всеотдайност, въпреки огромната задължителна преподавателска работа, която лежеше на неговите плещи. Този семинар, заедно със спецкурсовете, които ежегодно четеше, оформи център, около който се групираха студентите с афинитет към науката. По този начин, с талант и себераздаване, Тагамлицки хвърли мост между миналото и бъдещето и създаде новата българска школа. Ако сега, пред прага на новото столетие, се запитаме кой е математикът, допринесъл най-много за развитието на факултета през последните петдесетина години, отговорът може да бъде само един: Ярослав Тагамлицки. И слава Богу, до този отговор се достига и без преброяване на цитати.

Въпреки неговите безспорни заслуги и към българската наука и към българското образование, Тагамлицки не беше удостоен с членство в БАН. Академичната номенклатура му отказа онова, което беше заслужил многократно. Но за нас, неговите ученици, които имахме привилегията да го познаваме и обичаме, той беше и е нещо повече от академик. За нас той е фигура с други измерения, от друг мащаб, учен и просветител в исконния смисъл на думите, борец за по-широки духовни хоризонти и за по-висока математическа култура. Мисля, че няма да сбъркам, ако кажа, че това мнение се споделяше и от най-широките кръгове на нашата математическа общественост.

Вече 14 години Тагамлицки не е между нас. Освен своето творчество той ни остави и своя пример. Днес, в трудните преходни години, този пример придобива особена стойност. Наистина какво бихме могли да противопоставим на моралната ерозия, този неизбежен спътник на недоимъка, освен примера на най-достойните? Достойно място между тях заема и нашият незабравим учител професор Ярослав Тагамлицки.

От името на организационния комитет обявявам днешната научна сесия за открита и давам думата на проф. Скордев да изнесе доклад за живота и научното дело на проф. Тагамлицки.

София,
20 декември 1997

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

ЖИВОТЪТ И ДЕЛОТО НА ЯРОСЛАВ ТАГАМЛИЦКИ

ВЛАДИМИР ЧАКАЛОВ, ДИМИТЪР СКОРДЕВ

Човекът, комуто е посветена настоящата научна сесия, се е родил на 11.09.1917 г. в руския град Армавир (градът е разположен там, където р. Кубан напуска предпланините на Кавказ и навлиза в равнината, разположена на север от тях). Кръщелното име на този човек е Ярослав-Роман Александрович Тагамлицкий. Всъщност, както той е споделял в разговор, малкото му име, избрано от родителите, е било само Ярослав (това, с което обикновено е назоваван по-нататък), но свещеникът, който е трябвало да извърши кръщението, се е възпротивил срещу предложеното име, според него езическо, и се е наложило да се направи компромис, като се добави и името Роман. Фамилното име пък произлизало от наименованието Тагамлик, носено от някакво неголямо населено място в тази част на страната. Както знаем, много скоро след споменатата рождена дата приближаващата своя край Първа световна война преминава в нови, още по-жестоки, многобройни и продължителни изпитания и сътресения за Русия. Те стават причина през 1921 г. цялото семейство — бащата Александър Михайлович (инженер по професия), майката Вера Леонидовна, малкият Ярослав и сестра му Галина (с една година по-голяма от него), да се пресели в България. Фактически България става родина на Ярослав Тагамлицки и през целия си съзнателен живот, до своя последен ден — 28.11.1983 г., той работи за нейното развитие и издигане.

След преселването си в България семейство Тагамлицки се установява в София, където систематично понася големи материални несгоди, допълнително утежнени от все по-влошаващо се здравословно състояние на бащата (по-подробни сведения за този период могат да се намерят в спомените на проф. Галина Тагамлицка „Моят брат Ярослав Тагамлицки“, включени в сборника „Ярослав Тагамлицки — учен и учител“, издание на „Наука и изкуство“ от 1986 г.). Основното образование на младия Тагамлицки минава без нещо да подсказва за големите заложби, скрити в него, но при постъпването му в известната Втора мъжка софийска гимназия нещата коренно се променят. По всичко личи, че се е получило едно изключително благоприятно съчетание, от една страна, на голямата природна надареност на съзрелия вече ученик и неудържимия му стремеж към науката и, от друга страна, на високото професионално ниво на учителите и тяхната възрожденска обич към професията и грижата им за обучаваните. Математическите интереси на Тагамлицки през този период вече далеч надхвърлят изучаваната гимназиална материя и той има и сериозни изяви на самостоятелно научно творчество, макар за негово разочарование получените резултати да се оказват известни от по-рано. Пак през този период Тагамлицки е редовен слушател на университетските лекции на гостувация през 1935 г. в София виден немски математик Ото Блументал. Впрочем не само в математиката и не само в науката е проявил своите заложби даровитият и ученолюбив младеж. От това време датира например и неговото голямо влечение към музиката, интересът към която и от естетическа, и от научна гледна точка не го напуска до края на живота му.

Гимназиалното си образование Ярослав Тагамлицки завършва през 1936 г. Същата година постъпва в специалността математика на тогавашния Физико-математически факултет на Софийския университет и бързо привлича вниманието на своите професори както с дълбочината, така и с обхвата на своите познания, а особено със своите забележителни творчески възможности. Още по време на следването си Тагамлицки написва три научни статии, публикувани през 1938 и 1939 г. (две от тях са отпечатани във Физико-математическото списание, а третата — в Юбилеен сборник на Физико-математическото дружество). И трите статии свидетелстват за зрялост, несвойствена даже за такъв добър студент, а третата показва и детайлно познаване на теорията на Лебеговия интеграл, която по това време изобщо не се преподава в Софийския университет.

През 1940 г. Ярослав Тагамлицки завършва висшето си образование и е командирован от Министерството на народната просвета за научна работа в Софийския университет. През 1942 и 1943 г. е на специализация в Лайпцигския университет при известните математици Кьобе и Ван дер Варден. Специализацията му завършва със защита на докторска работа, в която се обобщава една известна теорема на Кьобе от теорията на аналитичните функции. Тук проличава способността на Тагамлицки бързо да навлезе и да се задълбочи в област, която дотогава е била извън интересите му, и то до степен да придвижи напред изследванията на нейния създател.

След като се връща в България, Тагамлицки е призован да отбие военната си служба. Един драматичен и много опасен момент от това време е избавянето му (заедно с още няколко други войници) от немски плен на територията на Югославия през есента на 1944 г.

През 1945 г. Ярослав Тагамлицки е назначен за асистент към катедрата по диференциално и интегрално смятане, чийто ръководител е бъдещият академик проф. Кирил Попов. Това е начало на изключително интензивна научна и преподавателска работа, изпълнила живота на Тагамлицки до последния му ден. Много скоро в изследванията на младия асистент започва все по-осезателно да звучи един лайтмотив, който след време довежда до разработването на нова за България, твърде интересна и обещаваща научна област. Става дума за понятието неразложимост, за неговата роля в анализа и за възможностите за приложения на свойствата на неразложимите елементи. Така например в статията от 1946 г. „Функции, които удовлетворяват известни неравенства върху реалната ос“ по същество се доказва неразложимостта на показателната функция в един естествено възникващ конус от функции. В някои следващи работи се показват аналогични свойства и на безкрайните геометрични прогресии. В поредица от привидно разнородни резултати Тагамлицки със забележително прозрение вижда дълбоката същност, която ги обединява, и настоятелно се стреми към пълното ѝ разкриване и към намиране на други нейни прояви. Тези търсения дават своите резултати. През 1949 г. се появява статията му „Върху някои приложения на общата теория на линейните пространства с частично нареждане“. В нея той формулира първите си общи теореми, отнасящи се за линейни пространства, и в тези теореми се използват свойствата на неразложимите елементи. От споменатите теореми вече прозира идеята на общия метод, върху чието създаване работи Тагамлицки. Без все още да е формулирал този метод в явен вид, Ярослав Тагамлицки доказва по сходен начин поредица от интересни и нетривиални резултати за представяне на функции чрез безкрайни редове или чрез интеграли. Безспорно най-интересната му работа в тази област е изследването върху интерполационния ред на Абел, публикувано през 1950 г. в Годишника на Софийския университет и през 1951 г. в Докладите на Академията на науките на СССР. В това изследване се получава един дълбок резултат за представяне чрез сбор на сума на безкраен ред и на интеграл, като този резултат разкрива причините за многобройните случаи, в които интерполационен ред на Абел не представя функцията, на която съответства. За споменатото изследване Тагамлицки получава Димитровска награда през 1952 г. (преди това през 1947 г. за други негови изследвания от същата поредица му е дадена наградата за наука на Комитета за наука, изкуство и култура). През 1953 г. за научната и преподавателската си дейност е награден с орден „Кирил и Методий“, I степен.

Междувременно Ярослав Тагамлицки е избран за частен доцент (през 1947 г.) и за редовен доцент (през 1949 г.). От 1954 г. той е професор, завеждащ катедрата по диференциално и интегрално смятане във Физико-математическия факултет на Софийския университет. Тъй като по-нататъшните изрази на научно признание, които Тагамлицки получава от ръководните инстанции в българската наука, не са особено многобройни, ще ги споменем тук, за да не разкъсваме изложението и за да се съсредоточим върху това, което той самият е считал за най-важно. През 1958 г. му е присъдена втора докторска степен — съгласно новите тогава правила за научните степени в страната. През 1961 г. е избран за член-кореспондент на БАН. Едновременно с катедрата по диференциално и интегрално смятане ръководи и секцията по функционален анализ в Математическия институт на БАН. След обединяването на двете научни звена в края на 1970 г. е ръководител на възникналия в резултат на това обединение сектор по реален и функционален анализ в създадения тогава Единен център по математика и механика. През 1967 г. повторно е награден с орден „Кирил и Методий“, I степен, а през 1969 г. — с юбилеен медал „25 години народна власт“. През 1982 г., когато навършва 65 години, му е присъдено званието „Заслужил деятел на науката“.

През 1952 г. се появява статията на Тагамлицки „Върху геометрията на конусите в Хилбертовите пространства“, а през 1954 г. — обобщаващата нейния основен резултат статия „Върху едно обобщение на понятието неразложимост“. Същността на този резултат и в двете статии е следната: при определени условия, за да се съдържа един конус в друг, достатъчно е неразложимите елементи на първия конус да принадлежат на втория (разбира се, става дума за конуси в линейни пространства). Това може да бъде оприличено на твърдението на принципа на математическата индукция, което казва, че при определени условия, за да се съдържа множеството на всички естествени числа в дадено множество, достатъчно е най-малкото от естествените числа (0 или 1 в зависимост от терминологията) да принадлежи на въпросното множество. Резултатът на Тагамлицки, за който става дума, получава наименованието „теорема за конусите“ и с негова помощ самият Тагамлицки, а по-нататък и негови ученици доказват редица известни теореми от анализа като например теоремата на Хаусдорф за моментите, теоремата на Уидер за представяне на функции с Лапласов интеграл, теоремата на Бернщейн за интегрално представяне на регулярно монотонните функции и ред други интересни и важни теореми. Някои нови резултати също биват открити и доказани най-напред с помощта на теоремата за конусите, а след това и по по-пряк начин. През 1953–1954 г. Тагамлицки дава и един начин за изграждане на теорията на обобщените функции с помощта на теоремата за конусите и на понятието неразложимост.

Основно средище за работа на Тагамлицки с млади обещаващи математици по това време става създаденият към катедрата по диференциално и интегрално смятане студентски кръжок (дошлата от руския език дума „кръжок“ по това време се употребява в смисъл на днешното „семинар“ и в разни случаи означава различни по своето ниво неща). В годините на извънредно силна научна изолация на България след Втората световна война създаденият от Тагамлицки кръжок става всъщност първият и в немалък срок единствен научен семинар от високо ниво във факултета. Това създава една неповторима и необичайна за времето си атмосфера, при която на младите математици още от студентската скамейка се дава възможност да станат пълноценни и равноправни сътрудници в научната дейност на своя изпълнен с ентусиазъм ръководител. Много от активно работещите през следващите десетилетия математици се изграждат като творци в науката именно преди всичко благодарение на работата си в кръжока на Тагамлицки.

Разбира се, кръжокът, за който става дума, до известна степен смекчава някои последици от научната изолация, за която споменахме, но за съжаление той не успява напълно да ги компенсира. Оказва се, че информацията за някои резултати, известни по света, по онова време идва в България с голямо закъснение. Един такъв резултат е теоремата на Крейн и Милман, публикувана през 1940 г. в списанието „Studia Mathematica“. Днес не би представлявало проблем веднага да се забележи, че теоремата за конусите може да се получи като следствие от теоремата на Крейн и Милман, стига да не обръщаме специално внимание на въпроса за използването и неизползването на аксиомата за избора. На времето обаче за това се оказват нужни няколко години и въпросното обстоятелство се изяснява едва през 1957 г. Все пак на Тагамлицки и на неговите ученици от онова време остава утешението, че са направили редица нови и нетривиални приложения на един общ метод, независимо дали методът ще се основава на теоремата за конусите или направо на теоремата на Крейн и Милман. Не е за подценяване и това, че най-добре се разбират, овладяват и прилагат онези постижения на математиката, до които човек е успял да достигне или да се доближи самостоятелно.

Един следващ кръг от изследвания на Тагамлицки е насочен към обобщение на понятието изпъкналост. През 1963 г. се появява статията му „Върху принципа за отделимост в абелевите асоциативни пространства“. В нея се дава аксиоматизация на понятието отсечка, при която аксиоматизация става достъпна задачата да се обобщи обичайният принцип за отделяне на непресичащи се изпъкнали множества с помощта на полупространства. При това, за разлика от някои по-ранни и неизвестни по онова време в България изследвания на други автори, не се забранява отсечка със съвпадащи краища да съдържа точки, различни от тях. Това дава възможност за разглеждане на по-широк кръг от модели и за повече приложения на доказаната от Тагамлицки теорема за отделимост. Например, както той посочва, можем, когато е дадена една комутативна полугрупа, под отсечка, определена от два елемента на полугрупата, да разбираме едноточковото множество, чийто елемент е произведението на дадените два елемента, и това веднага дава една теорема за отделимост на кои да е две непресичащи се подполугрупи. Изследванията на Тагамлицки върху принципа за отделимост са продължени от неговия блестящ ученик Иван Проданов, за съжаление твърде рано покосен от смъртта.

Продължавайки изследванията си, Тагамлицки успява да формулира едно далеч отиващо обобщение на теоремата на Крейн и Милман, което за разлика от нея се отнася не за локално изпъкнали линейни пространства, а за общи топологични пространства. Това обобщение получава името „топологична индукция“ (кратка публикация на Тагамлицки върху него се съдържа в трудовете на международен топологически симпозиум, проведен през 1968 г., а малко по-подробно изложение на резултата, подготвено на основата на материали на Тагамлицки — в трудовете на семинара на Г. Шоке в Париж за 1970/1971 г.). В своя статия от 1975 г. Тагамлицки използва топологичната индукция, за да докаже твърдение, което е в духа на един принцип на Брауер за максимум, но е приложимо и в някои случаи, когато този принцип не е приложим.

Изследванията на Тагамлицки върху понятието неразложимост, неговото обобщение и приложенията му събуждат значителен интерес в чужбина, особено след проведената през 1956 г. в София научна сесия с международно участие. Покани да публикува систематично изложение на тези резултати той получава от Франция, Германия и САЩ, но за съжаление не се стига до реализирането им. Не по-малко може да се съжалява, че проф. Тагамлицки не публикува и редица други свои резултати, като например доказания от него диагонален принцип за обобщени редици, съдържащ като частен случай теоремата на Тихонов за компактност, резултатите в теорията на многообразията и др. (след смъртта на проф. Тагамлицки бе оформена публикация върху диагоналния принцип, като за основа послужиха някои негови ръкописи, приложени към годишни научни отчети на сектора по реален и функционален анализ). Не можем със сигурност да кажем коя е причината за това, че през последните двадесет години от живота си проф. Тагамлицки публикува твърде рядко, в несъответствие с продължаващата негова висока научна активност. Допускаме, че тя е във високия критерий за оценяване на научните работи, който той прилага особено безкомпромисно към себе си. Имаме някои основания да предположим, че Тагамлицки е смятал редица свои изследвания за незавършени, защото се е стремял и се е надявал да намери много по-убедителни техни приложения за решаване на съществени задачи, поради което е отлагал публикуването им.

Не можем да не отбележим, че научните интереси на Тагамлицки съвсем не се ограничават в рамките на математиката. В неговата научна дейност се наблюдават черти на някогашните учени-енциклопедисти — преди всичко жив и активен интерес към всеки важен научен проблем, независимо от научната област, към която спада. Сред областите, на които той е отделил немалко време, са теоретичната физика, археологията, науката за древните езици, медицината. Споменахме по-рано и за научните му интереси в областта на музиката.

В живота на проф. Ярослав Тагамлицки не по-малко, а може би и още по-голямо място от научните изследвания заема преподаването на науката. Той винаги е имал много голяма лекционна заетост (от порядъка на 10 и повече часа седмично) и е полагал огромни грижи, за да направи своята преподавателска работа по-резултатна (включително и при студентите с по-ограничени възможности). По своя инициатива той се натоварва например с даване и преглеждане на многобройни домашни работи, които му помагат да следи и насочва развитието на всеки от обучаваните студенти (даваните задачи често биват два вида — едните са предназначени за всички студенти, а другите, значително по-трудни, са само за желаещите и успешното им решаване се награждава с похвала на следващата лекция). Учебникът на проф. Тагамлицки по диференциално и интегрално смятане също е проникнат от такава грижа за обучаваните, като същевременно е първият издържан в научно отношение учебник по тази дисциплина у нас. За научното израстване на по-изявените студенти изключително голяма роля изиграват ежегодно четените от проф. Тагамлицки специални курсове, особено тези по функционален анализ, отразяващи най-активните в момента собствени изследвания на лектора.

Съществено място в дейността на Тагамлицки заемат и въпросите на преподаването в средното училище. Неведнъж той е изнасял посрещани с голям интерес лекции пред ученици. Много грижи е полагал за методическото разработване на определени по-трудни въпроси, какъвто е например въпросът за преподаването на основите на математическия анализ в средното училище. През учебната 1973–1974 г. той даже сам води занятия в една столична гимназия в съответствие с методиката, която е предложил. За сериозното му и задълбочено отношение към тези въпроси свидетелстват няколко негови публикации с методически характер (след смъртта на проф. Тагамлицки бяха подготвени още някои публикации, отразяващи част от методическото му наследство в областта на средното образование).

Времето, с което разполагаме, и нашите скромни възможности не позволяват да дадем достатъчно пълна картина на това, което беше за науката, за университета и за нас незабравимият наш учител проф. Ярослав Тагамлицки. Настоящият доклад е само един малък опит да изразим своето дълбоко преклонение пред забележителното му дело и пред светлата му памет. Ние вярваме, че ярката следа, която Тагамлицки остави след себе си, ще пребъде и направеното от него ще се ползва и от идващите поколения, лишени от щастието да имат непосредствен досег с този забележителен човек.

София,
20 декември 1997 г.

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

A POLYNOMIAL PROBLEM

PAVEL G. TODOROV

We show that the roots of the equation (5) with respect to z are among the roots of the equation (6). Therefore the roots of the given equation (5) are determined by means of a check of the roots of the resolvent equation (6). Some examples and applications are given.

Keywords: two polynomial equations in two variables, common roots, Sylvester method of elimination, determinants, a check of the roots of the resolvent equation in the given equation

1991 Math. Subject Classification: primary 11C20, 12D10, 30C15

EXPOSITION OF THE PROBLEM

First we shall prove the following

Theorem 1. *Let*

$$p \equiv P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0, \quad a_n \neq 0, \ n \geq 1, \tag{1}$$

and

$$Q(z) = b_m z^m + b_{m-1} z^{m-1} + \cdots + b_1 z + b_0, \quad b_m \neq 0, \ m \geq 1, \tag{2}$$

and let

$$q \equiv Q(\bar{z}) = b_m \bar{z}^m + b_{m-1} \bar{z}^{m-1} + \cdots + b_1 \bar{z} + b_0, \tag{3}$$

and

$$\bar{q} \equiv \overline{Q(\bar{z})} = \bar{b}_m z^m + \bar{b}_{m-1} z^{m-1} + \cdots + \bar{b}_1 z + \bar{b}_0. \tag{4}$$

Then all roots of the equation

$$Q(\bar{z}) = P(z) \tag{5}$$

with respect to z are roots of the determinant (resolvent) equation

$$\begin{matrix} n\ \text{rows}\left\{\begin{matrix} \\ \\ \\ \\ \end{matrix}\right. \\ m\ \text{rows}\left\{\begin{matrix} \\ \\ \\ \\ \end{matrix}\right. \end{matrix}\begin{vmatrix} b_m & b_{m-1} & \dots & b_1 & b_0-p & 0 & 0 & \dots & & 0 \\ 0 & b_m & \dots & b_2 & b_1 & b_0-p & 0 & \dots & & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & & \\ 0 & 0 & \dots & b_m & b_{m-1} & \dots & & \dots & b_1 & b_0-p \\ \bar{a}_n & \bar{a}_{n-1} & \dots & \bar{a}_1 & \bar{a}_0-\bar{q} & 0 & 0 & \dots & & 0 \\ 0 & \bar{a}_n & \dots & \bar{a}_2 & \bar{a}_1 & \bar{a}_0-\bar{q} & 0 & \dots & & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & & & \\ 0 & 0 & \dots & \bar{a}_n & \bar{a}_{n-1} & \dots & & \dots & \bar{a}_1 & \bar{a}_0-\bar{q} \end{vmatrix} \tag{6}$$

as well, but, conversely, not always all roots of the equation (6) *are roots of the equation* (5) *as well, where the determinant is of order $n+m$.*

The determinant equation (6) *has*:

(i) *exactly n^2 roots if $n > m$,*

(ii) *exactly $n^2 = m^2$ roots if $n = m$ and $|a_m| \neq |b_m|$, and less than $n^2 = m^2$ roots if $n = m$ and $|a_m| = |b_m|$, both under the condition that all the equations*

$$\bar{a}_s = b_s e^{-i\varphi},\ 1 \leq s \leq m, \qquad a_0 = b_0 \pm i r_0 e^{i\frac{\varphi}{2}}, \tag{7}$$

where $\varphi \equiv \operatorname{Arg} a_m + \operatorname{Arg} b_m \pmod{2\pi}$, $r_0 \geq 0$ and the signs ± are taken singly, cannot exist simultaneously, and

(iii) *exactly m^2 roots if $n < m$.*

Proof. Let us examine the equations

$$b_m\zeta^m + b_{m-1}\zeta^{m-1} + \cdots + b_1\zeta + b_0 - p = 0 \tag{8}$$

and

$$\bar{a}_n\zeta^n + \bar{a}_{n-1}\zeta^{n-1} + \cdots + \bar{a}_1\zeta + \bar{a}_0 - \bar{q} = 0. \tag{9}$$

According to the classical Sylvester method of elimination, the two equations (8) and (9) have a common root ζ only if z is a root of the eliminating equation (6), and conversely (see the Sylvester method, for example, in Dickson's book [1, p. 164]). Hence, if a common root ζ of the two equations (8) and (9) is equal to $\bar{z}$, where z is a root of the resolvent (determinant) equation (6), then z is a root of the given equation (5) as well, taking into account the same multiplicity of z as a root of the determinant (resolvent) equation (6). If a common root ζ of the two equations (8) and (9) is not equal to $\bar{z}$, where z is a root of the determinant equation (6), then z is not a root of the given equation (5) as well.

If $n = m$, the condition in (ii) (see (7)) ensures that the equation (6) is not an identity with respect to z. Indeed, for $n = m$, the determinant in (6) is identically equal to zero with respect to z only if the two equations (8) and (9) are reduced to one equation, i.e. keeping in mind (1)–(4), if we have the identity

$$\bar{p} - \bar{q} \equiv \lambda(q - p) \quad (\zeta = \bar{z}) \tag{10}$$

for some complex (or real) number $\lambda \neq 0$ which does not depend on z and ζ. Thus from (10) we obtain the equations

$$\bar{a}_s = \lambda b_s, \quad 1 \leq s \leq m, \tag{11}$$

and the identity

$$\bar{a}_0 - \bar{q} \equiv \lambda(b_0 - p). \tag{12}$$

Further, from (12) it follows that

$$\bar{b}_s = \lambda a_s, \quad 1 \leq s \leq m, \tag{13}$$

and

$$\bar{b}_0 - \bar{a}_0 = \lambda(a_0 - b_0). \tag{14}$$

Now, from (11) and (13) for $s = m$, we obtain $|a_m| = |b_m|$ and hence $|\lambda| = 1$, i.e.

$$\lambda = e^{-i\varphi}, \tag{15}$$

where $\varphi \equiv \operatorname{Arg} a_m + \operatorname{Arg} b_m \pmod{2\pi}$. Therefore from (11) and (15) we get the first equations in (7). Finally, if we set $a_0 - b_0 = r_0 e^{i\alpha}$, $r_0 \geq 0$, α being real (α is arbitrary if $r_0 = 0$), from (14) and (15) we find $2\alpha = \pi + \varphi + 2k\pi$, $k = 0, \pm 1, \pm 2, \ldots$, if $r_0 > 0$, i.e. we obtain the second equations in (7).

Now we shall determine the degree of the resolvent equation (6) with respect to z. For the solution of this problem we shall use the fact that each summand in (6) consists of a product of elements of different columns and rows.

(i) Let $n > m$. Let k be a non-negative integer such that $0 \leq k \leq m$. If we take k times the binomial $\bar{a}_0 - \bar{q}$ and $n - k$ times the binomial $b_0 - p$, then we obtain the expression $(b_0 - p)^{n-k}(\bar{a}_0 - \bar{q})^k$ in which the highest degree of z is $n(n-k) + mk$. But $n(n-k) + mk \leq n^2$ for the considered n, m and k with equality sign only for $k = 0$. Thus we proved that the determinant development of (6) includes only one summand of the form $(-1)^{nm}\bar{a}_n^m(b_0 - p)^n$ (the sign is $(-1)^{nm}$ since the number of the inversions of the permutation of the columns in order $m+1$, $m+2$, $\ldots$, $m+n$, 1, 2, $\ldots$, m is nm), i.e. the resolvent equation (6) is exactly of degree n^2 with respect to z.

(ii) Let $n = m$ and the equations (7) not exist simultaneously. Then if we take k times $(0 \leq k \leq m)$ the binomial $\bar{a}_0 - \bar{q}$ and $m - k$ times the binomial $b_0 - p$, we obtain the expression $(b_0 - p)^{m-k}(\bar{a}_0 - \bar{q})^k$ in which the highest degree of z is $m(m-k) + mk = m^2$. Hence the determinant development of (6) contains the sum

$$\sum_{k=0}^{m} \sum_{k} (-1)^{\nu_{mk}} b_m^k (\bar{a}_0 - \bar{q})^k \bar{a}_m^{m-k} (b_0 - p)^{m-k}, \tag{16}$$

where the number of the summands in the inner sum is equal to the number of the combinations of m elements of the class k, and ν_{mk} is equal to the number of the inversions of the columns to which the considered non-zero elements of the determinant in (6) belong. Now we shall determine the coefficient of z^{m^2} and the exponent ν_{mk} in (16) with the help of the following method: From (1)–(4) we obtain the limit equations

$$\lim_{z\to\infty} \frac{\bar{a}_0 - \bar{q}}{z^m} = -\bar{b}_m \tag{17}$$

and

$$\lim_{z\to\infty} \frac{b_0 - p}{z^m} = -a_m. \tag{18}$$

From (16)–(18) it follows that the coefficient of z^{m^2} is $(-1)^m\Delta_{2m}(b_m,a_m)$, where

$$\Delta_{2m}(b_m,a_m)=\sum_{k=0}^{m}\sum_{k}(-1)^{\nu_{mk}}|b_m|^{2k}|a_m|^{2m-2k}. \tag{19}$$

On the other hand, we can determine directly the coefficient of z^{m^2} from (6). If we take out a factor z^m of each one of the last m columns of the determinant in (6) and set $z\to\infty$, then by means of (17)–(18) we obtain that the coefficient of z^{m^2} is $(-1)^m\Delta_{2m}(b_m,a_m)$, where

$$\Delta_{2m}(b_m,a_m)\equiv\Delta_{2m}\begin{pmatrix} b_m, & \dots, & b_1\\ \bar{a}_m, & \dots, & \bar{a}_1\end{pmatrix}$$

$$\equiv\begin{array}{c|ccccccccc|} & b_m & b_{m-1} & \dots & b_1 & a_m & 0 & 0 & \dots & 0\\ m & 0 & b_m & \dots & b_2 & 0 & a_m & 0 & \dots & 0\\ \text{rows} & & \dots & \dots & \dots & \dots & \dots & & & \\ & 0 & 0 & \dots & b_m & 0 & 0 & 0 & \dots & a_m\\ & \bar{a}_m & \bar{a}_{m-1} & \dots & \bar{a}_1 & \bar{b}_m & 0 & 0 & \dots & 0\\ m & 0 & \bar{a}_m & \dots & \bar{a}_2 & 0 & \bar{b}_m & 0 & \dots & 0\\ \text{rows} & & \dots & \dots & \dots & \dots & \dots & & & \\ & 0 & 0 & \dots & \bar{a}_m & 0 & 0 & 0 & \dots & \bar{b}_m\end{array}, \tag{20}$$

and the determinant is of order $2m$. Now we develop the determinant (20) with respect to the first column and again we develop the obtained two subdeterminants with respect to the m-th columns, respectively. Thus we obtain the recurrence relation

$$\Delta_{2m}\begin{pmatrix} b_m, & \dots, & b_1\\ \bar{a}_m, & \dots, & \bar{a}_1\end{pmatrix}=\left(|b_m|^2-|a_m|^2\right)\Delta_{2m-2}\begin{pmatrix} b_m, & \dots, & b_2\\ \bar{a}_m, & \dots, & \bar{a}_2\end{pmatrix} \tag{21}$$

for $m\geq 2$, where

$$\Delta_2\begin{pmatrix} b_m\\ \bar{a}_m\end{pmatrix}=\begin{vmatrix} b_m & a_m\\ \bar{a}_m & \bar{b}_m\end{vmatrix}=|b_m|^2-|a_m|^2. \tag{22}$$

From (21)–(22), by induction on m, we get the formula

$$\Delta_{2m}(b_m,a_m)=\left(|b_m|^2-|a_m|^2\right)^m \tag{23}$$

for $m\geq 1$, keeping in mind the notations (20). Hence the resolvent equation (6) for $n=m$ is exactly of degree m^2 if $|a_m|\neq|b_m|$, and of degree less than m^2 if $|a_m|=|b_m|$. Further we compare (19) with the binomial expansion of (23). This yields the formula

$$\nu_{mk}=m-k. \tag{24}$$

By means of the formula (24) we find that the part (16) of the development of the determinant (6) for $n=m$ has the form

$$[b_m(\bar{a}_0-\bar{q})-\bar{a}_m(b_0-p)]^m=\left[b_m\bar{a}_0-\bar{a}_m b_0+\sum_{s=0}^{m}(\bar{a}_m a_s-b_m\bar{b}_s)z^s\right]^m, \tag{25}$$

keeping in mind (1)–(4). Finally, from (25) for $s = m$ again our assertion for the degree becomes evident.

(iii) Let $n < m$. Then we interchange the roles of n and m, i.e. we examine the case $m > n$ as in point (i). Hence the resolvent equation (6) is exactly of degree m^2 with respect to z.

This completes the proof of Theorem 1.

Now we shall examine the equation (5) for $n = m$ under the conditions (7). In this case from (5) and (7), keeping in mind (1)–(4), we obtain the corresponding two equations

$$\sum_{s=1}^{m} b_s \bar{z}^s = e^{i\varphi} \sum_{s=1}^{m} \bar{b}_s z^s \pm i r_0 e^{i\frac{\varphi}{2}}, \tag{26}$$

which coincide with their conjugate equations

$$\sum_{s=1}^{m} \bar{b}_s z^s = e^{-i\varphi} \sum_{s=1}^{m} b_s \bar{z}^s \mp i r_0 e^{-i\frac{\varphi}{2}},$$

respectively, if the last equations are multiplied by $e^{i\varphi}$.

Theorem 2. *The two equations* (26) *are indeterminate, i.e. they have infinitely many roots* z.

Proof. We set

$$b_s = r_s e^{i\beta_s}, \quad 1 \leq s \leq m, \tag{27}$$

where $r_s \geq 0$ ($r_m > 0$), β_s are real (β_s is arbitrary if the corresponding $r_s = 0$), and

$$z = \rho e^{i\psi}, \tag{28}$$

where $\rho \geq 0$, ψ is real (ψ is arbitrary if $\rho = 0$). Then by means of (27) and (28) the two equations (26) become

$$\sum_{s=1}^{m} \rho^s r_s e^{i(\beta_s - s\psi)} = \sum_{s=1}^{m} \rho^s r_s e^{i(\varphi - \beta_s + s\psi)} \pm i r_0 e^{i\frac{\varphi}{2}},$$

which, after multiplication by $e^{-i\frac{\varphi}{2}}$, takes the form

$$2\sum_{s=1}^{m} \rho^s r_s \sin\left(s\psi - \beta_s + \frac{\varphi}{2}\right) \pm r_0 = 0. \tag{29}$$

The equations (29) are indeterminate with respect to ρ and ψ, depending on r_0, φ, r_s and β_s ($1 \leq s \leq m$).

This completes the proof of Theorem 2.

EXAMPLES AND APPLICATIONS

1. In particular, if $m = 1$, $b_1 = 1$, $b_0 = 0$ ($\bar{q} = Q(z) = z$) and $n \geq 1$ ($p = P(z)$), the equation (5) is reduced to the equation

$$\bar{z} = P(z), \tag{30}$$

keeping in mind (1). According to (6), the resolvent equation of (30) is the equation

$$D_{n+1}(\bar{a}_n, \bar{a}_{n-1}, \dots, \bar{a}_1, \bar{a}_0 - z) \equiv$$

$$\equiv \begin{array}{r} n \\ \text{rows} \end{array}\left\{\begin{array}{l} \\ \\ \\ \\ \end{array}\right. \!\!\! \begin{vmatrix} 1 & -p & 0 & \dots & \dots & 0 & 0 \\ 0 & 1 & -p & 0 & \dots & 0 & 0 \\ \multicolumn{7}{c}{\dotfill} \\ 0 & 0 & 0 & \dots & \dots & 1 & -p \\ \bar{a}_n & \bar{a}_{n-1} & \dots & \dots & \dots & \bar{a}_1 & \bar{a}_0 - z \end{vmatrix} = 0, \quad \begin{array}{l} \\ \\ \\ \\ 1 \text{ row} \end{array} \tag{31}$$

where the determinant is of order $n+1$. If we develop the determinant in (31) by the first column, then we obtain the recurrence relation

$$D_{n+1}(\bar{a}_n, \bar{a}_{n-1}, \dots, \bar{a}_1, \bar{a}_0 - z) = D_n(\bar{a}_{n-1}, \bar{a}_{n-2}, \dots, \bar{a}_1, \bar{a}_0 - z) + \bar{a}_n p^n \tag{32}$$

for $n \geq 2$, where

$$D_2(\bar{a}_1, \bar{a}_0 - z) = \begin{vmatrix} 1 & -p \\ \bar{a}_1 & \bar{a}_0 - z \end{vmatrix} = \bar{a}_0 - z + \bar{a}_1 p. \tag{33}$$

From (32) and (33) by induction on n we get the resolvent equation (31) of the equation (30) in the form

$$\bar{a}_n p^n + \bar{a}_{n-1} p^{n-1} + \dots + \bar{a}_1 p + \bar{a}_0 - z = 0, \tag{34}$$

keeping in mind (1). We shall note that the equation (34) follows directly from (30) with the help of the conjugate equation $z = \overline{P(z)}$ as well. If $n > 1$, the resolvent equation (34) is of degree n^2 and hence the given equation (30) has at most n^2 roots determined by (34). If $n = 1$, the resolvent equation (34) ($p = a_1 z + a_0$) is

$$\left(|a_1|^2 - 1\right) z + \bar{a}_1 a_0 + \bar{a}_0 = 0, \quad a_1 \neq 0. \tag{35}$$

Thus:

(I) If $|a_1| \neq 1$, from (35) it follows that the equation (30) ($n = 1$) has only one root which is

$$z = -\frac{\bar{a}_1 a_0 + \bar{a}_0}{|a_1|^2 - 1};$$

(II) If $|a_1| = 1$, i.e. $a_1 = e^{i\varphi}$, φ is real, the resolvent equation (35) is reduced to

$$0.z + e^{-i\varphi} a_0 + \bar{a}_0 = 0. \tag{36}$$

Now:

(II$_1$) If $e^{-i\varphi} a_0 + \bar{a}_0 \neq 0$, i.e. $a_0 \neq \pm i r_0 e^{i\frac{\varphi}{2}}$, $r_0 \geq 0$, the resolvent equation (36), and hence the given equation (30) for $n = 1$ and $a_1 \doteq e^{i\varphi}$, i.e. the equation $\bar{z} = e^{i\varphi} z + a_0$, has not a root;

(II$_2$) If $e^{-i\varphi} a_0 + \bar{a}_0 = 0$, i.e. $a_0 = \pm i r_0 e^{i\frac{\varphi}{2}}$, $r_0 \geq 0$, the resolvent equation (36) is the identity

$$0.z + 0 = 0.$$

This is so, since for $m = 1$ the two equations (7) exist simultaneously ($b_1 = 1$, $b_0 = 0$, $\bar{a}_1 = e^{-i\varphi}$, $a_0 = \pm i r_0 e^{i\frac{\varphi}{2}}$). In this case the given equation (30) ($n = 1$) yields the two equations

$$\bar{z} = e^{i\varphi} z \pm i r_0 e^{i\frac{\varphi}{2}}, \tag{37}$$

which coincide with their conjugate equations

$$z = e^{-i\varphi}\bar{z} \mp i r_0 e^{-i\frac{\varphi}{2}},$$

respectively, if the last equations are multiplied by $e^{i\varphi}$. If we set $z = \rho e^{i\psi}$, $\rho \geq 0$, ψ is real (ψ is arbitrary if $\rho = 0$), then from (37), after multiplication by $e^{-i\frac{\varphi}{2}}$, we obtain the corresponding two indeterminate equations

$$2\rho \sin\left(\psi + \frac{\varphi}{2}\right) \pm r_0 = 0,$$

which yield the unknown values ρ and ψ, depending on r_0 and φ.

2. In particular, if $m = 2$, $b_2 = 1$, b_1 is arbitrary, $b_0 = 0$ ($\bar{q} = \overline{Q(\bar{z})} = z^2 + \bar{b}_1 z$) and $n = 2$ ($p = P(z) = a_2 z^2 + a_1 z + a_0$, $a_2 \neq 0$), the equation (5) is reduced to the equation

$$\bar{z}^2 + b_1 \bar{z} = a_2 z^2 + a_1 z + a_0. \tag{38}$$

For this case the equations (7) ($m = 2$) are

$$\bar{a}_2 = e^{-i\varphi}, \quad \bar{a}_1 = b_1 e^{-i\varphi}, \quad a_0 = \pm i r_0 e^{i\frac{\varphi}{2}} \tag{39}$$

with $r_0 \geq 0$ and an arbitrary real φ. From (38) and (39) we obtain the two equations

$$\bar{z}^2 + b_1 \bar{z} = e^{i\varphi} z^2 + \bar{b}_1 e^{i\varphi} z \pm i r_0 e^{i\frac{\varphi}{2}}, \tag{40}$$

which coincide with their conjugate equations

$$z^2 + \bar{b}_1 z = e^{-i\varphi}\bar{z}^2 + b_1 e^{-i\varphi}\bar{z} \mp i r_0 e^{-i\frac{\varphi}{2}},$$

respectively, if the last equations are multiplied by $e^{i\varphi}$. If we set $z = \rho e^{i\psi}$, $\rho \geq 0$, ψ is real (ψ is arbitrary if $\rho = 0$), then from (40), after multiplication by $e^{-i\frac{\varphi}{2}}$, we obtain the corresponding two indeterminate equations

$$2\rho^2 \sin\left(2\psi + \frac{\varphi}{2}\right) + 2\rho r_1 \sin\left(\psi - \beta_1 + \frac{\varphi}{2}\right) \pm r_0 = 0,$$

which yield the unknown values ρ and ψ, depending on r_0, φ, $r_1 = |b_1|$ and $\beta_1 = \operatorname{Arg} b_1$ (β_1 is arbitrary if $r_1 = 0$).

In the general case, if the equations (39) do not exist simultaneously, then according to (6) ($m = n = 2$) the resolvent equation of (38) is the equation

$$(\bar{a}_0 - \bar{q} + \bar{a}_2 p)^2 + (\bar{a}_2 b_1 - \bar{a}_1)\left[b_1(\bar{a}_0 - \bar{q}) + \bar{a}_1 p\right] = 0, \tag{41}$$

where

$$\bar{a}_0 - \bar{q} + \bar{a}_2 p = \left(|a_2|^2 - 1\right) z^2 + \left(a_1\bar{a}_2 - \bar{b}_1\right) z + a_0\bar{a}_2 + \bar{a}_0 \tag{42}$$

and

$$b_1(\bar{a}_0 - \bar{q}) + \bar{a}_1 p = (\bar{a}_1 a_2 - b_1) z^2 + \left(|a_1|^2 - |b_1|^2\right) z + a_0\bar{a}_1 + \bar{a}_0 b_1. \tag{43}$$

If $|a_2| \neq 1$, then from (41)–(43) it follows that the resolvent equation (41) is of degree 4 and hence the given equation (38) has at most four roots z. If $|a_2| = 1$, then from (41)–(43) it follows that the resolvent equation (41) is of degree at most 2 and hence the given equation (38) has at most two roots z.

3. In particular, for $a_5 = 0$, $0 \leq s \leq n-1$, $a_n \neq 0$, $b_s = 0$, $0 \leq s \leq m-1$, $b_m \neq 0$, $n \geq m \geq 1$ and $|a_m| \neq |b_m|$, if $n = m$ ($p = P(z) = a_n z^n$, $\bar{q} = \overline{Q(\bar{z})} = \bar{b}_m z^m$), from (5) and (6) we obtain the equation

$$b_m \zeta^m = a_n z^n \qquad (\zeta = \bar{z}) \tag{44}$$

and its resolvent equation

$$E_{nm}(a_n, b_m, z) \equiv$$

$$\equiv \begin{array}{r|cccccccccc|} & b_m & 0 & \dots & 0 & -a_n z^n & 0 & 0 & \dots & 0 \\ n & 0 & b_m & \dots & 0 & 0 & -a_n z^n & 0 & \dots & 0 \\ \text{rows} & & \dots & \dots & \dots & \dots & \dots & & & \\ & 0 & 0 & \dots & b_m & 0 & 0 & 0 & \dots & -a_n z^n \\ & \bar{a}_n & 0 & \dots & 0 & -\bar{b}_m z^m & 0 & 0 & \dots & 0 \\ m & 0 & \bar{a}_n & \dots & 0 & 0 & -\bar{b}_m z^m & 0 & \dots & 0 \\ \text{rows} & & \dots & \dots & \dots & \dots & \dots & & & \\ & 0 & 0 & \dots & \bar{a}_n & 0 & 0 & 0 & \dots & -\bar{b}_m z^m \end{array} = 0, \tag{45}$$

where the determinant $E_{nm}(a_n, b_m, z)$ is of order $n + m$. The equation (45) is a result of the elimination of ζ from the equation (44) and the conjugate equation

$$\bar{a}_n \zeta^n = \bar{b}_m z^m \qquad (\zeta = \bar{z}). \tag{46}$$

Now we can eliminate ζ by means of another method. Namely, let

$$d \equiv (n, m) \qquad (1 \leq d \leq m) \tag{47}$$

denote the greatest common divisor of the numbers n and m, i.e.

$$n = n_1 d \quad \text{and} \quad m = m_1 d, \tag{48}$$

where n_1 ($1 \leq n_1 \leq n$) and m_1 ($1 \leq m_1 \leq m$) are the corresponding quotients which are relatively prime positive integers, i.e. their greatest common divisor $(n_1, m_1) = 1$. Since the product $n_1 m_1 d$ is the least common multiple of the numbers n and m, from (44), (46) and (48) we obtain the equations

$$\zeta^{n_1 m_1 d} = \left(\frac{a_n}{b_m}\right)^{n_1} z^{n_1^2 d} \tag{49}$$

and

$$\zeta^{n_1 m_1 d} = \left(\frac{\bar{b}_m}{\bar{a}_n}\right)^{m_1} z^{m_1^2 d}. \tag{50}$$

From (49) we obtain d equations

$$\zeta^{n_1 m_1} = \varepsilon^k z^{n_1^2} \sqrt[d]{\left(\frac{a_n}{b_m}\right)^{n_1}}, \qquad k = 1, \dots, d, \tag{51}$$

where

$$\varepsilon = e^{i\frac{2\pi}{d}} \tag{52}$$

and the radical is taken arbitrarily. Hence from (50)–(52) we obtain d equations of the form

$$\left(\frac{a_n}{b_m}\right)^{n_1} z^{n_1^2 d} - \left(\frac{\bar{b}_m}{\bar{a}_n}\right)^{m_1} z^{m_1^2 d} = 0, \tag{53}$$

which yield all roots z of the equation (44). Thus from (53) and (48) we get the resolvent equation

$$\left[\left(\frac{a_n}{b_m}\right)^{\frac{n}{d}} z^{\frac{n^2}{d}} - \left(\frac{\bar{b}_m}{\bar{a}_n}\right)^{\frac{m}{d}} z^{\frac{m^2}{d}}\right]^d = 0 \tag{54}$$

of the equation (44), keeping in mind the multiplicity of the roots z, where d is given by (47). Further, from the comparison of the equivalent equations (45) and (54) it follows that

$$E_{nm}(a_n, b_m, z) = \mu_{nm}\left(a_n^{\frac{n}{d}} \bar{a}_n^{\frac{m}{d}} z^{\frac{n^2}{d}} - b_m^{\frac{n}{d}} \bar{b}_m^{\frac{m}{d}} z^{\frac{m^2}{d}}\right)^d, \tag{55}$$

where μ_{nm} is a factor which does not depend on z. Now we shall determine μ_{nm}. From (55) we obtain

$$\left.\frac{E_{nm}(a_n, b_m, z)}{z^{m^2}}\right|_{z=0} = (-1)^d \mu_{nm} b_m^n \bar{b}_m^m \tag{56}$$

for $n > m \geq 1$, and

$$\left.\frac{E_{mm}(a_m, b_m, z)}{z^{m^2}}\right|_{z=0} = \mu_{mm}\left(|a_m|^2 - |b_m|^2\right)^m \tag{57}$$

for $n = m \geq 1$, keeping in mind that $d = (m, m) = m$. On the other hand, from (45) we obtain

$$\left.\frac{E_{nm}(a_n, b_m, z)}{z^{m^2}}\right|_{z=0} = (-1)^m b_m^n \bar{b}_m^m \tag{58}$$

for $n > m \geq 1$, and

$$\left.\frac{E_{mm}(a_m, b_m, z)}{z^{m^2}}\right|_{z=0} = (-1)^m \left(|b_m|^2 - |a_m|^2\right)^m \tag{59}$$

for $n = m \geq 1$, keeping in mind (20) (for $a_s = b_s = 0$, $1 \leq s \leq m-1$, if $m \geq 2$) and (23). If we compare (56) with (58) and (57) with (59), we obtain

$$\mu_{nm} = (-1)^{m-d}, \qquad n \geq m \geq 1. \tag{60}$$

Thus from (55) and (60) we get the formula

$$E_{nm}(a_n, b_m, z) = (-1)^{m-d}\left(a_n^{\frac{n}{d}} \bar{a}_n^{\frac{m}{d}} z^{\frac{n^2}{d}} - b_m^{\frac{n}{d}} \bar{b}_m^{\frac{m}{d}} z^{\frac{m^2}{d}}\right)^d \tag{61}$$

for the value of the determinant in (45) for $n \geq m \geq 1$ and d given by (47).

In particular, for $n = rm$, $r = 1, 2, \dots$ $(m \geq 1)$ we have $d = (rm, m) = m$ and hence the formula (61) is reduced to the formula

$$E_{rm,m}(a_{rm}, b_m, z) = \left(a_{rm}^r \bar{a}_{rm} z^{r^2 m} - b_m^r \bar{b}_m z^m\right)^m. \tag{62}$$

In particular, from (44) for $n = m \geq 1$ and (62) for $r = 1$ it follows that all roots z of the equation

$$b_m \bar{z}^m = a_m z^m, \qquad |a_m|^2 - |b_m|^2 \neq 0, \tag{63}$$

are represented by the multiple root $z = 0$ of order m^2 of the resolvent equation

$$\left(|a_m|^2 - |b_m|^2\right)^m z^{m^2} = 0. \tag{64}$$

The resolvent equation (64) can be directly obtained if we determine $\bar{z}$ from (63), which yields

$$\bar{z} = ze^{-i\frac{2k\pi}{m}} \sqrt[m]{\frac{a_m}{b_m}}, \qquad k = 0, 1, \dots, m-1,$$

for any value of the radical, and set these values of $\bar{z}$ in the conjugate equation of (63), namely in

$$\bar{b}_m z^m = \bar{a}_m \bar{z}^m.$$

Thus we obtain m equations of the form

$$\left(|a_m|^2 - |b_m|^2\right) z^m = 0$$

which, when multiplied, yield (64).

OTHER EXAMPLES

The next simple cases illustrate the application of example 1.

(A) Consider the equation

$$\bar{z} = z. \tag{65}$$

The conjugate equation of (65) is $z = \bar{z}$ and hence the resolvent equation is the identity $z = z$, i.e. the equation

$$0.z = 0. \tag{66}$$

The solutions of (66) are all complex numbers, but the solutions of (65) are only all real numbers, because the root $\zeta = z$ of the equation $\zeta - z = 0$ is equal to $\bar{z}$ if and only if z is a real number. This result is in accordance with Theorem 2 and example 1, item (II$_2$), for $\varphi = 0$ and $r_0 = 0$.

(B) Consider the equation

$$\bar{z} = z^2. \tag{67}$$

The equation (67) and its conjugate equation form the two equations

$$\zeta - z^2 = 0, \qquad \zeta^2 - z = 0. \tag{68}$$

From (6), or directly from (68), we obtain the resolvent equation

$$z(z^3 - 1) = 0. \tag{69}$$

All solutions $z = 0,\ 1,\ e^{i\frac{2\pi}{3}},\ e^{i\frac{4\pi}{3}}$ of (69) are roots of (67), because for these z the corresponding common root ζ of the two equations (68) is equal to the conjugate value $\bar{z}$, respectively.

(C) Consider the equation

$$\bar{z} = z^3 + z. \tag{70}$$

The equation (70) and its conjugate equation form the two equations

$$\zeta - (z^3 + z) = 0, \qquad \zeta^3 + \zeta - z = 0. \tag{71}$$

From (6), or directly from (71), we obtain the resolvent equation

$$0 = (z^3 + z)^3 + z^3 = z^3(z^2 + 2)\frac{z^6 - 1}{z^2 - 1}, \qquad z^2 \neq 1, \tag{72}$$

with the roots

$$z_{1,2,3} = 0, \quad z_{4,5} = \pm i\sqrt{2}, \quad z_6 = e^{i\frac{\pi}{3}}, \quad z_7 = e^{i\frac{2\pi}{3}}, \quad z_8 = e^{i\frac{4\pi}{3}}, \quad z_9 = e^{i\frac{5\pi}{3}}. \tag{73}$$

For the values $z = z_k$, $k = 1, 2, 3, 4, 5$, in (73), the common roots ζ of the two equations (71) are equal to $\zeta = \bar{z}_k$, $k = 1, 2, 3, 4, 5$, respectively. Hence the roots $z_{1,2,3}$ (a triple root) and $z_{4,5}$ of (72) are roots of (70) as well. For the values $z = z_k$, $k = 6, 7, 8, 9$, in (73), the two equations (71) take the forms

$$\zeta - z_{7,6,9,8} = 0, \qquad \zeta^3 + \zeta - z_{6,7,8,9} = 0, \tag{74}$$

respectively. The common roots ζ of the two equations (74) are equal to $\zeta = z_{7,6,9,8} \neq \bar{z}_{6,7,8,9}$, respectively. Hence the roots $z_{6,7,8,9}$ of the resolvent equation (72) are not roots of the given equation (70). Thus all roots of (70) are only the roots $z_{1,2,3,4,5}$ in (73).

(D) Consider the equation

$$\bar{z} = z^4. \tag{75}$$

The equation (75) and its conjugate equations form the two equations

$$\zeta - z^4 = 0, \qquad \zeta^4 - z = 0. \tag{76}$$

From (6), or directly from (76), we obtain the resolvent equation

$$z(z^{15} - 1) = 0. \tag{77}$$

But only the solutions

$$z = 0,\ 1,\ e^{i\frac{2\pi}{5}},\ e^{i\frac{4\pi}{5}},\ e^{i\frac{6\pi}{5}},\ e^{i\frac{8\pi}{5}}$$

of (77) are the unique solutions of (75), because only for these z the corresponding common root ζ of the two equations (76) is equal to the conjugate value $\bar{z}$, respectively.

The examples (B)–(D) are in accordance with Theorem 1.

REFERENCES

1. Dickson, L. E. Elementary Theory of Equations. John Wiley and Sons, Inc., New York, 1914.

Received February 6, 1998

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bontchev Str., Block 8
1113 Sofia, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

DIRECT BUILDING OF MINIMAL AUTOMATON FOR A GIVEN LIST

STOYAN MIHOV

This paper presents a method for direct building of minimal acyclic finite states automaton which recognizes a given finite list of words in lexicographical order. The size of the temporary automata which are necessary for the construction is less than the size of the resulting minimal automata plus the length of one of the longest words in the list. This property is the main advantage of our method.

Keywords: acyclic automata, automata building

1991/95 Math. Subject Classification: 68Q68

1. INTODUCTION

The standard methods for building minimal finite states (FS) automaton are building temporary automata which generally are huge compared to the resulting minimal automaton. This grounds the interest in the development of more direct methods. Building an acyclic minimal FS automaton that recognizes a given list sorted in lexicografical order is of special interest for practical applications. A linear algorithm for that case is presented in [1, 2]. The Revuz' method in the first stage builds a tree-like deterministic FS automaton. Then at the second stage this automaton is minimized efficiently. The drawback of this method is that the tree-like automaton is huge in respect of the resulting minimal automata. To make this method more efficient, Roche [3] proposes to divide the list into parts for which to build the corresponding minimal automata. After that those automata are united. At the end it is necessary to minimize the result.

It is claimed [1] that a method for direct building of minimal FS automaton for a given list does not exist. We will show that in general this statement is not valid. We shall present our new method for direct building of minimal automaton.

2. FORMAL BACKGROUND AND NOTATIONS

Definition 1. A deterministic FS automaton is a tuple $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$, where:

- Σ is a finite alphabet;
- S is a finite set of states;
- $s \in S$ is the starting state;
- $F \subseteq S$ is the set of final states;
- $\mu : S \times \Sigma \to S$ is a partial function called the transition function.

The function μ is extended naturally over $S \times \Sigma^*$ by induction:

$$\begin{cases} \mu^*(r, \varepsilon) = r, \\ \mu^*(r, \sigma a) = \begin{cases} \mu(\mu^*(r,\sigma), a), & \text{in case } \mu^*(r,\sigma) \text{ and } \mu(\mu^*(r,\sigma),a) \text{ are defined,} \\ \text{not defined,} & \text{otherwise,} \end{cases} \end{cases}$$

where $r \in S$, $\sigma \in \Sigma^*$, $a \in \Sigma$.

We shall work with a definition of FS automata with a partial transition function. The only difference from the definition with a total transition function is the absence of the necessity to introduce a dead state (a state r, for which $\forall a \in \Sigma\ (\mu(r,a) = r)$). Later we use $!\mu(r,\sigma)$ to denote that $\mu(r,\sigma)$ is defined and when writing $\mu^*(r,\sigma) \cong x$, we mean $!\mu(r,\sigma)\ \&\ \mu(r,\sigma) = x$.

Definition 2. Let $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ be a deterministic FS automaton. Then the set $L(\mathcal{A}) \subseteq \Sigma^*$, defined as

$$L(\mathcal{A}) = \{\sigma \in \Sigma^* \mid !\mu^*(s,\sigma)\ \&\ \mu^*(s,\sigma) \in F\},$$

is called the language of the automaton $\mathcal{A}$ or the language recognized by $\mathcal{A}$.

Two automata $\mathcal{A}$ and $\mathcal{A}'$ are called equivalent when $L(\mathcal{A}) = L(\mathcal{A}')$. An automaton is called acyclic when $\forall r \in S\ \forall \sigma \in \Sigma^+\ (\mu^*(r,\sigma) \not\cong r)$. The language of an acyclic FS automaton is finite.

Definition 3. Let $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ be a deterministic FS automaton:

1. The state $r \in S$ is called reachable from $t \in S$ when $\exists \sigma \in \Sigma^*\ (\mu^*(t,\sigma) \cong r)$.

2. We define the subautomaton starting in $s' \in S$ as $\mathcal{A}|_{s'} = \langle \Sigma, S', s', F \cap S', \mu|_{S' \times \Sigma} \rangle$, where $S' = \{r \in S \mid r \text{ is reachable from } s'\}$.

3. Two states $s_1, s_2 \in S$ are called equivalent when $L(\mathcal{A}|_{s_1}) = L(\mathcal{A}|_{s_2})$.

Definition 4. The deterministic FS automaton $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ with language $L(\mathcal{A})$ is called minimal (with language $L(\mathcal{A})$) when for every other deterministic FS automaton $\mathcal{A}' = \langle \Sigma, S', s', F', \mu' \rangle$ with language $L(\mathcal{A}') = L(\mathcal{A})$, it holds that $|S| \leq |S'|$.

From the classical FS theory the next theorem is well-known.

Theorem 5. *A deterministic FS automaton with non-empty language is minimal if and only if every state is reachable from the starting state, from every state a final state is reachable and there are no different equivalent states. There exists an unique (up to isomorphism) minimal automaton for a given language.*

3. METHOD DESCRIPTION

Further we assume that a finite alphabet Σ is given and there is a linear order in Σ. This order induces a lexicographical order in Σ^*.

Definition 6. Let $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ be an acyclic deterministic FS automaton with language $L(\mathcal{A})$. Then the automaton $\mathcal{A}$ is called *minimal except for the word* $\omega \in \Sigma^*$ when the following conditions hold:

1. Every state is reachable from the starting state and from every state a final state is reachable.

2. ω is a prefix of the last word in the lexicographical order of $L(\mathcal{A})$.

In that case we can introduce the notations

$$\omega = w_1^{\mathcal{A}} w_2^{\mathcal{A}} \ldots w_k^{\mathcal{A}}, \quad \text{where } w_i^{\mathcal{A}} \in \Sigma \text{ for } i = 1, 2, \ldots, k, \tag{1}$$

$$t_0^{\mathcal{A}} = s; \;\; t_1^{\mathcal{A}} = \mu(t_0^{\mathcal{A}}, w_1^{\mathcal{A}}); \;\; t_2^{\mathcal{A}} = \mu(t_1^{\mathcal{A}}, w_2^{\mathcal{A}}); \;\; \ldots; \;\; t_k^{\mathcal{A}} = \mu(t_{k-1}^{\mathcal{A}}, w_k^{\mathcal{A}}), \tag{2}$$

$$T = \{t_0^{\mathcal{A}}, t_1^{\mathcal{A}}, \ldots, t_k^{\mathcal{A}}\}. \tag{3}$$

3. In the set $S \setminus T$ there are no different equivalent states.

4. $\forall r \in S \, \forall i \in \{0, 1, \ldots, k\} \forall a \in \Sigma \left(\mu(r, a) \cong t_i \leftrightarrow (i > 0 \& r = t_{i-1} \& a = \omega_i^{\mathcal{A}})\right)$.

Further, when working with minimal except for a given word automaton, we use the notations (1)–(3). In case the notation is not ambiguous, we write t_i, w_i instead of $t_i^{\mathcal{A}}$, $w_i^{\mathcal{A}}$. Clearly, if an automaton is minimal except for two different words, one is a prefix of the other.

Proposition 7. *Let the automaton* $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ *be minimal except for* ω. *Then:*

1. $\forall r \in S \setminus T \, \forall a \in \Sigma \, (!\mu(r, a) \rightarrow \mu(r, a) \in S \setminus T)$;

2. $\mu^*(s, \sigma) \cong t_i \rightarrow \sigma = w_1 w_2 \ldots w_i$.

The proof of this proposition is derived directly from Definition 6.

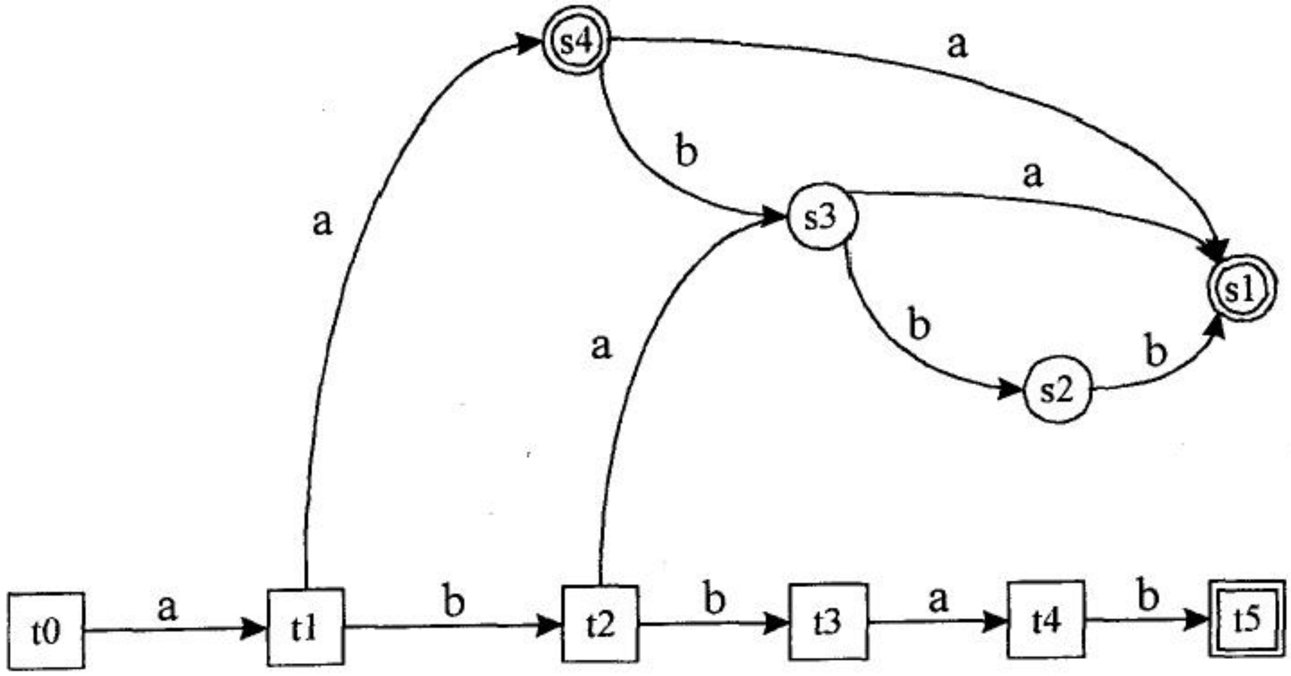

Fig. 1. The FS automaton minimal except for *abbab*

Example 1. On Fig. 1 an acyclic FS automaton over the alphabet $\{a, b\}$ is given. The language of the automaton is $\{aa, aaa, aaba, aabbb, abaa, ababb, abbab\}$. This automaton is minimal except for *abbab*.

Proposition 8. *An automaton which is minimal except for the empty word* ε *is minimal.*

Proof. Every state is reachable from the starting state and from every state a final state is reachable. Hence, to prove that the automaton is minimal, we have to show that there are no equivalent states. From the definition we know that there are no equivalent states in $S \setminus \{s\}$.

Let assume that $r \in S$ and s are equivalent and $r \neq s$. Let ω be one of the longest words recognized by the automaton. There exists longest word(s), because the language is finite. The states r and s are equivalent, hence $\omega \in L(\mathcal{A}|_r)$. The state r is reachable from s. Hence there exists $\sigma \in \Sigma^*$ and $\mu^*(s, \sigma) \cong r$. Then the word $\sigma\omega \in L(\mathcal{A})$ and from $\sigma \neq \varepsilon$ we have that $|\sigma\omega| > |\omega|$. This contradicts with the fact that ω is the longest word in the language. □

Lemma 9. *Let the automaton* $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ *be minimal except for* $\omega = w_1 w_2 \dots w_k$, $\omega \neq \varepsilon$. *Let there be no state equivalent to* t_k *in the set* $S \setminus T$. *Then* $\mathcal{A}$ *is also minimal except for the word* $\omega' = w_1 w_2 \dots w_{k-1}$.

Proof. We have to check the conditions of Definition 6. The conditions 1 and 2 are obviously satisfied. Condition 3 follows from the fact that in $S \setminus \{t_0, t_1, \dots, t_k\}$ there are no states equivalent to t_k. Condition 4 follows also directly from condition 4 of the definition for minimality except for ω. □

Lemma 10. *Let the automaton* $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ *be minimal except for*

$\omega = w_1 w_2 \dots w_k$, $\omega \neq \varepsilon$. Let the state $p \in S \setminus T$ be equivalent to the state t_k. Then the automaton $\mathcal{A}' = \langle \Sigma, S', s, F', \mu' \rangle$ defined as follows:

$$S' = S \setminus \{t_k\},$$

$$F' = F \setminus \{t_k\},$$

$$\mu'(r,a) = \begin{cases} \mu(r,a), & \textit{in case } r \neq t_{k-1} \vee a \neq w_k \textit{ and } \mu(r,a) \textit{ is definied,} \\ p, & \textit{in case } r = t_{k-1},\ a = w_k, \\ \textit{not defined,} & \textit{otherwise,} \end{cases}$$

is equivalent to the automaton $\mathcal{A}$ and is minimal except for the word $\omega' = w_1 w_2 \dots w_{k-1}$.

Proof. The automaton $\mathcal{A}$ is equivalent to $\mathcal{A}'$, because the new automaton is derived from the old one by removing the state t_k, and the only transition to t_k (refer to Proposition 7) is exchanged with a transition to an equivalent state. Conditions 1–3 from the definition for minimality except for ω' are trivially satisfied. Condition 4 is obviously satisfied for $t_0, t_1, \dots, t_{k-2}$, and holds also for t_{k-1}, because $\mu'(t_{k-1}, w_k) \cong p \in S \setminus T$. □

Theorem 11. *Let the automaton $\mathcal{A} = \langle \Sigma, S, s, F, \mu \rangle$ be minimal except for $\omega' = w_1 w_2 \dots w_m$. Let $\psi \in L(\mathcal{A})$ be the last word in the lexicographical order of the language of the automaton. Let ω be a word which is greater in lexicographical order than ψ. Let ω' be the longest common prefix of ψ and ω. In that case we can denote $\omega = w_1 w_2 \dots w_m w_{m+1} \dots w_k$, $k > m$. Then the automaton $\mathcal{A}' = \langle \Sigma, S', s, F', \mu' \rangle$ defined as follows:*

$$t_{m+1}, t_{m+2}, \dots, t_k \textit{ are new states such that } S \cap \{t_{m+1}, t_{m+2}, \dots, t_k\} = \emptyset,$$

$$S' = S \cup \{t_{m+1}, t_{m+2}, \dots, t_k\},$$

$$F' = F \cup \{t_k\},$$

$$\mu'(r,a) = \begin{cases} t_{m+1}, & \textit{in case } r = t_m,\ a = w_{m+1}, \\ \mu(r,a), & \textit{in case } r \in S,\ !\mu(r,a) \textit{ and } r \neq t_m \vee a \neq w_{m+1}, \\ t_{i+1}, & \textit{in case } r = t_i,\ m+1 \leq i \leq k-1,\ a = w_{i+1}, \\ \textit{is not defined,} & \textit{otherwise,} \end{cases}$$

is minimal except for ω and recognizes the language $L(\mathcal{A}) \cup \{\omega\}$.

Proof. First we shall show that $L(\mathcal{A}') = L(\mathcal{A}) \cup \{\omega\}$. We have to show that the automaton $\mathcal{A}'$ includes $\mathcal{A}$. Clearly, $S' \supset S$ and $F' \supset F$. We have to check that $\mu' \supset \mu$. Considering the definition of μ', it is clear that the only problem could be the case $\mu'(t_m, w_{m+1})$. But $\mu(t_m, w_{m+1})$ is not defined, because otherwise either ω' could not be the longest common prefix of ψ and ω or ω could not be greater in lexicographical order than ψ — the last word in the lexicographical order of the language of the automaton.

Hence we have that $L(\mathcal{A}') \supset L(\mathcal{A})$. New words could be recognized only by passing the additional new states. From the definition of μ' we have that they are reachable only from t_m. The only new word in the language $L(\mathcal{A}'|_{t_m})$ is $w_{m+1}w_{m+2}\dots w_k$. From Proposition 7 we have that t_m is reachable from s only by the word ω'. Hence the only new word in $L(\mathcal{A}')$ is the word $\omega' w_{m+1}w_{m+2}\dots w_k = \omega$.

We have to check that $\mathcal{A}'$ is minimal except for ω. Let us consider the conditions of Definition 6. Conditions 1–3 are obviously satisfied. Condition 4 is satisfied for $t_0, t_1, \dots, t_{m-1}$ from the definition of minimality except for ω' of the automaton $\mathcal{A}$, for the states $t_m, t_{m+1}, \dots, t_k$ the condition clearly holds because of the definition of μ'. □

Method for direct building of minimal FS automaton for a given list. Let a non-empty finite list of words L in lexicographical order be given. Let $\omega^{(i)}$ denote the i-th word of the list. We start with the minimal automaton which recognizes only the first word of the list. This automaton can be built trivially and it is also minimal except for $\omega^{(1)}$. Using it as basis, we carry out an induction on the words of the list. Let us assume that the automaton $\mathcal{A}^{(n)} = \langle \Sigma, S, s, F, \mu\rangle$ with language $L^{(n)} = \{\omega^{(i)} \,|\, i = 1, 2, \dots, n\}$ has been built and that $\mathcal{A}^{(n)}$ is minimal except for $\omega^{(n)}$. We have to build the automaton $\mathcal{A}^{(n+1)}$ with language $L^{(n+1)} = \{\omega^{(i)} \,|\, i = 1, 2, \dots, n+1\}$ which is minimal except for $\omega^{(n+1)}$.

Let ω' be the longest common prefix of the words $\omega^{(n)}$ and $\omega^{(n+1)}$. Using several times Lemma 9 and Lemma 10 (corresponding to the actual case), we build the automaton $\mathcal{A}' = \langle \Sigma, S', s, F', \mu'\rangle$ which is equivalent to $\mathcal{A}^{(n)}$ and is minimal except for ω'. Now we can use Theorem 11 and build the automaton $\mathcal{A}^{(n+1)}$ with language $L^{(n+1)} = L^{(n)} \cup \{\omega^{(n+1)}\} = \{\omega^{(i)} \,|\, i = 1, 2, \dots, n+1\}$ which is minimal except for $\omega^{(n+1)}$.

In this way by induction we build the minimal except for the last word of the list automaton with language the list L. At the end, using again Lemma 9 and Lemma 10, we build the automaton equivalent to the former one which is minimal except for the empty word. From Proposition 8 we have that it is the minimal automaton for the list L. The check of state equivalence needed to distinguish between Lemma 9 and Lemma 10 is performed efficiently using the following property:

$$t_k \text{ is equivalent to } r \in S \setminus T$$

$$\longleftrightarrow \; ((t_k \in F \leftrightarrow r \in F)\,\&\, \forall a \in \Sigma((\neg!\mu(t_k, a)\,\&\,\neg!\mu(r, a)) \\ \vee (!\mu(t_k, a)\,\&\,!\mu(r, a)\,\&\,\mu(t_k, a) = \mu(r, a))).$$

□

Clearly, all the temporary automata built during the construction of the resulting minimal automaton have less states than the resulting automaton plus the size of the longest word of the list. This is the main advantage of our method.

Example 2. To illustrate our method, let us consider the following example. On Fig. 1 the automaton recognizing the list $\{aa, aaa, aaba, aabbb, abaa, ababb,$

abbab} is given. This automaton is minimal except for the last word of the list — *abbab*. Let the next word be *baa*. The longest common prefix of those two words is ε. We have first to construct the automaton equivalent to the one on Fig. 1, which is minimal except for ε, by using Lemma 9 and Lemma 10. First we have to apply Lemma 9 twice and then to apply Lemma 10 three times. At the end, using Theorem 11, we construct the automaton which is minimal except for *baa*, given on Fig. 2. In this way we added the next word of the list to the language of the temporary automaton.

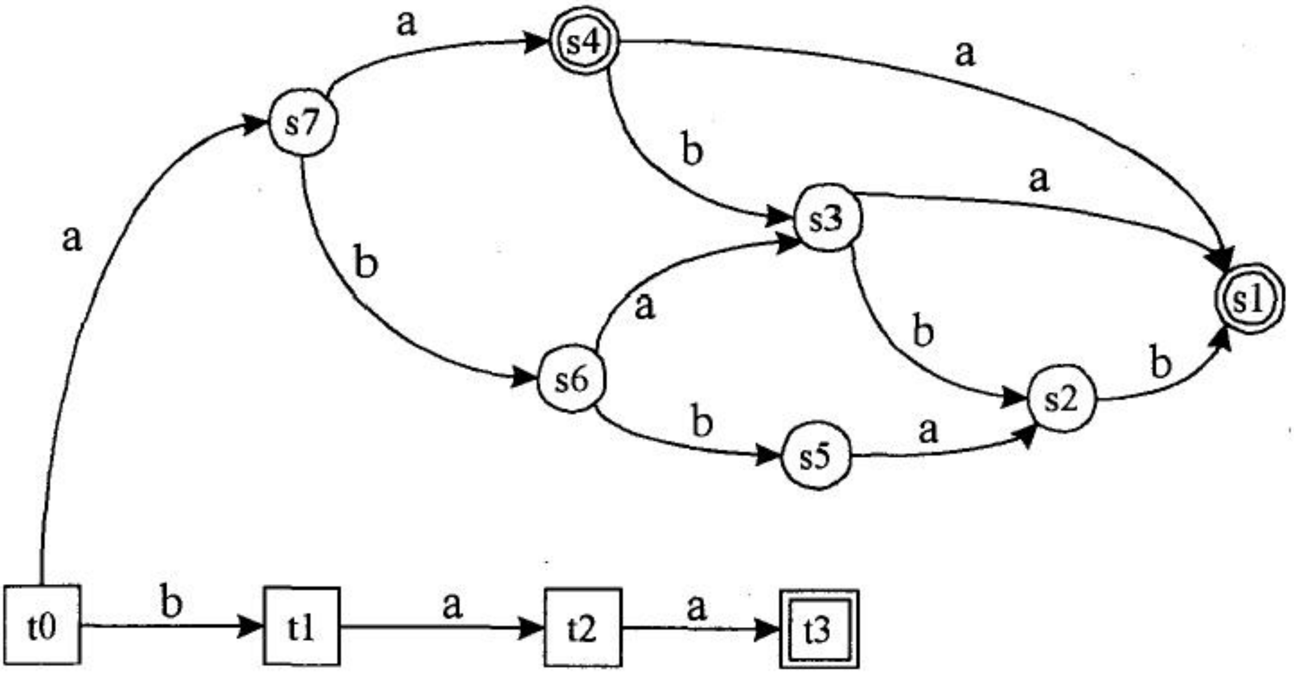

Fig. 2. The FS automaton minimal except for *baa*

4. CONCLUSION

The presented method for direct building of minimal automata can be extended for building minimal automata with labels on the final states and for automata which are returning the index of the recognized word in the list (for a presentation of those kinds of automata see, for example, [3]. In this way the method could be widely applied for building of large grammatical dictionaries, for indexing of huge lists, etc. The algorithms based on the method are distinguished with an excellent memory efficiency.

ACKNOWLEDGEMENTS. I would like to thank to professor Dimiter Skordev and Anton Zinoviev for the very valuable remarks on the presentation of our method, and to Bogdan Alexandrov who is a co-author of the idea for the presented method.

REFERENCES

1. Dominique Revuz. Dictionaires et lexiques, méthodes et algorithmes. Ph.D. dissertation, Université Paris 7, Paris, 1991.

2. Dominique Revuz. Minimization of acyclic deterministic automata in linear time. *Theoretical Computer Science*, **92**, 1, 1992.
3. Emmanuel Roche. Dictionary Compression Experiments. Technical report, LADL, Université Paris 7, Paris, 1992.

Received February 26, 1998

Linguistic Modelling Laboratory
LPDP — Bulgarian Academy of Sciences
E-mail: `stoyan@lml.acad.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

ON THE ALGEBRAIC PROPERTIES OF CONVEX BODIES AND RELATED ALGEBRAIC SYSTEMS

SVETOSLAV MARKOV

To the memory of Prof. Y. Tagamlitzki

The algebraic system of convex bodies with addition and multiplication by scalar is studied. A new operation for convex bodies, called inner addition, is introduced. New distributivity relations for convex bodies, called resp. quasidistributive and q-distributive law, are formulated and proved. The convex bodies form a quasilinear system with respect to addition and multiplication by scalar. The latter is isomorphically embedded in a q-linear system, which is an abelian group with respect to addition and,obeys the q-distributive law. A result of H. Radstróm for convex bodies is generalized.

Keywords: convex bodies, quasilinear systems

1991/95 Math. Subject Classification: main 52A01, secondary 13C99

1. INTRODUCTION

Notation. Let $\mathbb{E} = \mathbb{E}^n$, $n \geq 1$, be an n-dimensional real Euclidean vector space with origin 0. A convex compact subset of $\mathbb{E}$ is called *convex body* (*in* $\mathbb{E}$) or just a *body*; a convex body need not have necessarily interior points, e. g. a line segment and a single point in $\mathbb{E}$ are convex bodies [20]. The class of all convex bodies of $\mathbb{E}$ will be denoted by $\mathcal{K} = \mathcal{K}(\mathbb{E})$; in this work the empty set is not an element of $\mathcal{K}$. The elements of $\mathbb{E}$ are called one-point sets or degenerate bodies. The field of reals is denoted by $\mathbb{R}$.

The set $\mathcal{K}$ is closed under the operations

$$\begin{aligned} A+B &= \{c \mid c=a+b,\ a\in A,\ b\in B\},\ \ A,B\in\mathcal{K}, && (1.1)\\ \alpha * B &= \{c \mid c=\alpha b,\ b\in B\},\ \ B\in\mathcal{K},\ \alpha\in\mathbb{R}, && (1.2)\end{aligned}$$

called resp. (*Minkowski*) *addition* and *multiplication by scalar*. Operation (1.1) is well-known operation in algebra, see e. g. [1, Ch. I]. Operation (1.2) is not so familiar; it is used in comparatively new areas like set-valued, convex and interval analysis. The symbol "$*$" in (1.2) will not be omitted throughout the paper in order to avoid confusion with the multiplication by scalar in a linear system. The latter will be further called a *linear multiplication by scalar* and will be denoted by "$\cdot$"; the dot "$\cdot$" may be omitted as in the expression "αb" in (1.2).

Addition. We recall some properties of (1.1). For A, B, $C\in\mathcal{K}$ we have

$$\begin{aligned} (A+B)+C &= A+(B+C), && (1.3)\\ A+B &= B+A, && (1.4)\end{aligned}$$

hence $(\mathcal{K},+)$ is a commutative (abelian) semigroup. There exists a neutral element in $\mathcal{K}$ — the origin 0 of $\mathbb{E}^n$ — such that for all $A\in\mathcal{K}$

$$A+0=A, \qquad (1.5)$$

hence $(\mathcal{K},+)$ is an abelian monoid (cf. [9, Ch. 2], which will be also denoted $(\mathcal{K},0,+)$ (to avoid misunderstandings, we shall usually denote the algebraic systems together with their operations).

It has been proved (see, e. g., [16, Lemma 2], or [20, p. 41] that the monoid $(\mathcal{K},+)$ is cancellative, that is, for $A,B,X\in\mathcal{K}$ the cancellation law holds:

$$A+X=B+X \Longrightarrow A=B. \qquad (1.6)$$

The extension method. We recall that an abelian group is an ordered quadruple $(\mathcal{G},+,0,-)$ satisfying relations (1.3)–(1.5) and (1.7). An abelian monoid $(S,+,0)$ turns into a (abelian) group if there exists an operation opp : $S\longrightarrow S$ such that

$$\mathrm{opp}(A)+A=0 \ \text{ for all } \ A\in S; \qquad (1.7)$$

instead of opp(A) we shall also write $-A$ or (whenever needed to avoid confusion) $-_S A$. Abelian cancellative monoids and abelian cancellative groups play important role in this study, for brevity we shall write "a. c." instead of "abelian cancellative".

In the a. c. monoid $(\mathcal{K},+,0)$ there is no opposite, hence the latter is not a group. However, there is a standard algebraic construction, further referred to as "the extension method", which allows us to embed isomorphically every a. c. monoid $(\mathcal{Q},+,0)$ into an a. c. group $(\mathcal{G},+,0,-)$ (see, e. g., [1]–[4], [7], [8]). Briefly, the extension method consists in the following: define $\mathcal{G}=(\mathcal{Q}\times\mathcal{Q})/\rho$ as the set of pairs

(A,B), A, $B \in \mathcal{Q}$, factorized by the equivalence relation $\rho : (A,B)\rho(C,D) \Longleftrightarrow A+D = B+C$. Addition in $\mathcal{G}$ is defined by means of: $(A,B)+(C,D) = (A+C, B+D)$. We shall denote the equivalence class in $\mathcal{G}$, represented by the pair (A,B), again by (A,B), thus $(A,B) = (A+X, B+X)$. The null element of $\mathcal{G}$ is the class (Z,Z); due to the existence of null element in $\mathcal{Q}$, we have $(Z,Z) = (0,0)$. The opposite element to $(A,B) \in \mathcal{G}$ is $-(A,B) = (B,A)$; indeed $(A,B) + (-(A,B)) = (A,B) + (B,A) = (A+B, B+A) = (0,0)$. Instead of $(A,B) + (-(C,D))$ we write $(A,B) - (C,D)$; we have $(A,B) - (C,D) = (A,B) + (D,C) = (A+D, B+C)$.

To embed isomorphically $\mathcal{Q}$ into $\mathcal{G}$, we identify $A \in \mathcal{Q}$ with the equivalence class $(A,0) = (A+X, X)$, $X \in \mathcal{Q}$. Thus all "proper elements" of $\mathcal{G}$ are pairs (U,V), U, $V \in \mathcal{Q}$, such that $V+Y = U$ for some $Y \in \mathcal{Q}$, i. e. $(U,V) = (V+Y, V) = (Y,0)$.

The group $(\mathcal{G}, +, 0, -)$ obtained by the extension method is minimal in the sense that if $(\mathcal{G}', +, 0, -)$ is any group in which $(\mathcal{Q}, +, 0)$ is embedded, then $(\mathcal{G}, +, 0, -)$ is isomorphic to a subgroup of $(\mathcal{G}', +, 0, -)$ containing $(\mathcal{Q}, +, 0)$. The group $(\mathcal{G}, +, 0, -)$ is unique up to isomorphism; we shall call it the *extension group induced by* $(\mathcal{Q}, +, 0)$.

Multiplication by scalar. Recall now some properties of (1.2). For $A, B \in \mathcal{K}$, $\gamma, \delta \in \mathbb{R}$ we have

$$\gamma * (A+B) = \gamma * A + \gamma * B, \tag{1.8}$$

$$\gamma * (\delta * A) = (\gamma\delta) * A, \tag{1.9}$$

$$1 * A = A, \tag{1.10}$$

where $\gamma\delta$ denotes the (linear) product of $\gamma, \delta \in \mathbb{R}$. The set of convex bodies together with operations (1.1), (1.2) will be denoted $(\mathcal{K}, +, \mathbb{R}, *)$ or $(\mathcal{K}, \mathbb{E}, +, \mathbb{R}, *)$.

Property (1.8) is known as "first distributive law". The so-called "second distributive law" is characteristic for any linear (vector) system, e. g. in the linear system $(\mathbb{E}^n, +, \mathbb{R}, \cdot)$ we have for every $C \in \mathbb{E}^n$

$$(\alpha+\beta)\cdot C = \alpha \cdot C + \beta \cdot C, \quad \alpha,\ \beta \in \mathbb{R}. \tag{1.11}$$

We recall that a system $(\mathcal{G}, +, \mathbb{R}, \cdot)$ is linear if: i) $(\mathcal{G}, +, 0, -)$ is an abelian group; ii) for all $a, b, c \in \mathcal{G}$, $\alpha, \beta, \gamma \in \mathbb{R}$

$$\left\{\begin{array}{l} \gamma \cdot (a+b) = \gamma \cdot a + \gamma \cdot b; \\ \gamma \cdot (\delta \cdot a) = (\gamma\delta) \cdot a; \\ 1 \cdot a = a; \\ (\alpha+\beta) \cdot c = \alpha \cdot c + \beta \cdot c. \end{array}\right. \tag{1.12}$$

The last relation in (1.12) is the second distributive law. Recall that in a linear system we have $0 \cdot a = 0$ and $(-1) \cdot a = -a$, hence we may omit the symbols "0" and "$-$" in the notation of a linear system.

The second distributive law (1.11) is not valid in $(\mathcal{K}, +, \mathbb{R}, *)$, apart of certain special cases. For example, for $C \in \mathcal{K}$ and equally signed scalars α, β we have

$$(\alpha+\beta) * C = \alpha * C + \beta * C, \ \ \alpha\beta \geq 0. \tag{1.13}$$

Convex bodies are a. c. monoids with scalar operator satisfying (1.8)–(1.10), (1.13), see e. g. [16, 17]. It has been shown that such algebraic structure, called $\mathbb{R}$-semigroup with cancellation law, is characteristic for convex bodies [18].

Denote by $(\mathcal{L}, +, 0, -)$ the group induced by the semigroup of convex bodies $(\mathcal{K}, +, 0)$, $\mathcal{L} = (\mathcal{K} \times \mathcal{K})/\rho$. The following question arises:

Question 1. Can we embed isomorphically $(\mathcal{K}, +, \mathbb{R}, *)$ in a linear system $(\mathcal{L}, +, \mathbb{R}, \cdot)$ with $\mathcal{L} = (\mathcal{K} \times \mathcal{K})/\rho$? In other words, can we isomorphically extend (1.2) in $\mathcal{L}$, so that $\mathcal{L}$ (which is a group under addition) becomes a linear system, that is (1.12) are valid in $\mathcal{L}$?

H. Rådström shows that if we define a multiplication by scalar "$\cdot$" in $\mathcal{L}$ in terms of the multiplication by scalar (1.2) in $\mathcal{K}$ using the relation

$$\gamma \cdot (A, B) = \begin{cases} (\gamma * A, \gamma * B), & \gamma \geq 0, \\ (|\gamma| * B, |\gamma| * A), & \gamma < 0, \end{cases} \tag{1.14}$$

then $(\mathcal{L}, +, \mathbb{R}, \cdot)$ is a linear system, that is relations (1.12) hold true in $\mathcal{L}$ (see [16, Theorem 1]).

Formula (1.14) does not induce an isomorphic embedding of $(\mathcal{K}, +, \mathbb{R}, *)$ into the linear system $(\mathcal{L}, +, \mathbb{R}, \cdot)$. To see this, recall that under an isomorphic embedding the element $U \in \mathcal{K}$ is identified with $(U, 0) \in \mathcal{L}$, hence the element $U = \gamma * A \in \mathcal{K}$ is identified with $(\gamma * A, 0) \in \mathcal{L}$. Therefore the equality

$$\gamma \cdot (A, 0) = (\gamma * A, 0) \tag{1.15}$$

should hold true for all $A \in \mathcal{K}$, $\gamma \in \mathbb{R}$. However, (1.15) does not hold true for $\gamma < 0$, $A \in \mathcal{K} \setminus \mathbb{E}$. Indeed, from (1.14)

$$\gamma \cdot (A, 0) = ((-\gamma) * 0, (-\gamma) * A) = (0, -\gamma * A) \neq (\gamma * A, 0),$$

where the last inequality follows from $A + (-1) * A \neq 0$ for $A \in \mathcal{K} \setminus \mathbb{E}$.

An isomorphic extension of the multiplication by scalar (1.2) in $\mathcal{L}$ is given by the expression

$$\gamma * (A, B) = (\gamma * A, \gamma * B), \;\; A, \; B \in \mathcal{K}, \;\; \gamma \in \mathbb{R}. \tag{1.16}$$

For nonnegative scalars, (1.14) and (1.16) coincide, and an embedding theorem for convex cones holds true (see [16, Theorem 2]). Note that: i) the system $(\mathcal{K}, +, \mathbb{R}, *)$ is not linear, and ii) the induced via (1.16) system $(\mathcal{L}, +, \mathbb{R}, *)$ is not linear as well. We shall investigate in more detail the algebraic properties of these two systems of convex bodies. In particular, we shall point our attention towards extending relation (1.13) to include the case $\alpha\beta < 0$ and shall consider the following question:

Question 2. Can we embed isomorphically $(\mathcal{K}, +, \mathbb{R}, *)$ into $(\mathcal{L}, +, \mathbb{R}, *)$, where the operation "$*$" in $(\mathcal{L}, +, \mathbb{R}, *)$ is defined by (1.16), and what are the properties of the system $(\mathcal{L}, +, \mathbb{R}, *)$?

To answer this question we shall formulate some new algebraic properties of the original system $(\mathcal{K}, +, \mathbb{R}, *)$. Since we know that (1.8)–(1.10) hold true, what remains to be studied is distributivity. We therefore concentrate our attention to distributive relations, both in $\mathcal{K}$ and $\mathcal{L}$. We first prove a modification of (1.11) in $\mathcal{K}$, called quasidistributive law, which completes (1.13). We then find out the distributivity relation in $\mathcal{L}$ induced by (1.16) and call it "q-distributive" law. With the establishment of the distributivity relations in $\mathcal{K}$ and $\mathcal{L}$ we are able to give abstract definitions of $\mathcal{K}$ and $\mathcal{L}$ as algebraic systems, arriving thus to the concept of "quasilinear" and "q-linear" systems. We study the isomorphic embedding of the quasilinear system of convex bodies into the q-linear system of factorized pairs of convex bodies. We shall show that in a q-linear system relation (1.14) defines a linear multiplication by scalar, hence every q-linear system involves a linear one. Some of our results related to intervals, i. e. for convex bodies in $\mathbb{E}^1$, are published in [11, 12].

2. MINKOWSKI SUBTRACTION

A set A of the form $A = x + B$, for $x \in \mathbb{E}$, $B \in \mathcal{K}$, is called a *translate* of B (by the vector x). Clearly, if A is a translate of B by x, then B is a translate of A by $-x$, $B = A - x$.

Let $A, B \in \mathcal{K}$. The expression

$$A \overset{*}{-} B = \bigcap_{b \in B} (A - b) \tag{2.17}$$

is introduced for convex bodies and studied by H. Hadwiger (see, e. g., [5, 6]) under the name Minkowski difference. We consider (2.17) as a partial operation defined whenever the right-hand side is not empty. The following equivalent presentation of (2.17) holds:

$$A \overset{*}{-} B \;=\; \{x \in \mathbb{E} \mid x + B \subset A\}. \tag{2.18}$$

Expression (2.18) says that $A \overset{*}{-} B$ is the set of all vectors x such that the translate of B by x belongs to A. If there exists at least one $t \in \mathbb{E}$ such that $t + B \subset A$, then $A \overset{*}{-} B$ is well defined and $t \in A \overset{*}{-} B$; in this case we shall write $B \leq_M A$. As usual, we shall write $B =_M A$ if both $B \leq_M A$ and $A \leq_M B$ hold, that is, there exist $t, s \in \mathbb{E}$ such that $t + B \subset A$ and $s + A \subset B$. In other words, $B =_M A$ iff there exists $p \in \mathbb{E}$ such that $A + p = B$ (then $A = B - p$), that is A and B are translates of each other. In particular, $B \subset A$ implies $B \leq_M A$. From (2.18) we have for A, $B \in \mathcal{K}$ [6]

$$(A \overset{*}{-} B) + B \subset A, \tag{2.19}$$

$$(A + B) \overset{*}{-} B = A. \tag{2.20}$$

For $A, B \in \mathcal{K}$ we say that B is a *summand of* A if there exists $X \in \mathcal{K}$ such that $A = B + X$ (then X is a summand of A, too). Thus, we see from (2.19) that if B is a summand of A, then $A \overset{*}{-} B$ is a summand of A (see [20, Lemma 3.1.8]):

$$(A \overset{*}{-} B) + B = A. \tag{2.21}$$

In other words, if for $A, B \in \mathcal{K}$ some of the equations $A + X = B$, $B + Y = A$ is solvable, then the corresponding solution is $X = B \overset{*}{-} A$, resp. $Y = A \overset{*}{-} B$. The following equality has been established in [6]:

$$\lambda * (A \overset{*}{-} B) = \lambda * A \overset{*}{-} \lambda * B. \tag{2.22}$$

According to (1.13) for $\alpha\beta \geq 0$ the expression $(\alpha + \beta) * C$ can be written as a sum of the two terms of $\alpha * C$ and $\beta * C$. Can we express $(\alpha + \beta) * C$ in a similar way in the case $\alpha\beta < 0$? The answer is positive. H. Hadwiger [6] proves the following equality:

$$(\lambda - \mu) * A = \lambda * A \overset{*}{-} \mu * A, \quad \lambda > \mu > 0. \tag{2.23}$$

Relation (2.23) can be rewritten in the following form, cf. also [14:

$$(\alpha + \beta) * C = \alpha * C \overset{*}{-} (-\beta) * C, \quad \alpha > 0, \; -\alpha < \beta < 0. \tag{2.24}$$

Hence, for $\alpha\beta < 0$, (2.24) can be written more symmetrically as

$$(\alpha + \beta) * C = \begin{cases} \alpha * C \overset{*}{-} (-\beta) * C, & \text{if } |\alpha| \geq |\beta|, \\ \beta * C \overset{*}{-} (-\alpha) * C, & \text{if } |\alpha| < |\beta|. \end{cases}$$

3. SUMMABILITY

To formulate and prove a generalization of (1.13) in $\mathcal{K}$, which is valid for all α, $\beta \in \mathbb{R}$, we first concentrate on some further properties of the convex bodies, related to Minkowski subtraction. For our purposes we shall make use of the cancellation law (1.6): $A + X = B + X \Longrightarrow A = B$ for $A, B, X \in \mathcal{K}$.

For given $A, B \in \mathcal{K}$, if there exists an $X \in \mathcal{K}$ such that B is a summand of A (i.e. $B + X = A$), then, due to (1.5) and (1.6), we have in $\mathcal{L}$ the presentation $(A, B) = (B + X, B) = (X, 0)$.

Proposition 1. *For $A, B \in \mathcal{K}$, if B is a summand of A, then there exists a unique $X \in \mathcal{K}$ such that $A = B + X$.*

Proof. By assumption, there is some $X \in \mathcal{K}$ such that $A = B + X$. Assume that $X' \in \mathcal{K}$, with $X' \neq X$, is such that $A = B + X'$. Then we have $B + X = B + X'$, which by the cancellation law (1.6) implies $X = X'$, a contradiction. □

Proposition 2. *Let $A, B \in \mathcal{K}$. The equality $A + B = 0$ implies $A, B \in \mathbb{E}$ and $B = -A$.*

In what follows we shall symbolically denote the relation "B is a summand of A" by $B \leq_\Sigma A$, or $A \geq_\Sigma B$; "$\leq_\Sigma$" is a partial order in $\mathcal{K}$. The assertion "B is not a summand of A" will be denoted by $B \not\leq_\Sigma A$. Obviously, $B \leq_\Sigma A$ implies $B \leq_M A$; however, the inverse is not true.

Generally speaking, for every A, $B \in \mathcal{K}$ there exist four possibilities: 1) $B \leq_\Sigma A$ and $A \not\leq_\Sigma B$, denoted $B <_\Sigma A$; 2) $A \leq_\Sigma B$ and $B \not\leq_\Sigma A$, denoted $A <_\Sigma B$; 3) $A \leq_\Sigma B$ and $B \leq_\Sigma A$, denoted $A =_\Sigma B$; 4) $A \not\leq_\Sigma B$ and $B \not\leq_\Sigma A$. Note that $A =_\Sigma B$ is equivalent to $A =_M B$, that is A and B are translates of each other.

In cases 1)–3) we say that the pair $(A, B) \in \mathcal{L}$ is *Σ-comparable.* We shall further denote the set of all Σ-comparable pairs by $\mathcal{L}_\Sigma$. Clearly, if $(A, B) \in \mathcal{L}_\Sigma$, then at least one of the expressions $A \overset{*}{-} B$, $B \overset{*}{-} A$ is well defined.

In case 3) there exists a unique $X \in \mathcal{K}$ such that $A = B + X$ and a unique $Y \in \mathcal{K}$ such that $B = A + Y$. Summing up both equations, we obtain $A + B = (B+X)+(A+Y) = (A+B)+X+Y$, and by (1.6), $X+Y = 0$. By Proposition 2, the solutions X, Y are opposite to each other; they belong to the set $\mathbb{E}$ of degenerate convex bodies (one-point sets). Thus, in case 3) A is a translate of B by the vector P, and, conversely, B is a translate of A by $-P$, that is $B = A + (-P)$, where $-P = -_{\mathbb{E}}P$ is the opposite of P in $\mathbb{E}$ (the point sets P and $-P$ are symmetric with respect to the origin 0 of $\mathbb{E}$).

We summarize the above arguments in the next proposition.

Proposition 3 (T-property). *Let $A, B \in \mathcal{K}(\mathbb{E})$, $(A, B) \in \mathcal{L}_\Sigma$. For the equations*

$$B + X = A, \tag{3.25}$$

$$A + Y = B \tag{3.26}$$

exactly one of the following three possibilities holds true:

1) Case $B <_\Sigma A$: there exists a unique nondegenerate convex body $X \in \mathcal{K} \setminus \mathbb{E}$ satisfying (3.25); equation (3.26) is not solvable.

2) Case $A <_\Sigma B$: there exists a unique nondegenerate convex body $Y \in \mathcal{K} \setminus \mathbb{E}$ satisfying (3.26); equation (3.25) is not solvable.

3) Case $A =_\Sigma B$: both (3.25) and (3.26) are solvable for X, resp. Y, and $Y = -X \in \mathbb{E}$.

From the cancellation law it follows that for arbitrary $A, B \in \mathcal{K}$ each of the equations (3.25), (3.26) may have at most one solution.

Proposition 4. *Let for $A, B, C \in \mathcal{K}$, $A + B = C$ and $0 \in A$. Then $B \subset C$.*

Proof. Equation $A + B = C$, that is $\bigcup_{a \in A} a + B = C$, means $a + B \subset C$ for all $a \in A$. Hence for $a = 0$, $B = 0 + B \subset C$. □

In the next section we study an operator in $\mathcal{K}$ called "negation", which plays an important role for the algebraic description of the properties of the set of convex bodies.

4. NEGATION

Substituting $\alpha = -1$ in (1.2) we obtain the operator

$$(-1) * A = \{-a \mid a \in A\}, \ \ A \in \mathcal{K}, \tag{4.27}$$

called *negation*, which will be denoted by $\neg A = (-1) * A$ or $\mathrm{neg}(A)$. Obviously, $\neg(\gamma * A) = (-1) * (\gamma * A) = (-\gamma) * A$.

The following properties of negation are easily verified:

$$\neg(\neg A) = A, \ \ A \in \mathcal{K}, \tag{4.28}$$

$$\neg(A + B) = (\neg A) + (\neg B), \ \ A, \ B \in \mathcal{K}, \tag{4.29}$$

$$\neg P + P = 0 \Longleftrightarrow P \in \mathbb{E} \Longleftrightarrow \neg P = -P, \tag{4.30}$$

$$\neg A = 0 \Longleftrightarrow A = 0, \ \ A \in \mathcal{K}. \tag{4.31}$$

Properties (4.28)–(4.29) mean that negation is a dual automorphism (involution). Property (4.30) means that a convex body $P \in \mathcal{K}$ satisfies $\neg P + P = 0$ if and only if P is a degenerate, in which case the negation $(\neg)$ coincides with the opposite operator $(-)$ in the set $\mathbb{E}$ of degenerate (one-point) elements of $\mathcal{K}$, i. e. $-P + P = 0$. Of course, $\neg A + A = 0$ does not hold in $\mathcal{K} \setminus \mathbb{E}$, since nondegenerate convex bodies have no opposite elements. We see that negation isomorphically extends the opposite from $\mathbb{E}$ to $\mathcal{K}$.

For brevity, we shall denote for $A, \ B \in \mathcal{K}$

$$A \neg B \equiv A + (\neg B) = A + (-1) * B = \{a - b \mid a \in A, \ b \in B\}; \tag{4.32}$$

the operation $A \neg B$ is called an (*outer*) *subtraction*.

Remarks. Instead of the symbol "$\neg$" we may use "$-$" as it is well adopted in the literature on interval and set-valued analysis (see, e. g., [10, 19]); however it should be kept in mind that there is no opposite operator in $\mathcal{K}$, and that $A \neg A \neq 0$ for $A \in \mathcal{K} \setminus \mathbb{E}$. Since the notation "$-$" is usually associated with the equality $A - A = 0$, to avoid confusion we write "$\neg$" instead of "$-$". Using "$\neg$", we also avoid confusion with the opposite in $\mathcal{L}$. In mathematical morphology the outer subtraction (4.32) is called *dilatation*, whereas the Minkowski subtraction is called *erosion* [15].

Definition. $A \in \mathcal{K}$ is called *symmetric* (*with respect to the origin*) if $x \in \mathbb{E}$, $x \in A$, implies $-x \in A$.

Obviously, $A \in \mathcal{K}$ is symmetric if and only if $A = \neg A$. For $A \in \mathcal{K}$ the set $A \neg A$ is called the difference body of A (see [20, p. 127]). The set of all symmetric convex bodies is denoted by $\mathcal{K}_S$, that is $\mathcal{K}_S = \{A \in \mathcal{K} \mid A = \neg A\}$.

Proposition 5. *For $A \in \mathcal{K}$ we have $A \neg A \in \mathcal{K}_S$.*

Proof. Let $x \in \mathbb{E}$ be such that $x \in A \neg A$. Then, from (4.32) $-x \in \neg(A \neg A) = (\neg A)\neg(\neg A) = (\neg A) + A = A \neg A$, using properties (4.28), (4.29). □

Proposition 6. *The following two conditions for symmetricity of $A \in \mathcal{K}$ are equivalent:*

i) $A = \neg A$;

ii) *there exists $Z \in \mathcal{K}$ such that $A = Z \neg Z$.*

Proof. i) Let $A = \neg A$. Assume $t \in \mathbb{E}$ and set $Z = A/2 + t$, where $A/2 = (1/2) * A$. Using $A = \neg A$, we obtain $\neg Z = \neg A/2 - t = A/2 - t$. Hence $Z \neg Z = Z + (\neg Z) = (A/2 + t) + (A/2 - t) = A$.

ii) Assume that $A = Z \neg Z$ for some $Z \in \mathcal{K}$. Then we have $\neg A = \neg(Z \neg Z) = \neg Z + Z = Z \neg Z = A$. □

Definition. $A \in \mathcal{K}$ is called *t-symmetric*, with *center* $t \in \mathbb{E}$, if $(A - t) \in \mathcal{K}_S$.

In other words, a t-symmetric element is a translate by t of a symmetric element.

Proposition 7. *Every t-symmetric convex body A is a translate of its negation* $\neg A$.

Proof. Let $A \in \mathcal{K}$ be t-symmetric. We have to show that there exists $P \in \mathbb{E}$ such that $\neg A + P = A$. Since A is t-symmetric, $A - t$ is symmetric, that is $(A - t) = \neg(A - t) = \neg A + t$. This implies $\neg A + 2t = A$, hence A is a translate by $2t$ of $\neg A$; we found $P = 2t$. □

Remark. Let $A \in \mathcal{K}$ be t-symmetric, i. e. $(A - t) \in \mathcal{K}_S$. By Proposition 6 there exists $Z \in \mathcal{K}$ such that $A - t = Z \neg Z$. To find an expression for Z, fix $s \in \mathbb{E}$ and set $Z = (A - t)/2 + s$; we obtain $Z = A/2 + s'$, $s' \in \mathbb{E}$. Thus $A - t = Z \neg Z = A/2 \neg A/2 = (A \neg A)/2$. We thus have $A - t = (A \neg A)/2$, that is for any t-symmetric element $A \in \mathcal{K}$ its symmetric translate by $-t$ is $(A \neg A)/2$.

5. INNER OPERATIONS

Inner addition in $\mathcal{K}$. *Inner sum $A +^- B$* is defined for (some) $A, B \in \mathcal{K}$ by

$$A +^- B = \begin{cases} \bigcap_{b \in B} (A + b), & \text{if } \neg B \leq_M A, \\ \bigcap_{a \in A} (B + a), & \text{if } \neg A <_M B. \end{cases}$$

Remark. The inner sum is defined whenever one of the conditions in the right-hand side is fulfilled. Note that $\neg B \leq_M A$ is equivalent to $B \leq_M \neg A$. Note also that if both $\neg B \leq_M A$ and $\neg A \leq_M B$ hold, that is $\neg B =_M A$, then $\neg B = A + t$ for some $t \in \mathbb{E}$. In this case it can be shown that both intersections in the right-hand side of the above definition produce the same result. Therefore we can replace the second condition above by $\neg A \leq_M B$ (or $A \leq_M \neg B$).

Inner difference $A -^- B$, for $A, B \in \mathcal{K}$, is defined by the equality $A -^- B \equiv A +^- (\neg B)$ [11, 12].

In the situation when $(A, \neg B) \in \mathcal{L}_\Sigma$, resp. $(A, B) \in \mathcal{L}_\Sigma$, the inner addition, resp. inner subtraction, admits simple presentation. Namely, we have

$$A +^- B \;=\; \begin{cases} Y|_{\neg B + Y = A}, & \text{if } \neg B \leq_\Sigma A, \\ X|_{\neg A + X = B}, & \text{if } \neg A \leq_\Sigma B, \end{cases} \tag{5.33}$$

$$A -^- B \;=\; \begin{cases} Y|_{B + Y = A}, & \text{if } B \leq_\Sigma A, \\ X|_{A \neg X = B}, & \text{if } A \leq_\Sigma B. \end{cases} \tag{5.34}$$

The inner operations (5.33) and (5.34) are related by $A +^- B = A -^- (\neg B)$. Note that $A \leq_M B$ does imply $\neg A \leq_M \neg B$, but does not necessarily imply $\neg A \leq_M B$. Due to this fact, for some A, $B \in \mathcal{K}$ it may happen that $A +^- B$ is defined, but $A -^- B$ is not, or vice versa.

A relation between the inner operations and the Minkowski difference is given by

$$\begin{aligned} A +^- B &= (A \stackrel{*}{-} (\neg B)) \cup (B \stackrel{*}{-} (\neg A)), \\ A -^- B &= (A \stackrel{*}{-} B) \cup \neg (B \stackrel{*}{-} A). \end{aligned}$$

Inner addition is commutative, $A +^- B = B +^- A$; other important property is $A -^- A = 0$.

Proposition 8. *Let* $(A, \neg B) \in \mathcal{L}_\Sigma$. *Then* $A +^- B \leq_\Sigma A + B$ *and* $A +^- B \subset A + B$.

Proof. From (5.33) we immediately see that $A +^- B$ is a summand of $A + B$. Indeed, if $\neg B \leq_\Sigma A$, we have $\neg B + (A +^- B) = A$, hence $B \neg B + (A +^- B) = A + B$. If $\neg A \leq_\Sigma B$, then $\neg A + (A +^- B) = B$, and hence $A \neg A + (A +^- B) = A + B$. Since in both cases the other summand contains 0 (indeed, $A \neg A \ni 0$ and $B \neg B \ni 0$), we have $A +^- B \subset A + B$, using Proposition 4 and $A +^- B \leq_\Sigma A + B$ as well. □

Remark. The proof can be generalized (cf. [15]) for the more general case when either $B \leq_M A$ or $A \leq_M B$ (in which case it may happen that $(A, \neg B) \notin \mathcal{L}_\Sigma$). Most of the results in the sequel can be extended to this more general case.

Proposition 9. *Let* $(A, \neg B) \in \mathcal{L}_\Sigma$. *Then*

$$(A, \neg B) = \begin{cases} (A +^- B, 0), & \text{if } \neg B \leq_\Sigma A, \\ (0, \neg (A +^- B)), & \text{if } \neg B >_\Sigma A. \end{cases}$$

Proof. From (5.33), if $\neg B \leq_\Sigma A$, then $\neg B + (A +^- B) = A$. Hence $(A, \neg B) = (\neg B + (A +^- B), \neg B) = (A +^- B, 0)$. The case $\neg B >_\Sigma A$ is treated analogously, using that $B >_{sig} \neg A$ implies $B = \neg A + (A +^- B)$, hence $\neg B = A + \neg (A +^- B)$.

Proposition 10. *Let $C \in \mathcal{K}$, $\alpha \in \mathbb{R}$, $\alpha \geq 0$. Then*

$$\begin{aligned} C &\geq_\Sigma \alpha * C, \quad \text{if} \quad 0 \leq \alpha \leq 1, \\ C &\leq_\Sigma \alpha * C, \quad \text{if} \quad \alpha \geq 1. \end{aligned}$$

Proof. Let $0 \leq \alpha \leq 1$. We have to verify that $\alpha * C$ is a summand of C, that is $\alpha * C + X = C$ for some $X \in \mathcal{K}$. Take $X = (1-\alpha) * C$. Substituting $\beta = 1 - \alpha \geq 0$ in

$$\alpha * C + \beta * C = (\alpha + \beta) * C, \ \ \alpha\beta \geq 0,$$

we obtain $\alpha * C + (1-\alpha) * C = (\alpha + 1 - \alpha) * C = C$, showing that $C \geq_\Sigma \alpha * C$, for $\alpha \in [0,1]$. Let $\alpha \geq 1$. We look for Y such that $\alpha * C = Y + C$. Taking $Y = (\alpha - 1) * C$, we see that $C \leq_\Sigma \alpha * C$ for $\alpha \geq 1$. □

The above proposition shows that for $\alpha \in (0,1)$ the solution of $C = \alpha * C + X$ is $X = (1-\alpha) * C$.

Proposition 11. *Let $\alpha, \beta \in \mathbb{R}$, $C \in \mathcal{K}$. If $\alpha\beta > 0$, then $(\alpha * C, \beta * C) \in \mathcal{L}_\Sigma$. If $\alpha\beta < 0$, then $(\neg\alpha * C,\ \beta * C) \in \mathcal{L}_\Sigma$.*

Proof. The case $\alpha\beta = 0$ is obvious. Let $\alpha\beta > 0$, say $\alpha \geq \beta > 0$. We shall show that the pair $(\alpha * C, \beta * C)$ is Σ-comparable, and $\beta * C \leq_\Sigma \alpha * C$. By Proposition 10 we have that C and $(\alpha/\beta) * C$, $\alpha/\beta > 1$ are Σ-comparable with $(\alpha/\beta) * C \geq_\Sigma C$, that is $C + X = (\alpha/\beta) * C$ is solvable. Then $\beta * C + Y = \alpha * C$ is solvable, i. e. $\alpha * C \geq_\Sigma \beta * C$. The other subcases of $\alpha\beta > 0$ are treated similarly. The case $\alpha\beta < 0$ is reduced to the previous case by setting $\alpha = -\gamma$. □

Proposition 12. *Let α, $\beta \in \mathbb{R}$, $\alpha\beta < 0$, $C \in \mathcal{K}$. Then*

$$(\alpha + \beta) * C = \alpha * C +^- \beta * C.$$

Proof. Without a loss of generality we may assume that $\alpha > 0$, $\beta < 0$. Denote $-\beta = \gamma > 0$. Using (5.33), we obtain

$$\begin{aligned} \alpha * C +^- \beta * C &= \begin{cases} Y|_{\gamma * C + Y = \alpha * C}, & \text{if } \gamma * C \leq_\Sigma \alpha * C, \\ X|_{\alpha * C + (\neg X) = \gamma * C}, & \text{if } \alpha * C \leq_\Sigma \gamma * C; \end{cases} \\ &= \begin{cases} (\alpha - \gamma) * C, & \gamma \leq \alpha, \\ \neg(\gamma - \alpha) * C, & \alpha \leq \gamma; \end{cases} \\ &= (\alpha - \gamma) * C = (\alpha + \beta) * C. \end{aligned}$$

In the proof we make use of Proposition 10. □

Denote the sign of the real number $\alpha \in \mathbb{R}$ by $\sigma(\alpha) \in \{+,-\}$, that is:

$$\sigma(\alpha) = \begin{cases} +, & \text{if } \alpha \geq 0, \\ -, & \text{if } \alpha < 0. \end{cases}$$

Assuming $+^{+} = +$, we can combine Proposition 12 and relation (1.13) in the following general *quasidistributive law*: for every $\alpha, \beta \in \mathbb{R}$, $C \in \mathcal{K}$,

$$(\alpha + \beta) * C = \alpha * C +^{\sigma(\alpha\beta)} \beta * C. \tag{6.35}$$

7. QUASILINEAR SYSTEMS OF CONVEX BODIES

As already mentioned, due to (1.3)–(1.6) the system $(\mathcal{K}, +, 0)$ is an (additive) a. c. monoid. This system is a *proper* semigroup (i. e. not a group itself), which means that there exists at least one pair (A, B) such that $A + X = B$ has no solution for $X \in \mathcal{K}$. The monoid $(\mathcal{K}, +, 0)$ has a unique idempotent element e (such that $e + e = e$), which is the null element ($e = 0$). For a semigroup $(\mathcal{Q}, +)$ with only one idempotent element it is known that the set $\mathcal{Q}_0$ of all invertable elements $u \in \mathcal{Q}$ (i. e., such that $u + v = 0$ for some v) is a group $(\mathcal{Q}_0, +, 0, -)$, which is the unique maximal subgroup of the semigroup (see, e. g., [1, Section 1.7]). Recall that a subgroup $(\mathcal{M}, +)$ of a semigroup $(\mathcal{Q}, +)$ is called maximal (with respect to "$\subset$") if there is no other subgroup $(\mathcal{M}', +)$ of $(\mathcal{Q}, +)$ such that $\mathcal{M}' \supset \mathcal{M}$ and $\mathcal{M} \neq \mathcal{M}'$. If no doubt occurs, we shall further say "the subgroup of the monoid" instead of "the unique maximal subgroup of the monoid".

Using the above terminology, we can say that the system of convex bodies $(\mathcal{K}, +, 0)$ involves the group $(\mathbb{E}, +, 0)$, which is the (maximal) subgroup of $\mathcal{K}$ comprising all invertible elements of $\mathcal{K}$.

Given a semigroup $(\mathcal{Q}, +)$, we shall call $\pi : \mathcal{Q} \longrightarrow \mathcal{Q}$ an *involution in* $\mathcal{Q}$ if it is a dual automorphism, that is:

i) $\pi(\pi(A)) = A$ for $A \in \mathcal{Q}$;

ii) $\pi(A + B) = \pi(A) + \pi(B)$ for $A, B \in \mathcal{Q}$.

A proper a. c. monoid $(\mathcal{Q}, +, 0)$ with (maximal) subgroup $(\mathcal{Q}_0, +, 0, -)$ will be further denoted $(\mathcal{Q}, \mathcal{Q}_0, +)$. The subgroup $(\mathcal{Q}_0, +, 0, -)$ contains the trivial group, and, in particular, it may happen that $\mathcal{Q}_0 = \{0\}$. In $(\mathcal{Q}, \mathcal{Q}_0, +)$ we define negation as follows:

Definition. Let $(\mathcal{Q}, \mathcal{Q}_0, +)$ be a proper a. c. monoid. An involution neg : $\mathcal{Q} \to \mathcal{Q}$ is called *negation* in $(\mathcal{Q}, \mathcal{Q}_0, +)$ if it extends the operator opposite from $\mathcal{Q}_0$ to $\mathcal{Q}$: $\mathrm{neg}(P) = -_{\mathcal{Q}_0} P$ for $P \in \mathcal{Q}_0$ (i. e. $\mathrm{neg}(P) + P = 0$, $P \in \mathcal{Q}_0$).

It is easily seen that $\mathrm{neg}(A) = 0 \iff A = 0$ for $A \in \mathcal{Q}$, which corresponds to (4.31).

We shall further require that negation is unique (sufficient conditions for uniqueness will be discussed elsewhere).

Definition [12]. A proper a. c. monoid $(\mathcal{Q}, \mathcal{Q}_0, +)$ with unique operator negation "neg $= \neg$" is called a *quasimodule* and is denoted by $(\mathcal{Q}, \mathcal{Q}_0, +, \neg)$.

Remark. Note that a quasimodule $(\mathcal{Q},\mathcal{Q}_0,+,\neg)$ is not a group, but it possesses the same number of basic operations as a group does: one binary "+", one unary "$\neg$", and one nullary operation "0", and the algebraic properties of $(\mathcal{Q},\mathcal{Q}_0,+,\neg)$ are close to those of a group.

The relation "$\leq_\Sigma$" is defined in a general semigroup in the same manner as it is done in the semigroup $(\mathcal{K},+)$, see Sections 2, 3. Inner addition "$+^-$" and inner subtraction "$-^-$" in a quasimodule are partial operations defined by (5.33)–(5.34), hence the quasidistributivity law (6.35) mentioned in the next definition makes sense.

Definition. Multiplication by scalar $*:\mathbb{R}\times\mathcal{Q}\longrightarrow\mathcal{Q}$, in a quasimodule $(\mathcal{Q},\mathcal{Q}_0,+,\neg)$ is a scalar operator over $\mathbb{R}$ satisfying relations (1.8)–(1.10), (6.35), and such that $(-1)*A=\neg A$ for all $A\in\mathcal{Q}$.

The last assumption $(-1)*A=\neg A$ means that negation is a special case of multiplication by scalar. Note also that the multiplication by integers $n*A = A+A+\cdots+A$ is consistent with the multiplication by (real) scalar (1.2), hence the symbol "$*$" makes sense in expressions like $2*A=A+A$; we also have $\neg(n*A)=(-1)*(n*A)=(-n)*A$.

Definition. A quasimodule endowed with multiplication by scalar is called *quasilinear system (over the field* $\mathbb{R}$*)*, or $\mathbb{R}$*-quasimodule*, and is denoted by $(\mathcal{Q},\mathcal{Q}_0,+,\mathbb{R},*)$.

We borrow the notion "quasilinear" from [13], where this notion is used to denote a similar algebraic system of convex bodies over $\mathbb{E}^1$, that is intervals.

If the subgroup of a quasimodule $\mathcal{Q}$ is $\mathcal{Q}_0=(\{0\},+)$, then negation and identity coincide. Due to $\neg A=A$, in a quasimodule with (maximal) subgroup 0 we have $A+B=A\neg B$ and $A+^-B=A-^-B$.

Example 1. The subgroup of the monoid of convex bodies is $(\mathbb{E},+,0)$, hence the corresponding $\mathbb{R}$-quasimodule is $(\mathcal{K},\mathbb{E},+,\mathbb{R},*)$ (also to be further referred to as *quasilinear system of convex bodies*).

Example 2. The system of symmetric elements $(\mathcal{K}_S,0,+)$ is a quasimodule with subgroup $(\{0\},+)$. The quasilinear system of symmetric elements is $(\mathcal{K}_S,0,+,\mathbb{R},*)$. Due to $\neg B=(-1)*B=B$ it is easy to check that $\alpha*B=|\alpha|*B$ for $B\in\mathcal{K}_S$. Using (1.2), this implies $\alpha*B=\{\alpha x\mid x\in B\}=\{|\alpha|x\mid x\in B\}$.

Example 3 [17]. Another instructive example of a quasilinear system with (maximal) subgroup 0 is the system $(\mathbb{R}^+,0,+,\mathbb{R},*)$, where $(\mathbb{R}^+,+)$ is the semigroup of nonnegative reals with subgroup $(\{0\},+)$. The system $(\mathbb{R}^+,0,+,\mathbb{R},*)$ can be considered as a subsystem of $(\mathcal{K}_S,0,+,\mathbb{R},*)$ whenever $\mathcal{K}_S$ is replaced by a subset of symmetric bodies of the form $\mathcal{K}'_S=r*B$, $r\in\mathbb{R}^+$, with $B\in\mathcal{K}_S$, $B\neq 0$, fixed; then $\mathcal{K}'_S\cong\mathbb{R}^+$. The multiplication by scalar $*:\ \mathbb{R}\times\mathbb{R}^+\longrightarrow\mathbb{R}^+$ satisfies $\alpha*A=|\alpha|*A$. In this system we have $A+^-B=A-^-B=|A-B|$, where $|A-B|$

is defined in $\mathbb{R}^+$ by $\{A-B, \text{ if } B \leq A;\ B-A, \text{ if } A < B\}$. By definition, $A-B$ for $B \leq A$ is the solution of $B+X=A$.

8. THE $\mathcal{Q}$-LINEAR SYSTEM

Here we shall further stay within the framework of abstract algebraic systems. This approach allows us to summarize several important special cases, such as the ones considered in Examples 1–3.

Recall that every quasimodule $(\mathcal{Q}, \mathcal{Q}_0, +, \neg)$ is an a. c. monoid, and, according to the extension method mentioned in the introduction, the quasimodule induces an extension group with supporting set $\mathcal{G} = (\mathcal{Q} \times \mathcal{Q})/\rho$. The extension group has opposite $\text{opp}(A,B) = (B,A)$, which will be denoted symbolically by "$-_{\mathcal{G}}$" or just "$-$". Negation ($\text{neg} = \neg$) in the quasimodule induces a corresponding operator in the extension group $(\mathcal{G}, +)$ by means of $\text{neg}(A,B) = (\text{neg}(A), \text{neg}(B))$, $A, B \in \mathcal{Q}$, which will be again called *negation* (in $\mathcal{G}$) and denoted symbolically by "$\neg$": $\neg(A,B) = (\neg A, \neg B)$. The set of invertable elements $\mathcal{Q}_0$ of the monoid is isomorphic to a subset of $\mathcal{G}$ of the form $\mathcal{G}_0 = \{(P,0) \mid P \in \mathcal{Q}_0\}$, which is subgroup of the extension group: $(\mathcal{G}_0, +, 0, -) \cong (\mathcal{Q}_0, +, 0, -)$. We shall incorporate the important elements "$\mathcal{G}_0$", "$-$", "$\neg$" (and, of course, "$+$") in the notation of the extension group induced by the quasimodule $(\mathcal{Q}, \mathcal{Q}_0, +, \neg)$, writing thus $(\mathcal{G}, \mathcal{G}_0, +, -, \neg)$.

Let us first discuss in some detail the automorphisms in the extension group $(\mathcal{G}, \mathcal{G}_0, +, -, \neg)$. Denote the identity in $\mathcal{G}$ by "id" and the operator, which is a composition "$\circ$" of negation and opposite, by dual: $\text{dual} = \text{neg} \circ \text{opp}$, that is $\text{dual}(a) = \text{neg}(\text{opp}(a))$ for $a \in \mathcal{G}$; the operator $\text{dual}(a)$ is called *dualization* (or *conjugation*). Since negation and opposite are involutions, dualization is also involution. Any two of the four involutions id, neg, opp and dual in $\mathcal{G}$ are composed to each other according to Table 1.

TABLE 1
Composition table for the involutions in $\mathcal{G}$

o	id	neg	opp	dual
id	id	neg	opp	dual
neg	neg	id	dual	opp
opp	opp	dual	id	neg
dual	dual	opp	neg	id

Let $(\mathcal{G}, \mathcal{G}_0, +, \neg)$ with $\mathcal{G} = (\mathcal{Q} \times \mathcal{Q})/\rho$ and $\mathcal{G}_0 \cong \mathcal{Q}_0$ be the extension group generated by the quasimodule $(\mathcal{Q}, \mathcal{Q}_0, +, \neg)$, and let π be any of the operators opposite or negation in $\mathcal{G}$. As already mentioned, π is an involution in the sense that:

C1) $\pi(\pi(a)) = a$ for $a \in \mathcal{G}$;

C2) $\pi(a+b) = \pi(a) + \pi(b)$ for $a, b \in \mathcal{Q}$;

C3) $\pi(a) = 0 \Longleftrightarrow a = 0$ for $a \in \mathcal{G}$.

It is important to note that the both involutions "opp" and "neg" extend the operator "opp" from $\mathcal{Q}_0$ into $\mathcal{G}$, that is:

C4) $\pi(p) = -_{\mathcal{Q}_0}(p)$, i. e. $\pi(p) + p = 0$, for $p \in \mathcal{G}_0 \cong \mathcal{Q}_0$.

Since both opposite and negation satisfy conditions C1)–C4), it is interesting to formulate characteristic conditions for the distinction of these two operators. One such distinctive property is that $\mathrm{opp}(p) + p = 0$ for every $p \in \mathcal{G}$, whereas $\mathrm{neg}(p) + p \neq 0$ for $p \in \mathcal{G} \setminus \mathcal{G}_0$. We shall next consider another distinctive property.

The class $\mathcal{G}_\Sigma$ of Σ-comparable elements of $\mathcal{G}$ is

$$\mathcal{G}_\Sigma = \{(U,0) \mid U \in \mathcal{Q}\} \bigcup \{(0,V) \mid V \in \mathcal{Q}\}. \tag{8.36}$$

The function *type* (or *direction*) of an element of $\mathcal{G}_\Sigma$ is defined by

$$\tau(A,B) \quad = \quad \begin{cases} +, & \text{if } B \leq_\Sigma A, \\ -, & \text{if } A <_\Sigma B. \end{cases} \tag{8.37}$$

The form of presentation $(U,0)$, resp. $(0,V)$, appearing in (8.36) is similar to the form used for real numbers; indeed, we may write $(U;+)$ for $(U,0)$ and $(V;-)$ for $(0,V)$ as we do with positive, resp. negative, numbers.

An element $(A,B) \in \mathcal{G}_\Sigma$ is *proper* if $(A,B) = (U,0)$ for some $U \in \mathcal{Q}$. According to the extension method the element $W \in \mathcal{Q}$ is identified with the proper element $(W,0) \in \mathcal{G}_\Sigma$. Improper Σ-comparable elements are of the form $(A,B) = (0,V)$, $V \in \mathcal{Q} \setminus \mathcal{Q}_0$. A special case of proper elements are the degenerate $(U,0)$ with $U \in \mathcal{Q}_0$. Using the notation (8.37), $a \in \mathcal{G}_\Sigma$ is proper if $\tau(a) = +$, and improper if $\tau(a) = -$. According to Proposition 9, using inner addition we can present any Σ-comparable element of $\mathcal{G}$ in the "($\pm$)-form $(U,0)$ or $(0,V)$, resp. $(A;\pm)$.

If an element $a \in \mathcal{G}_\Sigma$ is proper, then $\mathrm{neg}(a)$ is also proper, since $a = (A,0)$ implies $\mathrm{neg}(A,0) = (\mathrm{neg}(A),0)$ for $A,B \in \mathcal{Q}$. If an element b is improper, then $\mathrm{neg}(b) = \mathrm{neg}(0,B) = (0,\mathrm{neg}(B))$, showing that negation preserves the type, that is:

C5) $\tau(\mathrm{neg(a)}) = \tau(\mathrm{a})$ for $a \in \mathcal{G}_\Sigma$.

On the other side, for a nondegenerate a:

C6) $\tau(\mathrm{opp(a)}) = -\tau(\mathrm{a})$ for $a \in \mathcal{G}_\Sigma$,

where $-\tau$ is defined by $-- = +$, $-+ = -$.

The operators identity "id" and "dual" satisfy C1)–C3) and, instead of C4), the condition:

C7) $\pi(p) = p$ for $p \in \mathcal{G}_0$.

However, unlike identity, dualization changes the type of a Σ-comparable element. We summarize the above observations as follows:

Proposition 13. *The quasimodule $(\mathcal{Q}, \mathcal{Q}_0, +, \neg)$ generates (by means of the extension method) a system $(\mathcal{G}, \mathcal{G}_0, +, -, \neg)$ such that:*

1) $\mathcal{G} = (\mathcal{Q} \times \mathcal{Q})/\rho$; $(\mathcal{G}_0, +, 0, -) \cong (\mathcal{Q}_0, +, 0, -)$; *the opposite in* $\mathcal{G}$ *is:* $\mathrm{opp}(A, B) = -(A, B) = (B, A)$, $A, B \in \mathcal{Q}$.

2) *Negation in* $\mathcal{G}$ *is given by* $\mathrm{neg}(A, B) = (\mathrm{neg}(A), \mathrm{neg}(B))$, $A, B \in \mathcal{Q}$; *dualization, which is a composition of negation and opposite, is:* $\mathrm{dual}(A, B) = \mathrm{neg}(B, A) = (\mathrm{neg}(B), \mathrm{neg}(A))$, $A, B \in \mathcal{Q}$. *Opposite and negation coincide on* $\mathcal{G}_0$ *and dual coincides on* $\mathcal{G}_0$ *with identity. Opposite and dualization change the type of the* Σ*-comparable elements, whereas negation does not influence the type. The four automorphisms on* $\mathcal{G}$*: identity, opposite, negation and dualization, obey composition rules according to Table* 1.

The following symbolic notation will be used: for $a \in \mathcal{G}$ we write $\mathrm{dual}(a) = a_-$, $a = a_+$; then a_σ is either a or $\mathrm{dual}(a)$ according to the value of $\sigma \in \{+, -\}$. In such notation we have for $U, V \in \mathcal{Q}$: $(U, V)_- = (\neg V, \neg U)$.

The multiplication by scalar "$*$" in the quasilinear system $(\mathcal{Q}, \mathcal{Q}_0, +, \mathbb{R}, *)$ induces a corresponding multiplication in the extension group $(\mathcal{G}, \mathcal{G}_0, +, -, \neg)$ by means of the relation

$$\gamma * (A, B) = (\gamma * A, \gamma * B), \; A, B \in \mathcal{Q}. \tag{8.38}$$

Applying the extension method to the quasilinear system $(\mathcal{Q}, \mathcal{Q}_0, +, \mathbb{R}, *)$ (that is, extending the multiplication by scalar), we obtain a new system with basic properties given in the next proposition; below we assume $\alpha, \beta, \gamma \in \mathbb{R}$, $a, b, c \in \mathcal{G}$ [12].

Proposition 14. *Let* $(\mathcal{Q}, \mathcal{Q}_0, +, \mathbb{R}, *)$ *be a quasilinear system,* $(\mathcal{G}, \mathcal{G}_0, +, -, \neg)$ *be the extension group according to Proposition* 13 , *and multiplication by scalar "$*$" be defined in* $\mathcal{G}$ *by* (8.38). *Then:*

i) $\neg a = (-1) * a$;

ii) $\alpha * (\beta * c) = (\alpha\beta) * c$;

iii) $\gamma * (a + b) = \gamma * a + \gamma * b$;

iv) $1 * a = a$;

v) $(\alpha + \beta) * c_{\sigma(\alpha+\beta)} = \alpha * c_{\sigma(\alpha)} + \beta * c_{\sigma(\beta)}$;

vi) $(-1) * a + a = 0$ *for* $a \in \mathcal{G}_0$.

Proof. Relations i)–iv) and vi) are obvious. To prove v), note that it is equivalent to v′) $(\alpha + \beta) * c = (\alpha * c + \beta * c_{\sigma(\alpha)\sigma(\beta)})_{\sigma(\alpha)\sigma(\alpha+\beta)}$; we shall prove v) in this latter form. Substitute $c = (U, V) \in \mathcal{G}$ with $U, V \in \mathcal{Q}$. The right-hand side of v′) is

$$r = (\alpha * (U, V) + \beta * (U, V)_{\sigma(\alpha)\sigma(\beta)})_{\sigma(\alpha)\sigma(\alpha+\beta)}.$$

If $\sigma(\alpha)\sigma(\beta) = +$, using that $\sigma(\alpha)\sigma(\alpha + \beta) = +$ as well, we see that r is identical to the left-hand side

$$l = (\alpha + \beta) * (U, V) = ((\alpha + \beta) * U, \; (\alpha + \beta) * V).$$

Consider now the case $\sigma(\alpha)\sigma(\beta) = -$. The right-hand side becomes

$$\begin{aligned} r &= (\alpha * (U,V) + \beta * (U,V)_-)_{\sigma(\alpha)\sigma(\alpha+\beta)} \\ &= (\alpha * (U,V) + \beta * (\neg V, \neg U))_{\sigma(\alpha)\sigma(\alpha+\beta)} \\ &= ((\alpha * U, \alpha * V) + ((-\beta) * V, (-\beta) * U)_{\sigma(\alpha)\sigma(\alpha+\beta)} \\ &= (\alpha * U + (-\beta) * V,\ \alpha * V + (-\beta) * U)_{\sigma(\alpha)\sigma(\alpha+\beta)}. \end{aligned}$$

We must now consider a number of subcases. Consider, e. g., the subcase $\sigma(\alpha) = +$, $\sigma(\beta) = -$, $\sigma(\alpha+\beta) = +$ (in this subcase we have $\alpha \geq -\beta > 0$). Adding the zero term $(-\beta) * (U+V, U+V) = (0,0)$ to the left-hand side and using the quasidistributive law (6.35), we obtain

$$\begin{aligned} l &= (\alpha+\beta) * (U,V) + (-\beta) * (U+V, U+V) \\ &= ((\alpha+\beta) * U, (\alpha+\beta) * V) + ((-\beta) * U + (-\beta) * V, (-\beta) * U + (-\beta) * V) \\ &= ((\alpha+\beta) * U + (-\beta) * U + (-\beta) * V,\ (\alpha+\beta) * V) + (-\beta) * V + (-\beta) * U) \\ &= (\alpha * U + (-\beta) * V,\ \alpha * V) + (-\beta) * U) = r. \end{aligned}$$

The rest of the cases are treated analogously. □

Relation v) (or v′)) will be called *q*-distributive law. The *q*-distributive law can be also written in the form $(\alpha+\beta)c = \alpha c_\lambda + \beta c_\mu$ with $\lambda = \sigma(\alpha)\sigma(\alpha+\beta)$, $\mu = \sigma(\beta)\sigma(\alpha+\beta)$.

Definition. The system obtained in Proposition 14 will be further denoted $(\mathcal{G}, \mathcal{G}_0, +, -, \mathbb{R}, *)$ and called *q-linear system.*

Proposition 14 is a generalization of Radström's embedding theorem [16] in two directions: a) no restrictions for the signs of the scalar multipliers in the second distributive law (that is in the quasi- and *q*-distributive laws) are required (leading to embedding of cones in Radström case), and b) our theorem is formulated for abstract algebraic systems, comprising the system of convex bodies as special case. Clearly, (8.38) isomorphically extends multiplication by scalar from $\mathcal{Q}$ into $\mathcal{G}$; briefly, Proposition 14 says that a quasilinear system can be isomorphically embedded into a *q*-linear system. Thus the proposition answers fully the questions posed in the introduction.

Proposition 15. *Let $(\mathcal{G}, \mathcal{G}_0, +, -, \mathbb{R}, *)$ be a q-linear system and the operation "$\cdot$": $\mathbb{R} \times \mathcal{G} \longrightarrow \mathcal{G}$ be defined by*

$$\alpha \cdot c = \alpha * c_{\sigma(\alpha)}, \quad \alpha \in \mathbb{R},\ c \in \mathcal{G}. \tag{8.39}$$

Then $(\mathcal{G}, +, \mathbb{R}, \cdot)$ is a linear system.

Proof. Let us check that "$\cdot$" satisfies the axioms for linear multiplication.

1. Let us prove that $\alpha \cdot (\beta \cdot d) = (\alpha\beta) \cdot d$. Substitute $c = d_{\sigma(\beta)}$ in the relation $\alpha * (\beta * c) = (\alpha\beta) * c$ to obtain $\alpha * (\beta * d_{\sigma(\beta)}) = (\alpha\beta) * d_{\sigma(\beta)}$. Using (8.39), we

have $\alpha * (\beta \cdot d) = (\alpha\beta) * d_{\sigma(\beta)}$. "Dualizing" by $\sigma(\alpha)$, we obtain $\alpha * (\beta \cdot d)_{\sigma(\alpha)} = (\alpha\beta) * d_{\sigma(\beta)\sigma(\alpha)} = (\alpha\beta) * d_{\sigma(\beta\alpha)}$, or $\alpha \cdot (\beta \cdot d) = (\alpha\beta) \cdot d$, for all $d \in \mathcal{G}$, $\alpha, \beta \in \mathbb{R}$.

2. To prove the relation $\gamma \cdot (a + b) = \gamma \cdot a + \gamma \cdot b$, substitute $a = c_{\sigma(\gamma)}$, $b = d_{\sigma(\gamma)}$ in $\gamma * (a + b) = \gamma * a + \gamma * b$. We obtain $\gamma * (c_{\sigma(\gamma)} + d_{\sigma(\gamma)}) = \gamma * c_{\sigma(\gamma)} + \gamma * d_{\sigma(\gamma)}$, or $\gamma * (c + d)_{\sigma(\gamma)} = \gamma * c_{\sigma(\gamma)} + \gamma * d_{\sigma(\gamma)}$. This implies that $\gamma \cdot (c + d) = \gamma \cdot c + \gamma \cdot d$ for all $c, d \in \mathcal{G}$, $\gamma \in \mathbb{R}$.

The relations $1 \cdot a = a$, $(\alpha + \beta) \cdot c = \alpha \cdot c + \beta \cdot c$ and $(-1) \cdot a + a = 0$ can be proved similarly. □

We proved that the system $(\mathcal{G}, +, \mathbb{R}, \cdot)$ is a linear system (and, hence, "$\cdot$" is a linear multiplication by scalar). The operation "$\cdot$" is involved in the q-linear system — therefore the latter can be written in the form $(\mathcal{G}, \mathcal{G}_0, +, \mathbb{R}, *, \cdot)$.

9. CONCLUSIONS

Algebraic properties of convex bodies with respect to Minkowski operations for addition and multiplication by real scalar are studied. To this end two new operations, called inner addition, resp. inner subtraction, are introduced, and a new analogue of the second distributive law, called quasidistributive law, is proved. With the latter the system of convex bodies becomes a quasilinear system. A quasilinear system of convex bodies can be isomorphically embedded into a q-linear system, having group properties with respect to addition. The quasidistributive law induces in the q-linear system a corresponding distributivity relation, called q-distributive law. A q-linear system has much algebraic structure and is rather close to a linear system and differs from the latter by:

i) existence of two new automorphic operators — "negation" and "dualization" — in addition to the familiar automorphism "opposite" (and, of course, identity);

ii) the distributivity relation (q-distributive law) resembles the usual linear distributive law with the difference that the operator dualization is involved.

From our study it becomes clear that quasilinear and q-linear systems summarize some of the most characteristic algebraic properties of convex bodies. However, the following methodological question remains open: Which are the algebraic properties of the abstract systems corresponding to the notion of "convexity"? In our abstract study we circumvent this question by stepping directly on the fundament of abelian cancellative monoids — algebraic systems comprising well-known properties of convex bodies. Another approach could be to take into account that the set of convex bodies is a power set of certain type over a vector (or Euclidean) space (or lattice). For the latter approach results from [21] may be used, where the concept of convexity has been considered in abstract algebraic systems, which are more general than semigroups — the so-called associative spaces. Another similar approach offers the mathematical morphology (see, e. g., [15]), where vector lattices are used as fundament.

ACKNOWLEDGEMENTS. The present article arose from a talk at a seminar commemorating Prof. Y. Tagamlitzki, organized by Prof. D. Skordev and Prof. V.

Chakalov. The author is grateful to the anonymous referee for his extraordinary careful reading of the paper and remarks. This work was partially supported by the Bulgarian National Science Fund under Grant No I-507/95.

REFERENCES

1. Clifford, A. H., G. B. Preston. The Algebraic Theory of Semigroups. I, AMS, 1964.
2. Cohn, P. M. Universal Algebra. Harper and Row, 1965.
3. Fuchs, L. Abelian Groups. Publ. House Hung. Acad. Sci., Budapest, 1958.
4. Fuchs, L. Infinite Abelian Groups. I and II. Academic Press, 1970 and 1973.
5. Hadwiger, H. Minkowskische Addition und Subtraktion beliebiger Punktmengen und die Theoreme von Rirhard Schmidt. *Math. Z.*, **53**, Heft 3, 1953, 210–218.
6. Hadwiger, H. Vorlesungen über Inhalt, Oberfläche und Isoperimetrie. Springer, Berlin, 1957.
7. Kaucher, E. Interval Analysis in the Extended Interval Space $I\mathbb{R}$. *Computing Suppl.*, **2**, 1980, 33–49.
8. Kurosh, A. G. Lectures on General Algebra. Chelsea, 1963.
9. MacLane S., G. Birkhoff. Algebra. II ed., Macmillan, N. Y., 1979.
10. Markov, S. On Two Interval-arithmetic Structures and Their Properties. *Ann. Univ. Sofia, Fac. Math. Inf.*, **89**, 1995, 129–164 (in Bulgarian).
11. Markov, S. Isomorphic Embedding of Abstract Interval Systems. *Reliable Computing*, **3**, 3, 1997, 199–207.
12. Markov, S. On the Algebra of Intervals and Convex Bodies. *J. UCS*, **4**, 1, 1998, 34–47.
13. Mayer, O. Algebraische und metrische Strukturen in der Intervallrechnung und einige Anwendungen. *Computing*, **5**, 1970, 144–162.
14. Ohmann, D. Ungleichungen für die Minkowskische Summe und Differenz konvexer Körper. *Comment. Math. Helvet.*, **30**, 1956, 297–307.
15. Popov, A. A Relation between Morphological and Interval Operations. *Reliable Computing*, **4**, 2, 1998, 167–178.
16. Rådström, H. An Embedding Theorem for Spaces of Convex Sets. *Proc. Amer. Math. Soc.*, **3**, 1952, 165–169.
17. Ratschek, H., G. Schröder. Über den quasilinearen Raum. Forsch.-Zentr. Graz, 1975.
18. Ratschek, H., G. Schröder. Representation of Semigroups as Systems of Compact Convex Sets. *Proc. Amer. Math. Soc.*, **65**, 1977, 24–28.
19. Rockafellar, R. T. Convex Analysis. Princeton, 1970.
20. Schneider, R. Convex Bodies: The Brunn-Minkowski Theory. Cambridge Univ. Press, 1993.
21. Tagamlitzki, Y. On the Separability Principle in Abelian Associative Spaces. *Bull. Inst. Math.*, Bulg. Acad. Sci., **7**, 1963, 169–183 (in Bulgarian).

Received February 27, 1998
Revised May 22, 1998

Institute of Mathematics and Computer Sciences
Bulgarian Academy of Sciences,
"Acad. G. Bonchev" st., Bl. 8
BG-1113 Sofia, BULGARIA
E-mail: `smarkov@iph.bio.acad.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

SINGULARLY PERTURBED DELAYED DIFFERENTIAL INCLUSIONS

TZANKO D. DONCHEV, IORDAN I. SLAVOV

The paper deals with singularly perturbed differential inclusions with time lag. The limit behaviour of the solution set when the singular parameter tends to zero is investigated. The limits of the fast solutions are considered as Radon probability measures. Then the upper semicontinuity of the solution set with respect to uniform convergence of the slow motions and to weak probability convergence of the fast motions is examined.

Keywords: singular perturbation, delayed differential inclusions, Radon measures
1991/95 Math. Subject Classification: main 34A60, 34K99, secondary 34D15

1. INTRODUCTION

The paper deals with singularly perturbed differential inclusions with time lag, having the form

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x_t, y_t), \quad x_0 = \varphi, y_0 = \psi, \tag{1}$$

where $x \in \mathbf{R}^n, y \in \mathbf{R}^m, t \in I \overset{\text{def}}{=} [0,1]$ and $\varepsilon > 0$ represents the singular perturbation. For any $z : [-\tau, 1] \to \mathbf{R}^k$ and $t \in [0,1]$ we let $z_t : [-\tau, 0] \to \mathbf{R}^k$ be defined by $z_t(s) = z(t+s)$, $-\tau \le s \le 0$. Here $\tau > 0$, H is a set-valued map from $I \times C([-\tau,0], \mathbf{R}^n) \times L^1([-\tau,0], \mathbf{R}^m)$ into $\mathbf{R}^{n+m}$ and $\varphi \in C([-\tau,0], \mathbf{R}^n), \psi \in C([-\tau,0], \mathbf{R}^m)$, where C and L^p, $1 \le p < \infty$, are the usual spaces of respectively continuous (equipped with the uniform norm) and p-integrable functions.

The limit behavior of the solution set when the small parameter ε tends to zero is investigated here. In the literature there are mainly three ways to deal with the problem.

1. *Reduction*. In this case we consider solution set $Z(\varepsilon)$, $\varepsilon > 0$, of (1) consisting of all AC (absolutely continuous) functions (x, y) satisfying (1) for a.e. $t \in I$. For $\varepsilon = 0^+$ it is natural to mean by $Z(0)$ the set of all pairs (x, y), with x-AC and y-integrable on I, satisfying for a.e. $t \in I$ the "degenerate" inclusion

$$\begin{pmatrix} \dot{x}(t) \\ 0 \end{pmatrix} \in H(t, x_t, y_t), \quad x_0 = \varphi, \ y_0 = \psi. \tag{2}$$

The connection between the inclusions (1) and (2) has been investigated in many papers when they are *ordinary* — [4, 7, 8, 13, 15, 16]. The LSC (lower semicontinuity) is proved first in [15] in the ordinary differential case and afterwards for more general systems in [5, 6]. The topology considered is $C \times L^2$. However, to prove the USC (upper semicontinuity) in this topology, one has to "expand" in some sense the set $Z(0)$, but then the LSC will be no longer valid. It is easy to prove USC in the weaker $C \times (L^2\!-\!weak)$ topology but under restrictive conditions. It was done in [4], where the first result concerning "reduction" technique for nonlinear differential inclusions is published.

Considering more general functional-differential inclusion than (1), namely

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x, y, x_t, y_t), \quad x_0 = \varphi, \ y_0 = \psi, \tag{3}$$

we proved in [5] under one-sided Lipschitz condition the USC of $Z(\varepsilon)$ at $\varepsilon = 0^+$ in $C \times (L^2\!-\!weak)$ topology. However, generally we do not have LSC. Making restrictive assumptions concerning the dependence of the right-hand side of (3) on y, we get in [5, 6] LSC in some partial cases.

2. *Averaging*. This approach is used mainly for systems in the form

$$\begin{aligned} &\dot{x}(t) \in F(t, x, y, u(t)), \quad x(0) = x^0, \\ &\varepsilon\dot{y}(t) \in G(x, y, u(t)), \quad y(0) = y^0. \end{aligned} \tag{4}$$

Here $u(t) \in U$ (U — compact subset), and $u(\cdot)$ plays the role of a control.

Fix x and consider the following *associated* system:

$$\begin{aligned} &x = \text{const}, \\ &\dot{y}(\tau) \in G(x, y(\tau), u(\tau)), \quad y(0) \in Q \subset \mathbf{R}^m, \quad u(\tau) \in U, \quad \tau \geq 0. \end{aligned} \tag{5}$$

For given x and t the Aumann's integral

$$\bar{V}(t, x, S, Q) = \text{cl}\left\{ \frac{1}{S} \int_0^S F(t, x, Y(\tau, x, S, Q), u(\tau))\, d\tau : u(\tau) \in U \right\},$$

where $Y(\tau, x, S, Q)$ is the solution set on the interval $[0, S]$ of (5) and "cl" denotes the closed hull, possesses a limit

$$\bar{V}(t,x) = \lim_{S\to\infty} \bar{V}(t,x,S,Q)$$

when certain conditions are met. Then it can be shown, see, e.g. [10, 11], that the "slow part", i.e. the projection of $Z(\varepsilon)$ on $\mathbf{R}^n$, converges in the C-topology to the solution set of the averaged inclusion

$$\dot{x}(t) \in \bar{V}(t,x), \quad x(0) = x^0, \quad t \in I.$$

Some other averaging results are obtained in [9, 12].

In the forthcoming paper [7] we combine the averaging technique with the notion of generalized solutions (introduced via Radon probability measures over a compact set K containing all "fast" solutions) and obtain that $Z(\varepsilon)$ has a limit at $\varepsilon = 0^+$ in $C \times [L^1(I, C(K))]^*$–*weak*$*$ topology.

3. *Invariant measures.* The fundamental theorem of Tikhonov [14] states that for single-valued H depending on (x, y) instead of (x_t, y_t), i.e. $H \equiv H(t, x, y)$, under appropriate conditions the unique solution of (1) converges as $\varepsilon \to 0$ to a special solution of (2) in $C(I, \mathbf{R}^n) \times C([\delta, 1], \mathbf{R}^m)$ for every $0 < \delta < 1$.

Its recent generalizations for systems of ordinary differential equations and control systems are done in [1, 2, 17]. They are based on the identification of the limits of the fast solutions y_ε with invariant measures of the associated system. The convergence in y_ε is in some statistical sense, while the slow part converges to a solution of specially defined "reduced" system.

We finish the introduction with some notations and definitions. For $A \subset \mathbf{R}^{n+m}$, we denote by $\hat{A}$ the projection of A on $\mathbf{R}^n$ and by $\tilde{A}$ the projection of A on $\mathbf{R}^m$. Throughout the paper $\langle \cdot, \cdot \rangle$ is the scalar product, $|\cdot|$ is the norm. For a set A denote by $\sigma(x, A) := \sup\limits_{y\in A} \langle x, y \rangle$ its support function and by $D_H(A, B)$ the Hausdorff distance between the sets A, B.

The multifunction F from the space X into the space Y is said to be U(pper) S(emi) C(ontinuous) (L(ower)S(emi)C(ontinuous)) at $x \in X$ when to every open $V \supset F(x)$ $(V \bigcap F(x) \neq \emptyset)$ there exists a neighbourhood $W \ni x$ such that $V \supset F(y)$ $(V \bigcap F(y) \neq \emptyset)$ for $y \in W$. All the concepts non-discussed in details in the sequel can be found in [3] or [18].

2. THE RESULTS

Suppose that:

A1. The map H is compact, convex valued, bounded on the bounded sets. Also $H(\cdot, \alpha, \beta)$ is measurable and $H(t, \cdot, \cdot)$ is USC.

A2. There exist constants $a, b, \mu > 0$ such that for every $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$ and a.e. $t \in I$:

$$\begin{aligned}\sigma(x, \hat{H}(t,\alpha,\beta)) &\leq a(1 + |\alpha(0)|^2 + \|\beta\|_C^2), \quad \alpha \in \Omega_1,\ \beta \in C([-\tau,0],\mathbf{R}^m),\\ \sigma(y, \tilde{H}(t,\alpha,\beta)) &\leq b(1 + \|\alpha\|_C^2) - \mu|\beta(0)|^2, \quad \alpha \in C([-\tau,0],\mathbf{R}^n),\ \beta \in \Omega_2.\end{aligned}$$

Here

$$\Omega_1 = \left\{\alpha \in C([-\tau,0],\mathbf{R}^n) : |\alpha(0)| = \|\alpha\|_C = \max_{-\tau \leq s \leq 0} |\alpha(s)|\right\},$$

$$\Omega_2 = \left\{\beta \in C([-\tau,0],\mathbf{R}^m) : |\beta(0)| = \|\beta\|_C = \max_{-\tau \leq s \leq 0} |\beta(s)|\right\}$$

and $\alpha(0) = x, \beta(0) = y$.

First, we prove the following lemma:

Lemma 1. *There exist constants $N_x, N_y, L > 0$ such that*

$$\|x^\varepsilon\|_C \leq N_x, \quad \|y^\varepsilon\|_C \leq N_y, \quad |H(t, x_t^\varepsilon, y_t^\varepsilon)| \leq L$$

for every $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon), \varepsilon > 0$ and $t \in I$.

Proof. Let $\varepsilon > 0$ be given and let $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon)$. Denote

$$p(t) = \max_{-\tau \leq s \leq 0} |x^\varepsilon(t+s)|^2, \quad q(t) = \max_{-\tau \leq s \leq 0} |y^\varepsilon(t+s)|^2.$$

From A2 it follows that

$$\begin{aligned}\langle x^\varepsilon(t), \dot{x}^\varepsilon(t)\rangle &\leq \sigma(x^\varepsilon(t), \hat{H}(t, x_t^\varepsilon, y_t^\varepsilon))\\ &\leq a(1 + |x^\varepsilon(t)|^2 + \|y_t^\varepsilon\|_C^2)\end{aligned}$$

when $|x^\varepsilon(t)| = \|x_t^\varepsilon\|_C := \max_{-\tau \leq s \leq 0} |x^\varepsilon(t+s)|$, and

$$\begin{aligned}\langle y^\varepsilon(t), \varepsilon\dot{y}^\varepsilon(t)\rangle &\leq \sigma(y^\varepsilon(t), \tilde{H}(t, x_t^\varepsilon, y_t^\varepsilon))\\ &\leq b(1 + \|x_t^\varepsilon\|_C^2) - \mu|y^\varepsilon(t)|^2\end{aligned}$$

when $|y^\varepsilon(t)| = \|y_t^\varepsilon\|_C := \max_{-\tau \leq s \leq 0} |y^\varepsilon(t+s)|$.

Obviously, $p(\cdot)$ and $q(\cdot)$ are absolutely continuous functions, hence a.e. differentiable. Then we have the following two possibilities for $p(t)$ and $q(t)$, respectively:

$$\begin{aligned}\dot{p}(t) &\leq 2a(1 + p(t) + q(t)) \text{ or } \dot{p}(t) \leq 0,\\ \varepsilon\dot{q}(t) &\leq 2b(1 + p(t)) - 2\mu q(t) \text{ or } \dot{q}(t) \leq 0,\end{aligned}$$

reasoning like in the proof of [5, Lemma 2.1]. It is not difficult to see that $p(t) \leq u(t)$, $q(t) \leq v(t)$, where

$$\begin{aligned}\dot{u}(t) &= 2a(1 + u(t) + v(t)), \quad u(0) = \max\left\{\|\varphi\|_C, \frac{\mu}{b}\|\psi\|_C\right\},\\ \varepsilon\dot{v}(t) &= 2b(1 + u(t)) - 2\mu v(t), \quad v(0) = \psi(0).\end{aligned}$$

By the first equation $\dot{u}(t) \geq 0$, $t \in I$, so $b(1+u(t))/\mu$ is increasing function. Then, since $v(0) \leq b(1+u(0))/\mu$, we have $v(t) \leq b(1+u(t))/\mu$, $t \in I$. Suppose the opposite, i.e. that there are $t_0 \in (0,1)$ and $\delta > 0$ such that $v(t_0) = b(1+u(t_0))/\mu$ and $v(t) > b(1+u(t))/\mu$, $t \in (t_0, t_0+\delta)$. Therefore, by the second equation of the above system, $\dot{v}(t) < 0$, $t \in (t_0, t_0+\delta)$, thus $v(t)$ decreases and

$$v(t) < v(t_0) = \frac{b}{\mu}(1+u(t_0)) \leq \frac{b}{\mu}(1+u(t)) \text{ for } t \in (t_0, t_0+\delta).$$

This is a contradiction.

Now, we get

$$\dot{u}(t) \leq 2a\left(1+u(t)+\frac{b}{\mu}(1+u(t))\right) = M(1+u(t)),$$

where $M = 2a(1+b/\mu)$. By virtue of Gronwall inequality one obtains

$$\begin{aligned} u(t) &\leq (M+u(0))\exp(M) = N_x^2, \\ v(t) &\leq \frac{b}{\mu}(1+u) \leq \frac{b}{\mu}(1+(M+u(0))\exp(M)) = \frac{b}{\mu}\left(1+N_x^2\right). \quad \square \end{aligned}$$

Remark 1. Obviously, we have that

$$N_x^2 = \exp(M)(M+u(0)), \quad N_y^2 = \frac{b}{\mu}\left(1+N_x^2\right),$$

where M and $u(0)$ are defined in the proof above. Furthermore, the boundedness for $\varepsilon = 0$ can be easily proven using Gronwall lemma.

Remark 2. We use A2 only to prove Lemma 1, so we could replace A2 by the requirement of boundedness of all solutions of (1), uniformly in $\varepsilon \geq 0$ and $t \in I$. Or we could assume A2 only locally — over the closed ball (in $\mathbf{R}^{n+m}$) with radius $\left(N_x^2+N_y^2\right)^{1/2}$ and centered at zero, which the solutions of (1) could not abandon.

We give a simple example where A2 is satisfied.

Example. Consider the following control system:

$$\begin{aligned} \dot{x}(t) &\in x_t + y_t + w(t), \quad x_0 \equiv 0, \\ \varepsilon\dot{y}(t) &\in x_t - 2f(y) \max_{-\tau \leq s \leq 0} |y(t+s)| + w(t), \quad y_0 \equiv 0, \end{aligned}$$

where $w(\cdot)$ is measurable, $w(t) \in [-1,1]$ a.e. in I, $f(0) = 0$ and $f(y) = y/|y|$, $y \neq 0$. Then, using the simple inequality $cd \leq (c^2+d^2)/2$, we get for α and β such that $\alpha(0) = x$, $\beta(0) = y$:

$$\langle x, \hat{H}(t,\alpha,\beta)\rangle = \langle \alpha(0), \alpha(\cdot)\rangle + \langle \alpha(0), \beta(\cdot)\rangle\langle \alpha(0), w(t)\rangle$$

$$\begin{aligned}
 & \leq & & \frac{3|\alpha(0)|^2}{2} + \frac{|\alpha(\cdot)|^2}{2} + \frac{|\beta(\cdot)|^2}{2} + \frac{|w(\cdot)|}{2} \\
 & \leq & & 2(1 + |\alpha(0)|^2 + \|\beta\|_C^2), \\
\langle y, \tilde{H}(t, \alpha, \beta) \rangle & = & & \langle \beta(0), \alpha(\cdot) \rangle - 2\langle \beta(0), f(\beta(0))|\beta(0)| \rangle + \langle \beta(0), w(t) \rangle \\
 & \leq & & 1 + \|\alpha\|_C^2 - |\beta(0)|^2 \ \text{ for } \ \beta \in \Omega_2.
\end{aligned}$$

Then $a = 2$, $b = \mu = 1$.

Theorem 1. *Let* A1, A2 *hold. Suppose in addition*

A3. For every $r \in \mathbf{R}^{n+m}$, $\alpha^i \to \alpha^0$ in $C([-\tau, 0], \mathbf{R}^n)$, and $\beta^i \to \beta^0$ in $L^1([-\tau, 0], \mathbf{R}^m)$-*weak*

$$\limsup_{i \to \infty} \sigma\left(r, H(t, \alpha^i, \beta^i)\right) \leq \sigma\left(r, H(t, \alpha^0, \beta^0)\right).$$

Then the map $\varepsilon \to Z(\varepsilon)$ *is upper semicontinuous at* $\varepsilon = 0^+$ *in* $C(I, \mathbf{R}^n) \times (L^1(I, \mathbf{R}^m)$-*weak*).

Proof. Suppose $\varepsilon_i \to 0$ and $(x^i, y^i) \in Z(\varepsilon_i)$ for $i = 1, 2, \ldots$ By Lemma 1 all sets $Z(\varepsilon)$, $\varepsilon \geq 0$, are contained in a $C(I, \mathbf{R}^n) \times L^1(I, \mathbf{R}^m)$—bounded set, so it is sufficient to prove that every cluster point of $\{(x^i, y^i)\}_{i=1}^{\infty}$ in $C(I, \mathbf{R}^n) \times (L^1(I, \mathbf{R}^m)$-*weak*) belongs to $Z(0)$. We denote where necessary a given sequence and its subsequences in the same way to simplify the notations.

Let (x^i, y^i) and (x_t^i, y_t^i), $i = 1, 2, \ldots$, be subsequences, converging to (x^0, y^0), respectively (x_t^0, y_t^0) in $C(I, \mathbf{R}^n) \times (L^1(I, \mathbf{R}^m)$-*weak*). Obviously, $\dot{x}^i(\cdot) \to \dot{x}^0(\cdot)$ in $L^1(I, \mathbf{R}^m)$-*weak*.

Let $r \in \mathbf{R}^n$ be arbitrary. Then by A3 we have

$$\limsup_{k \to \infty} \sigma\left(r, H(t, x_t^i, y_t^i)\right) \leq \sigma\left(r, H(t, x_t^0, y_t^0)\right) \quad \text{for a.e.} \quad t \in I$$

and with standard arguments (see [5]) one can show that (2) is fulfilled. □

Remark 3. We note that A3 is satisfied, for example, if for fixed (t, α) the map $H(t, \alpha, \cdot)$ has convex graph.

Reformulated Theorem 1 states that if $\{(x^\varepsilon, y^\varepsilon)\}_{\varepsilon > 0}$ is a generalized sequence of solutions of (1), then it has a subsequence converging in $C \times (L^1$-*weak*) to (x^0, y^0), where x^0 is AC, y^0 is in L^1 and

$$\begin{pmatrix} \dot{x}^0(t) \\ 0 \end{pmatrix} \in H(t, x_t^0, y_t^0), \quad x_0^0 = \varphi, \ y_0^0 = \psi, \tag{6}$$

for a.e. $t \in I$.

If A3 does not hold, we will not be able to claim the above result. But we will derive a close result considering the "fast" y-parts of $Z(\varepsilon)$ as measures over the *compact* set $K = \{y \in \mathbf{R}^m : |y| \leq N_y\}$ containing all y-solutions (N_y is the constant found in Lemma 1).

To this end let $\Re(K)$ be the set of all Radon probability measures on K and define the set of functions

$$\wp := \{\nu : I \to \Re(K) \mid \nu(\cdot) \textit{ is measurable}\}.$$

If every point $y \in K$ is considered as the Dirac measure δ_y concentrated at the point y (i.e. $\delta_y(\{y\}) = 1$), we can represent every measurable function $y : I \to K$ as $\bar{\nu}(\cdot) = \delta_{y(\cdot)}$, which is an element of $\wp$.

Let E be the space of all Caratheodory functions $f(\cdot,\cdot)$ on $I \times K$ with values in $\mathbf{R}^m$, i.e. $f(\cdot,y)$ is measurable, $f(t,\cdot)$ is continuous and integraly bounded. Then E is isometrically isomorphic to $L^1(I, C(K, \mathbf{R}^m))$ (see [18, Theorem I.5.25]). Moreover, from Dunford-Pettis theorem [18, Theorem IV.1.8], we know that $\wp$ with the weak norm topology is isomorphic to the space $[L^1(I, C(K, \mathbf{R}^m))]^*$ equipped with the weak$*$ topology. Then $\nu^i \to \nu$ for ν^i, $\nu \in \wp$ and $i = 1, 2, \dots$ if and only if

$$\int_I \Big(\int_K f(t,y)\nu^i(t)\,(dy) \Big)\, dt \to \int_I \Big(\int_K f(t,y)\nu(t)\,(dy) \Big)\, dt \quad \text{for every } f \in E,$$

which means that $y^i(\cdot) \in L^1(I, \mathbf{R}^m)$ converges to ν in $(L^1(I, C(K, \mathbf{R}^m))^*$-weak* if and only if

$$\lim_{i\to\infty} \int_I f\left(t, y^i(t)\right)\, dt = \int_I \Big(\int_K f(t,y)\nu(t)\,(dy) \Big)\, dt$$

for every $f \in E$.

Theorem 2. *Let* A1 *and* A2 *be fulfilled and let* $\{(x^\varepsilon, y^\varepsilon)\}_{\varepsilon>0}$ *be a generalized sequence of solutions of* (1) *with* $\varepsilon \to 0$. *Then there exists a subsequence* $\{(x^\varepsilon, y^\varepsilon)\}_{\varepsilon>0}$ *(denoted in the same way) such that* $x^\varepsilon \to x^0$ *in* C *and* $y^\varepsilon \to \nu$ *in the weak* topology of* $[L^1(I, C(K))]^*$ *as* $\varepsilon \to 0$.

Proof. Suppose $\varepsilon \to 0$ and $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon)$ for every $\varepsilon > 0$. The net $\{x^\varepsilon(\cdot)\}_{\varepsilon>0}$ is $C(I, \mathbf{R}^n)$ precompact due to Lemma 1 and to Arzela-Ascoli theorem. We know that $\{y^\varepsilon(\cdot)\}_{\varepsilon>0}$ is $[L^1(I, C(K, R^m))]^*-weak^*$ precompact [18, Theorem IV.2.1]. Therefore passing to subsequences if necessary, $(x^\varepsilon, y^\varepsilon)$ converges to (x^0, ν) in considered topology, where $\nu \in \wp$. □

Obviously we have $x_t^\varepsilon \to x_t^0$ in $C([-\tau, 0], \mathbf{R}^n)$ and $y_t^\varepsilon \to \nu_t$ in $L^1([-\tau, 0], \mathcal{L})$ for every $t \in I$, where $\mathcal{L} = [L^1(I, C(K, \mathbf{R}^m))]^* - weak^*$. But more important question is to define an inclusion corresponding to (6) which is satisfied by x^0 and ν (like in [7], where ordinary differential inclusions are considered). In some partial cases it is possible.

Consider first a functional-differential inclusion with constant time lag $\tau > 0$:

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x_t, y, y(t-\tau)), \quad x_0 = \varphi,\ y(s) = \psi(s),\ s \in [-\tau, 0]. \tag{7}$$

Theorem 3. *Suppose the following is true*:

A1′. The map H is compact, convex valued, bounded on the bounded sets. Also H is almost continuous, i.e. for every $\delta > 0$ there exists $I_\delta \subset I$ with measure greater than $1 - \delta$ such that H is continuous on $I_\delta \times \mathbf{R}^{m+2n}$.

A2′. There exist constants $a, b, \mu > 0$ such that for every $x \in \mathbf{R}^n$ and $y, v \in \mathbf{R}^m$

$$\begin{aligned}\sigma(x, \hat{H}(t, \alpha, y, v)) &\leq a(1 + |\alpha(0)|^2 + |y|^2 + |v|^2), \quad \alpha \in \Omega_1, \\ \sigma(y, \tilde{H}(t, \alpha, y, v)) &\leq b(1 + \|\alpha\|_C^2) - \mu|y|^2, \quad \alpha \in C([-\tau, 0], \mathbf{R}^n),\end{aligned}$$

for a.e. $t \in I$. Here $\alpha(0) = x$, $v(t) = y(t - \tau)$.

Then to every generalized sequence $\{(x^\varepsilon, y^\varepsilon)\}_{\varepsilon>0}$ of solutions of (1) *there exists a subsequence (denoted in the same way) such that $x^\varepsilon \to x^0$ and $y^\varepsilon \to \nu$ in the same topologies as in Theorem* 2 *and*

$$\begin{pmatrix} \dot{x}^0(t) \\ 0 \end{pmatrix} \in \int_{K\times K} H(t, x_t^0, z)\mu(t)(dz), \quad x_0 = \varphi, \tag{8}$$

where $\mu(t)$ is the measure product $\nu(t) \otimes \nu(t - \tau)$. Here $\nu(s) = \delta_{\psi(s)}$, $s \in [-\tau, 0]$.

Proof. Substitute $z(t) = (y(t),\ y(t - \tau))$. Then like in the proof of Theorem 2 we have $\varepsilon_i \to 0$ and $(x^i, y^i) \in Z(\varepsilon_i)$ for every $i = 1, 2, \dots$ such that (passing to subsequences if necessary) (x^i, z^i) converges to (x^0, μ) in considered topologies and $(\dot{x}^i(\cdot), \varepsilon_i\dot{y}^i(\cdot))$ converges to $(\dot{x}_0(\cdot), 0)$ in $L^1(I, \mathbf{R}^{n+m})$-*weak*.

Let $r \in \mathbf{R}^{n+m}$ be arbitrary and let $[s, t] \subset I$. For every i one has

$$\left\langle r, (x^i(t) - x^i(s), \varepsilon_i(y^i(t) - y^i(s)))\right\rangle \leq \int_s^t \sigma(r, H(\tau, x^i(\tau), z^i(\tau)))\, d\tau.$$

Due to [18, Theorem IV.2.9],

$$\lim_{i\to\infty} \int_s^t \sigma(r, H(\tau, x^i(\tau), z^i(\tau)))\, d\tau = \int_s^t \Big\{ \int_{K\times K} \sigma(r, H(\tau, x_0(\tau), z))\mu_0(\tau)\,(dz) \Big\}\, d\tau.$$

Combining the above two inequalities, we obtain

$$\left\langle r, (x^0(t) - x^0(s), 0)\right\rangle \leq \int_s^t \Big\{ \int_{K\times K} \sigma(r, H(\tau, x^0(\tau), z))\mu_0(\tau)\,(dz) \Big\}\, d\tau \tag{9}$$

for every $t \geq s \in I$. Consequently, $x_0(0) = x^0$ and x^0, μ satisfy (8). □

Take now a functional-differential inclusion with two variable time lags:

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x_t, y, y(t - h_1(t)), y(t - h_2(t))),$$

$$x_0 = \varphi, \quad y(s) = \psi(s), \quad s \in [-\tau, 0], \tag{10}$$

where $h_1(t), h_2(t) \in [0, \tau]$ and h_1, h_2 are Borel measurable functions on I. The measurability of $h_i(\cdot)$ is required to assure the existence of solutions of (10). We can formulate the same result for (10) like in Theorem 3.

Theorem 4. *Suppose that the following conditions are satisfied:*

A1″. The map H is compact, convex valued, bounded on the bounded sets. Also H is almost continuous, i.e. for every $\delta > 0$ there exists $I_\delta \subset I$ with measure greater than $1-\delta$ such that H is continuous on $I_\delta \times \mathbf{R}^{m+3n}$.

A2″. There exist constants $a, b, \mu > 0$ such that for every $t \in I$, $(x(t), y(t)) \in \mathbf{R}^{n+m}$

$$\begin{aligned}
&\sigma(x(t), \hat{H}(t, x_t, y, y(t-h_1(t)), y(t-h_2(t))))\\
&\qquad \le a(1 + |x(t)|^2 + |y(t)|^2 + |y(t-h_1(t))|^2 + |y(t-h_2(t))|^2),\\
&\sigma(y(t), \tilde{H}(t, x_t, y, y(t-h_1(t)), y(t-h_2(t))))\\
&\qquad \le b(1 + |x(t)|^2 + |y(t-h_1(t))|^2 + |y(t-h_2(t))|^2) - \mu|y(t)|^2.
\end{aligned}$$

A3′. If $\inf_{t\in I}\{h_1(t), h_2(t)\} = 0$, then $\mu > 2b$.

Then to every generalized sequence $\{(x^\varepsilon, y^\varepsilon)\}_{\varepsilon>0}$ *of solutions of* (10) *there exists a subsequence (denoted in the same way) such that* $x^\varepsilon \to x^0$ *and* $y^\varepsilon \to \nu$ *in the same topologies as in Theorem* 2 *and*

$$\begin{pmatrix} \dot{x}^0(t) \\ 0 \end{pmatrix} \in \int\limits_{K^3} H(t, x_t^0, z)\mu(t)\,(dz), \quad x_0 = \varphi, \tag{11}$$

where $\mu(t) = \nu(t) \otimes \nu(t-h_1(t)) \otimes \nu(t-h_2(t))$ *and* $\nu(s) = \delta_{\psi(s)}$, $s \in [-\tau, 0]$.

Proof. Using A2″ and A3′, we can prove a result analogous to Lemma 1, see, e.g. [5]. Then substituting $z(t) = (y(t), y(t-h_1(t)), y(t-h_2(t)))$ again, like in the previous proof, we have $\varepsilon_i \to 0$ and $(x^i, y^i) \in Z(\varepsilon_i)$, $i = 1, 2, \dots$ such that (passing to subsequences if necessary) (x^i, z^i) converges to (x^0, μ) in considered topologies and $(\dot{x}^i(\cdot), \varepsilon_i \dot{y}^i(\cdot))$ converges to $(\dot{x}_0(\cdot), 0)$ in $L^1(I, \mathbf{R}^{n+m})$-*weak*.

Now, we will show that (x^0, μ) satisfies (11). The proof is very similar to the previous one — we just have to prove (9) (with K^3 in the limits of the second integral instead of $K \times K$) for any $r \in \mathbf{R}^{n+m}$ and $[s, t] \subset I$.

Since H is almost continuous, we have by [18, Theorem IV.2.9]

$$\lim_{i\to\infty} \int_s^t \sigma(r, H(\tau, x^i(\tau), z^i(\tau)))\,d\tau \le \int_s^t \Big\{ \int_{K^3} \sigma(r, H(\tau, x_0(\tau), z))\mu_0(\tau)\,(dz) \Big\}\,d\tau.$$

Consequently,

$$\langle r, (x^0(t) - x^0(s), 0)\rangle \le \int_s^t \Big\{ \int_{K^3} \sigma(r, H(\tau, x^0(\tau), z))\mu_0(\tau)\,(dz) \Big\}\,d\tau$$

for every $r \in \mathbf{R}^{n+m}$ and $t \ge s \in I$, which finishes the proof. □

Obviously, we are able to extend the above result for inclusions with finite number of delays

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in H(t, x_t, y, y(t-h_1(t)), \dots, y(t-h_k(t))),\ x_0 = \varphi, y(s) = \psi(s), s \in [-\tau, 0],$$

where $h_j(t) \in [0, \tau]$, $j = 1, \ldots, k$, and h_j are Borel measurable functions on I. But proving the corresponding theorem for the general case (1) is an open question.

ACKNOWLEDGEMENTS. This work is partially supported by the National Foundation for Scientific Research at the Bulgarian Ministry of Science and Education under Grant MM-701/97.

REFERENCES

1. Artstein, A., V. Gaitsgory. Tracking fast trajectories along a slow dynamics: A singular perturbation approach. *SIAM J. Control & Optimization*, **35**, 1997, 1487–1507.
2. Artstein, A., A. Vigodner. Singularly perturbed ordinary differential equations with fast dynamics. *Proceedings of the Royal Society of Edinburgh*, **126A**, 1996, 541–569.
3. Deimling, K. Multivalued Differential Equations. De Gruyter, Berlin, 1992.
4. Dontchev, A., I. Slavov. Upper semicontinuity of solutions of singularly perturbed differential inclusions. In: *System Modelling and Optimization*, eds. H.-J. Sebastian and K. Tammer, *Lecture Notes in Control and Information Sciences*, Vol. 143, Springer-Verlag, Berlin etc., 1991, 273–280.
5. Donchev, T., I. Slavov. Singularly perturbed functional differential inclusions. *Set Valued Analysis*, **3**, 1995, 113–128.
6. Donchev, T., I. Slavov. Tikhonov's theorem for functional-differential inclusions. *Ann. Sof. Univ., Fac. Math. et Inf.*, **89**, 1, 1995.
7. Donchev, T., I. Slavov. Averaging method for one-sided Lipschitz differential inclusions with generalized solutions. *SIAM J. Control & Optimiztion* (submitted).
8. Dontchev, A., V. Veliov. Singular perturbation in Mayer's problem for linear systems. *SIAM J. Control Optimization*, **21**, 1983, 566–581.
9. Filatov, O. Averaging of differential inclusions with control. *Differential Equations*, **33**, 1997, 782–785 (in Russian).
10. Gaitzgory, V. Control of Systems with Fast and Slow Motions. Nauka, Moscow, 1991 (in Russian).
11. Grammel, G. Singularly perturbed differential inclusions: an averaging approach. *Set-Valued Analysis*, **4**, 1996, 361–374.
12. Plotnikov, V. Averaging Methods in Control Problems. Libid Kiev, Odessa, 1992 (in Russian).
13. Quincampoix, M. Singular perturbations for systems of differential inclusions. In: *Geometry in Nonlinear Control and Differential Inclusions*, Banach Center Publications, Warsaw, 1995, 341–348.
14. Tikhonov, A. Systems of differential equations containing a small parameter in the derivatives. *Mat. Sbornik*, **31(73)**, 1952, 575–586 (in Russian).
15. Veliov, V. Differential inclusions with stable subinclusions. *Nonlinear Anal. TMA*, **23**, 1994, 1027–1038.
16. Veliov, V. A generalization of the Tikhonov theorem for singularly perturbed differential inclusions. *J. Dynamical and Control systems*, **3**, 1997, 291–319.

17. Vigodner, A. Limits of singularly perturbed control problems with statistical dynamics of fast motions. *SIAM J. Control & Optimization*, **35**, 1997, 1–28.

18. Warga, R. Optimal Control of Differential and Functional Differential Equations. Academic Press, 1973.

Received February 27, 1998
Revised May 6, 1998

Tzanko Donchev
Department of Mathematics
University of Mining and Geology
BG-1100 Sofia, BULGARIA
E-mail: **donchev@staff.mgu.bg**

Iordan Slavov
Institute of Applied Mathematics
Technical University of Sofia, bl. 2
BG-1000 Sofia, BULGARIA
E-mail: **iis@vmei.acad.bg**

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

A SEPARATION THEOREM OF Y. TAGAMLITZKI IN ITS NATURAL GENERALITY

DIMITER SKORDEV

It is shown how the assumptions of a separation theorem of Y. Tagamlitzki can be weakened without any essential change of the proof. In contrast to the original version of the theorem, the obtained thus strengthened version is not an instance of Ellis' separation theorem.

Keywords: segment, convex set, half-space, separation theorem, axiomatization
1991/95 Math. Subject Classification: main 52A01, secondary 46A22

1. INTRODUCTION

In Y. Tagamlitzki's paper [3] an axiomatization of the notion of segment is used as a basis for an abstract approach to separation of convex sets. The axiomatization looks as follows.

A set K is supposed to be given, and a subset ab of K is supposed to be put into correspondence to any a and b in K in such a way that always $ab = ba$. By definition, $a/b = \{x \in K : a \in bx\}$.[1] The following denotations are adopted for any elements a and b of the set K and any its subsets A and B:

$$aB = \bigcup\{ab : b \in B\},\ Ab = \bigcup\{ab : a \in A\},\ AB = \bigcup\{ab : a \in A, b \in B\},$$

[1] We use this denotation instead of $\frac{a}{b}$ used in [3] (and, similarly, further for a/B, A/b, A/B). Another denotational difference is that we shall designate a set inclusion by $\subseteq$, whereas Tagamlitzki designates it by $\subset$.

$$a/B = \bigcup\{a/b : b \in B\},\ A/b = \bigcup\{a/b : a \in A\},\ A/B = \bigcup\{a/b : a \in A, b \in B\}.$$

The two operations considered so far will be called *multiplication* and *division*, respectively.

Two associativity laws are supposed to hold for any a, b, c in K, namely,

$$(ab)c = a(bc), \quad a(b/c) \subseteq (ab)/c$$

(the first of these conditions allows freely using expressions of the form abc for arbitrary a, b, c in K).

Remark. After quite a time from the appearance of the paper [3] it became known that a somewhat more restrictive but similar axiomatization of the notion of segment had been given earlier by W. Prenowitz in [2]. It is easy to see that

$$a(b/c) \subseteq (ab)/c$$

for all a, b, c in K iff Prenowitz' transposition law (cf. [2, pp. 4 and 7])

$$(a/b) \cap (c/d) \neq \emptyset \Rightarrow (ad) \cap (bc) \neq \emptyset$$

holds for any a, b, c, d in K. Having this in mind, one sees that Prenowitz's join spaces from [2] coincide with the structures satisfying Tagamlitzki's axioms plus the additional ones (not required in [3]) that $ab \neq \emptyset$, $a/b \neq \emptyset$, $aa = \{a\}$ and $a/a = \{a\}$ for any a, b in K. Therefore any join space is surely a model for Tagamlitzki's axiomatization. In particular, the elements of an arbitrary vector space K form such a model if one sets

$$ab = \{pa + qb \ : \ p > 0, q > 0, p + q = 1\}.$$

The converse is not true, since the other models indicated in [3] do not satisfy, in general, the whole set of conditions in Prenowitz' definition of join space. We should like especially to mention as an example of such other model the one (indicated on p. 173), where K is again a vector space, but we have

$$ab = \{\lambda a + \mu b \ : \ \lambda > 0, \mu > 0\}$$

(the conditions $aa = \{a\}$ and $a/a = \{a\}$ are violated in this model for any non-zero element of K).

To reduce the number of brackets, we accept the convention that multiplication and division have a higher priority than $\cap$ and $\cup$ (thus we could omit the brackets in Prenowitz' transposition law mentioned above).

A subset C of K is called *convex* if the condition $CC \subseteq C$ holds. A *half-space* is a non-empty convex subset S of K such that $K \setminus S$ is also convex and non-empty. The following separation theorem plays a central role in [3]:

Theorem 1.1 (Theorem 1 of [3]). *Let $abb \subseteq ab$ for any a, b in K.*[2] *Then for any two disjoint non-empty convex subsets A and B of K there is a half-space that contains A and does not meet B.*

[2] This condition is surely satisfied in join spaces, since then $abb = a(bb) = ab$. The model mentioned at the end of the remark preceding the theorem also satisfies the condition in question, and we again have the equality $abb = ab$ in this model.

As it became clear later, the above formulated result is an instance of a more general separation theorem of J. W. Ellis published in [1]. Ellis' approach is based on a direct axiomatization of the notion of convex subset of a given set (without axiomatizing the notion of segment), and it turns out that the family of all convex subsets of K in the situation considered in the above theorem satisfies the assumptions of Ellis' one. The present paper aims at showing that Tagamlitzki's proof actually establishes a result stronger than Theorem 1.1 and this result is no more an instance of Ellis' theorem. Namely, a reduction of the assumptions of Theorem 1.1 will be done in the next section without making essential changes in its proof from [3].

2. REDUCTION OF THE ASSUMPTIONS OF TAGAMLITZKI'S SEPARATION THEOREM

We are going to formulate now the stronger result mentioned at the end of the previous section.

First of all, we reduce the assumptions from the beginning of Section 1 by omitting the first associativity law. For the reader's convenience, we formulate now what is remaining from those assumptions. Namely, we suppose in/the present section a set K to be given and a subset ab of K to be put into correspondence to any a and b in K in such a way that always the equality $ab = ba$ and the inclusion $a(b/c) \subseteq (ab)/c$ hold, adopting the denotations introduced in Section 1 before the formulation of the associativity laws.

Clearly, the definition of convex set remains the same as in Section 1, but the absence of the first associativity law obliges us now to write all brackets in the expressions that are built up by more than one application of multiplication. In particular, Theorem 1.1 does not make sense now without specifying the meaning of its assumption that $abb \subseteq ab$ for any a, b in K. (Does abb mean $(ab)b$ or $a(bb)$?) The following modification of the theorem can be established with almost no change in the proof of Theorem 1 from [3].

Theorem 2.1. *Let $(ab)b \subseteq a(bb) \subseteq ab$ for any a, b in K. Then for any two disjoint non-empty convex subsets A and B of K there is a half-space that contains A and does not meet B.*

To see the workability of the mentioned proof in the new situation considered now, it is sufficient to note that there are only two steps in the proof needing a revision: the first of them is in the transition from $\xi \in S/(xx)$ to $\xi x \cap S \neq \emptyset$ (cf. p. 174), where one has to apply now the inclusion $\xi(xx) \subseteq \xi x$, and the second one is in proving that $((SS)/x)/x$ is a subset of $(SS)/(xx)$ — to prove this inclusion (used on p. 175), one should consider an arbitrary element ξ of $((SS)/x)/x$ and apply the inclusion $(\xi x)x \subseteq \xi(xx)$. Of course, as in the original proof one uses many times the equivalence

$$BX \cap A \neq \emptyset \Leftrightarrow X \cap A/B \neq \emptyset,$$

where A, B, X can be arbitrary subsets of K.[3]

A further reduction of the assumptions is possible at the cost of a quite small change that in fact even simplifies the proof. The change consists in using the set $(S/x)/x$ instead of the set $S/(xx)$. We shall formulate now a result obtainable thanks to the admissibility of this change, and we shall present its proof following Tagamlitzki's one as close as possible, making only the necessary changes (except for small differences in the denotations).

Theorem 2.2. *Let $(ab)b \subseteq ab$ for any a, b in K. Then for any two disjoint non-empty convex subsets A and B of K there is a half-space that contains A and does not meet B.*

Proof. Let A and B be disjoint non-empty convex subsets of K. By Zorn's Lemma, there is some maximal convex subset S of K containing A and not intersecting B. We set $T = K \setminus S$ for short.

Let x be an arbitrary element of K. We shall firstly prove that

$$(S/x)/x \subseteq S/x\,. \tag{2.1}$$

In fact, let $\xi \in (S/x)/x$. Then $\xi x \cap S/x \neq \emptyset$, and hence $(\xi x)x \cap S \neq \emptyset$. Making use of the inclusion $(\xi x)x \subseteq \xi x$, we conclude that $\xi x \cap S \neq \emptyset$, hence $\xi \in S/x$.

It is easy to see now that the set $S \cup S/x$ is convex. Indeed, we have

$$\begin{aligned}(S \cup S/x)(S \cup S/x) &= SS \cup S(S/x) \cup (S/x)S \cup (S/x)(S/x)\\ &= SS \cup S(S/x) \cup (S/x)(S/x) \subseteq SS \cup (SS)/x \cup ((S/x)S)/x\\ &\subseteq SS \cup (SS)/x \cup ((SS)/x)/x \subseteq S \cup S/x \cup (S/x)/x \subseteq S \cup S/x\end{aligned}$$

(the first two inclusions follow from the second associativity law, the convexity of S implies the inclusion next to the last, and (2.1) is applied to obtain the last one).

Let us consider now the particular case when $x \in B$. We shall show that S/x does not meet B in this case. In fact, if $S/x \cap B \neq \emptyset$, then $S \cap xB \neq \emptyset$, hence $S \cap BB \neq \emptyset$, and from here, by the convexity of B, the false conclusion $S \cap B \neq \emptyset$ follows. So S/x does not meet B and therefore the set $S \cup S/x$ also does not. By the convexity of $S \cup S/x$ and the maximality of S, we get the inclusion $S/x \subseteq S$. Thus we see that $S/x \cap T = \emptyset$, hence $x \notin S/T$. Since x can be any element of B, it follows that

$$S/T \cap B = \emptyset\,. \tag{2.2}$$

Consider now the convex set $S \cup S/x$, where x is an arbitrary element of T. By (2.2), this convex set does not meet B and hence, by the maximality of S, the inclusion $S/x \subseteq S$ holds. Since x can be an arbitrary element of T, we get the

[3] This equivalence has been observed by Ivan Prodanov about 1962, but in fact it is indicated earlier in [2] (cf. Theorem 5 of that paper).

inclusion $S/T \subseteq S$. Consequently, $S/T \cap T = \emptyset$, therefore $S \cap TT = \emptyset$, i.e. $TT \subseteq T$. So the convexity of T is established, and it remains to notice that S and T are non-empty because $A \subseteq S$ and $B \subseteq T$. □

The improved versions Theorem 2.1 and Theorem 2.2 of Tagamlitzki's Theorem 1 are not instances of the separation theorem from [1]. For any a, b in K let $[a, b]$ (*the convex closure* of the set $\{a, b\}$) be the intersection of all convex subsets of K containing both a and b as elements. In order the separation theorem from [1] to be applicable to the family $\mathcal{G}$ of the convex subsets of K, this family must satisfy the following condition: for any set X belonging to $\mathcal{G}$ and any a in K the union of all convex closures $[a, x]$, where $x \in X$, must belong to $\mathcal{G}$ too. We shall give now an example showing the existence of cases when this condition is not satisfied, but nevertheless the assumptions of Theorems 2.1 and 2.2 are fulfilled (of course, it is sufficient to check only the stronger assumptions — those of Theorem 2.1).

Example. Let K consist of five distinct elements p_1, p_2, p_3, p_4, p_5, and let the multiplication in K be defined by the condition that $x \in yz$ iff some of the three cases below is present:

(α) $x \in \{y, z\}$;

(β) $x = p_3$, $\{y, z\} = \{p_1, p_2\}$;

(γ) $x = p_5$, $\{y, z\} = \{p_3, p_4\}$.

The commutativity of the multiplication is obvious. To check the validity of the second associativity law, suppose a, b, c are elements of K, and x is an element of $a(b/c)$. We shall prove that x belongs to $(ab)/c$. We have $x \in ay$ for some y such that $b \in cy$, and we must show that $cx \cap ab \neq \emptyset$. If $x \in ay$ holds according to case (α), i.e. $x \in \{a, y\}$, then $a \in cx \cap ab$ in the case of $x = a$, and $b \in cx \cap ab$ in the case of $x = y$. The situation is similar if $b \in cy$ holds according to case (α). Now suppose that each of the statements $x \in ay$ and $b \in cy$ holds according to some of the cases (β), (γ). Since $\{a, y\} \cap \{c, y\} \neq \emptyset$, it is not possible that one of the both statements holds according to (β) and the other one holds according to (γ). Therefore $x = b$, hence the condition $cx \cap ab \neq \emptyset$ is satisfied again. We obviously have $bb = \{b\}$ for any b in K, therefore $a(bb) = ab$ for any a, b in K. We shall prove the inclusion $(ab)b \subseteq a(bb)$ by proving that $(ab)b \subseteq ab$. Suppose $x \in (ab)b$ for some a, b in K; we shall prove that $x \in ab$. We have $x \in yb$ for some $y \in ab$. But the cases of $x \in \{y, b\}$ or $y \in \{a, b\}$ are easy, and, on the other hand, it turns out to be not possible that each of the statements $x \in yb$ and $y \in ab$ holds according to some of the cases (β), (γ). So we have shown that all assumptions of Theorems 2.1 and 2.2 are satisfied in this example. Let us now consider the convex set $X = \{p_1, p_4\}$ and the union of all convex closures $[p_2, x]$, where $x \in X$. The union in question is $\{p_1, p_2, p_3\} \cup \{p_2, p_4\} = \{p_1, p_2, p_3, p_4\}$, and it is not convex due to $p_5 \in p_3p_4$.

Remark. Theorem 2.2 remains true if the inclusion $(ab)b \subseteq ab$ is replaced by the weaker one $(ab)b \subseteq ab \cup \{a, b\}$. To see this, it is sufficient to make the following changes in the proof:

- The sentence "Let x be an arbitrary element of K" must be replaced by "Let x be an arbitrary element of T".
- The inclusion (2.1) must become $(S/x)/x \subseteq S \cup S/x$.
- The third sentence after (2.1) must become "Making use of the inclusion $(\xi x)x \subseteq \xi x \cup \{\xi, x\}$, we conclude that $\xi x \cap S \neq \emptyset$ or $\xi \in S$, hence $\xi \in S \cup S/x$".

3. CONCLUDING REMARKS

We think it is quite possible that in the time of writing [3] Professor Tagamlitzki had already been aware of the possibility to prove a version of Theorem 1 in the absence of the first associativity law. In our opinion, he could have the following reasons not to mention this possibility in his paper:

- a lack of known interesting applications of such a generalization of the theorem;
- the fact that the rest of the paper anyway needs the first associativity law (Theorem 1 being mainly a tool for the considerations there);
- the lack of information about Ellis' separation theorem at that time.

There is, however, a chance that a generalization of this kind could be possibly applied in the future to some problems of interest, and also the other considerations from [3] perhaps could be generalized in some way for the case of absent first associativity law. If this happens, then the fact that Ellis' theorem does not completely cover the content of Tagamlitzki's result will turn out to be more essential than it could seem at the present moment.

REFERENCES

1. Ellis, J. W. A general set-separation theorem. *Duke Math. J.*, **19**, 1952, 417–421.
2. Prenowitz, W. A contemporary approach to classical geometry. *Amer. Math. Monthly*, **68**, No. 1, part II, 1961, vi+67pp.
3. Tagamlitzki, Y. Über die Trennbarkeitsbedingungen in den abelschen assoziativen Räumen. *Acad. des Sci. de Bulgarie, Bull. de l'Inst. de Math.*, **7**, 1963, 169–183 (in Bulgarian; Russian and German summaries).

Received March 4, 1998

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5 Blvd. J. Bourchier
BG-1164 Sofia, BULGARIA
E-mail: skordev@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

ASYMPTOTIC SOLUTION OF DEFINITE CLASS OF SINGULARLY PERTURBED LINEAR BOUNDARY-VALUE PROBLEMS FOR ORDINARY DIFFERENTIAL EQUATIONS

LJUDMIL KARANDJULOV

The singular perturbation for boundary problems for linear systems of ordinary differential equations is considered. Under suitable assumptions using generalized inverse matrix the unique asymptotic expansion with boundary function is constructed.

Keywords: boundary-value problems, singular perturbation, asymptotic solution, boundary functions

1991/1995 Math. Subject Classification: 34E15

1. INTRODUCTION

The theory of the singularly perturbed systems for ordinary differential equations is primarily due to the works of A. Tikhonov [1, 2] and N. Levinson (see [3, 19]) in the early 1950. The method and results of A. B. Vasil'eva [4, 5] and A. B. Vasil'eva, V. F. Butuzov [6, 7] widely make use of the construction asymptotic solution of a singularly perturbed differential systems. The questions connected with asymptotic calculation of relaxational oscillation are considered in the monographs [8, 9]. The method of regularization of singular perturbation is studied in [10]. A method of separation of differential equations for obtaining asymptotic decomposition similar to regularized decomposition is given in the papers [11,12]. In this paper the behavior of the solution at $\varepsilon \to 0$ is considered for a linear boundary-value problem

$$\varepsilon \dot{x} = Ax + \varepsilon A_1(t)x + \varphi(t), \quad t \in [a,b],\ 0 < \varepsilon < 1, \tag{1}$$

$$l(x) = h, \quad h \in \mathbb{R}^m, \tag{2}$$

where the coefficients of the system (1) and the equality (2) are subordinate to the conditions:

(H1) A is a constant $(n \times n)$-matrix, $\operatorname{Re}\lambda_i < 0$ $(i = \overline{1,n})$, $\lambda_i \in \sigma(A)$;

(H2) $A_1(t)$ is an $(n \times n)$-matrix, $A_1(t) \in C^\infty[a,b]$; φ is an n-vector function, $\varphi(t) \in C^\infty[a,b]$;

(H3) l is a linear m-dimensional bounded functional

$$l = \operatorname{col}(l_1, \ldots, l_m), \quad l \in (C([a,b]) \to \mathbb{R}^n, \mathbb{R}^m).$$

The condition (H1) shows that $\det A \neq 0$.

We consider the problem (1), (2) in the class of continuously differentiable functions. Then the domain $D(L_\varepsilon)$ of the operator

$$(L_\varepsilon x)(t) = \varepsilon\dot{x}(t) - Ax(t) - \varepsilon A_1(t)x(t)$$

consists of a continuously differentiable in $[a,b]$ functions, satisfying the boundary condition (2). At $\varepsilon = 0$ we obtain the degenerate equation $Ax_0(t) + \varphi(t) = 0$, which solution $x_0(t) = -A^{-1}\varphi(t)$ for arbitrary $\varphi(t) \in C^\infty[a,b]$ does not belong to the domain $D(L_\varepsilon)$ of the operator L_ε, since, in general, the condition $l(x_0) = h$ is not fulfilled.

Let the equation (1) is solvable for arbitrary $\varphi \in C^\infty[a,b]$. Then the dimension of the kernel of the operator L_ε is equal to the dimension n of the system (1) and the boundary-value problem (1), (2) is the Noetherian problem with index $n - m$: $\operatorname{ind}[L_\varepsilon, l] = n - m \neq 0$. It will be the Fredholm problem ($\operatorname{ind}[L_\varepsilon, l] = 0$) if and only if $m = n$ (see [13]).

We shall consider the case $m \neq n$. We use an asymptotic method of the boundary functions and construct an asymptotic series, satisfying the boundary-value problem (1), (2) at $\det A \neq 0$. The initial research in the case is made in [14].

In the Fredholmian case ($m = n$) an asymptotic integration of boundary-value problems for non-linear and weakly non-linear systems with two-point boundary conditions is studied in [6,7] on the basis of the method of boundary functions, and in [10] — on the basis of the regularization method.

The construction of an asymptotic solution of (1), (2) in the Notherian case ($m \neq n$) is represented on the basis of generalized inverse matrices and projectors [15–17, 13].

2. FORMALLY ASYMPTOTIC EXPANSION

We shall seek a formally asymptotic expansion of the solution of the problem (1), (2) in the form of the series

$$x(t,\varepsilon) = \sum_{i=0}^{\infty}[x_i(t) + \Pi_i(\tau)]\varepsilon^i, \quad \tau = \frac{t-a}{\varepsilon}, \tag{3}$$

where $x_i(t)$ and $\Pi_i(\tau)$ are unknown n vector-functions. By $\Pi_i(\tau)$ (see [6, 7]) we denote the boundary functions in a neighbourhood of the point $t = a$. They will be constructed so that when $0 < \varepsilon \leq \varepsilon_0$, the inequalities

$$\|\Pi_i(\tau)\| \leq \gamma_i \exp(-\alpha_i \tau), \tag{4}$$

where γ_i and α_i are positive constants for $i = 0, 1, 2, \dots$ and $\tau \geq 0$, hold in $[a, b]$.

Formally, by substituting (3) in (1), for $x_i(t)$ we obtain the recurrent expressions

$$x_i(t) = \begin{cases} -A^{-1}\varphi(t), & i = 0, \\ A^{-1}(Lx_{i-1})(t), & i = 1, 2, \dots, \end{cases} \tag{5}$$

where L is the differential operator $Lx = \dfrac{d}{dt}x - A_1(t)x$. The boundary functions are solutions of the differential equations

$$\frac{d}{d\tau}\Pi_i(\tau) = A\Pi_i(\tau) + f_i(\tau), \quad \tau \in [0, \tau_b], \;\; \tau_b = \frac{b-a}{\varepsilon}, \tag{6}$$

where

$$f_i(\tau) = \begin{cases} 0, & i = 0, \\ \displaystyle\sum_{q=i-1}^{0} \frac{1}{q!}\tau^q A_1^{(q)}(a)\Pi_{i-1-q}(\tau), & i = 1, 2, \dots \end{cases} \tag{7}$$

We substitute (3) in the boundary condition (2). Then the coefficients of the expansion (3) satisfy the boundary conditions

$$l(x_i) + l\left(\Pi_i\left(\frac{(\cdot) - a}{\varepsilon}\right)\right) = \begin{cases} h, & i = 0, \\ 0, & i = 1, 2, \dots \end{cases} \tag{8}$$

We denote $X(\tau) = \exp(A\tau)$ to be the normal fundamental matrix of the solutions of the linear system $\dfrac{dx}{d\tau} = Ax, \; \tau \in [0, \tau_b]$; $D(\varepsilon) = l(X) = l\left(X(\frac{(\cdot)-a}{\varepsilon})\right)$ is an $(m \times n)$-matrix.

Now consider two cases depending on the structure of the matrix $D(\varepsilon)$.

2.1. Let $D(\varepsilon) = D_0 + O\left(\varepsilon^s \exp\left(-\dfrac{\alpha}{\varepsilon}\right)\right)$, where $\alpha > 0$, $s \in \mathbb{N}$, D_0 is an $(m \times n)$-constant matrix.

All the expressions $\varepsilon^s \exp\left(-\dfrac{\alpha}{\varepsilon}\right)$ are exponentially small and it is possible to reject them, because they are of higher order of vanishing than an arbitrary degree of ε.

Let the following condition be fulfilled:

(H4) $\operatorname{rank} D_0 = n_1 < \min(m, n)$.

Denote by P and P^* the matrix orthoprojectors

$$P : \mathbb{R}^n \to \ker(D_0), \quad P^* : \mathbb{R}^m \to \ker(D_0^*), \quad D_0^* = D_0^T.$$

By D_0^+ we denote the unique Moore-Penrose inverse $(n \times m)$ -matrix of the matrix D_0 [15–17, 13]. Let P_d^* be a $(d \times m)$-matrix with $d = m - n_1$ linear independent rows from the matrix P^*, and let P_r be $r = n - n_1$ linear independent columns from the matrix P.

Consider the system (6–8) for $i = 0$. Then the boundary-value problem about $\Pi_0(\tau)$ has the form

$$\frac{d}{d\tau}\Pi_0(\tau) = A\Pi_0(\tau), \quad l(\Pi_0) = h - l(x_0). \tag{9}$$

We substitute the general solution of the system (9) $\Pi_0(\tau) = X(\tau)c_0$ in the boundary condition. Ignoring the exponentially small elements in the matrix $D(\varepsilon)$, we obtain by the algebraic system

$$D_0 c_0 = h_0, \tag{10}$$

where $h_0 = H - l(x_0)$, the n-vector c_0.

When the condition (H4) is fulfilled, the system (10) possesses a family of solutions

$$c_0 = P_r c_0^r + D_0^+ h_0$$

if and only if

$$P^* h_0 = 0 \implies P_d^* h_0 = 0.$$

Substituting c_0 in $\Pi_0(\tau) = X(\tau)c_0$, we obtain

$$\Pi_0(\tau) = X_r(\tau)c_0^r + g_0(\tau), \quad c_0^r \in \mathbb{R}^r, \tag{11}$$

where

$$X_r(\tau) = X(\tau)P_r - (n \times r)\text{-matrix}, \quad g_0(\tau) = X(\tau)D_0^+ h_0. \tag{12}$$

We define the vector $c_0^r \in \mathbb{R}^r$ by obtaining $\Pi_1(\tau)$. Consider the boundary-value problem with respect to $\Pi_1(\tau)$:

$$\frac{d}{d\tau}\Pi_1(\tau) = A\Pi_1(\tau) + f_1(\tau), \quad \tau \in [0, \tau_b], \quad l(\Pi_1) = -l(x_1), \tag{13}$$

where $f_1(\tau) = A_1(a)\Pi_0(\tau)$. Keeping in mind (11), (12), $f_1(\tau)$ will depend on the unknown vector c_0^r:

$$f_1(\tau, c_0^r) = A_1(a)X_r(\tau)c_o^r + A_1(a)g_0(\tau).$$

We substitute the general solution

$$\Pi_1(\tau) = X(\tau)c_1 + \int\limits_0^\tau X(\tau)X^{-1}(s)f_1(s)\,ds \tag{14}$$

of the differential system (13) in the boundary condition and ignoring the exponential small elements in the matrix $D(\varepsilon)$, obtain the system with respect to c_1:

$$D_0 c_1 = h_1(\varepsilon), \quad c_1 \in \mathbb{R}^n, \tag{15}$$

where

$$h_1(\varepsilon) = -l(x_1) - l\left(\int_0^{(\cdot)} X\left(\frac{(\cdot)-a}{\varepsilon}\right) X^{-1}(s) f_1(s, c_0^r)\, ds\right).$$

According to (H4), the system (15) has a solution

$$c_1 = P_r c_1^r + D_0^+ h_1(\varepsilon), \quad c_1^r \in \mathbb{R}^r,$$

if $P_d^* h_1(\varepsilon) = 0$.

From the last equality and the form of $h_1(\varepsilon)$ we obtain

$$\overline{D}(\varepsilon) c_0^r = P_d^* b_1(\varepsilon), \tag{16}$$

where

$$\overline{D}(\varepsilon) = P_d^* l\left(\int_0^{(\cdot)} X\left(\frac{(\cdot)-a}{\varepsilon}\right) X^{-1}(s) A_1(a) X_r(s)\, ds\right),$$

$$b_1(\varepsilon) = -l\left(\int_0^{(\cdot)} X\left(\frac{(\cdot)-a}{\varepsilon}\right) X^{-1}(s) A_1(a) g_0(s)\, ds\right) - l(x_1). \tag{17}$$

We assume that $\overline{D}(\varepsilon) = \overline{D}_0 + O\left(\varepsilon^p \exp\left(\frac{-\alpha}{\varepsilon}\right)\right)$, where $\alpha > 0$, $p \in \mathbb{N}$, $\overline{D}_0$ is a $(d \times r)$-constant matrix, and after ignoring the exponentially small elements, the system (16) takes the form

$$\overline{D}_0 c_0^r = P_d^* b_1(\varepsilon). \tag{18}$$

Let the following conditions be satisfied:

(H5) $\operatorname{rank} \overline{D}_0 = r$,

(H6) $\overline{P}_{d_1}^* P_d^* = 0$, $d_1 = d - r$,

where $\overline{P}^* : \mathbb{R}^d \to \ker(\overline{D}_0^*)$. Then the system (18) is always solvable and

$$c_0^r = \overline{D}_0^+ P_d^* b_1(\varepsilon). \tag{19}$$

We substitute (19) in (11) and obtain the resultant expression for $\Pi_0(\tau)$:

$$\Pi_0(\tau) = X_r(\tau) \overline{D}_0^+ P_d^* b_1(\varepsilon) + g_0(\tau). \tag{20}$$

Define the norm of the matrix $B = [b_{ij}]$ by means of the equality $\|B\| = \max_i \sum_{j=1}^n |b_{ij}|$. Keeping in mind the representation $b_1(\varepsilon)$ from (17) and the structure of the matrix

$X(\tau)$, it follows that there exists ε_0 and when $0 < \varepsilon \le \varepsilon_0$, the following inequalities are fulfilled:

$$\|b_1(\varepsilon)\| \le c_4, \quad c_4 > 0; \quad \|X_r(\varepsilon)\| \le c_1 \exp(-\alpha_1 \tau), \quad c_1 > 0, \quad \alpha_1 > 0;$$

$$\|D_0^+\| \le c_2, \quad c_2 > 0; \quad \|P_d^*\| \le c_3, \quad c_3 > 0;$$

$$\|g_0(\tau)\| \le c_5 \exp(-\alpha_2 \tau), \quad c_5 > 0, \quad \alpha_2 > 0.$$

Consequently, we can indicate positive constants γ_0, β_0 such that

$$\|\Pi_0(\tau)\| \le \gamma_0 \exp(-\beta_0 \tau),$$

that is the boundary function $\Pi_0(\tau)$ decreases exponentially.

It is obvious that $\Pi_i(\tau)$ $(i = 1, 2, \ldots)$ will be determined sequentially.

Assume that the boundary functions $\Pi_i(\tau)$ $(\overline{1, i-2})$ are defined. Then the vectors c_i $(\overline{0, i-2})$ are entirely defined. By means of $\Pi_i(\tau)$ we determine the vector c_{i-1}^r, which participates in the boundary function Π_{i-1}:

$$\Pi_{i-1}(\tau) = X_r(\tau) c_{i-1}^r + g_{i-1}(\tau), \quad c_{i-1}^r \in \mathbb{R}^r, \tag{21}$$

where $g_{i-1}(\tau) = g_{i-1}(\tau, c_{i-2}^r, \ldots, c_0^r)$.

We substitute the general solution of the system (6):

$$\Pi_i(\tau) = X(\tau) c_i + \int\limits_0^\tau X(\tau) X^{-1}(s) f_i(s, c_{i-1}^r, \ldots, c_0^r)\, ds, \quad c_{i-1}^r \in \mathbb{R}^n, \tag{22}$$

in (8) and obtain the algebraic system (ignoring the exponentially small elements in $D(\varepsilon)$)

$$D_0 c_i = h_i(\varepsilon, c_{i-1}^r, \ldots, c_0^r), \tag{23}$$

where

$$h_i(\varepsilon, c_{i-1}^r, \ldots, c_0^r)$$

$$= -l \left(\int\limits_0^{(\cdot)} X\left(\frac{(\cdot) - a}{\varepsilon} \right) X^{-1}(s) A_1(a) X_r(s)\, ds \right) c_{i-1}^r + b_i(\varepsilon, c_{i-2}^r, \ldots, c_0^r), \tag{24}$$

$$b_i(\varepsilon) = -l(x_i)$$

$$- l \left(\int\limits_0^{(\cdot)} X(\cdot) X^{-1}(s) \left[\sum_{q=i-1}^{1} \frac{1}{q!} s^q A_1^{(q)}(a) \Pi_{i-1-q}(s) + A_1(a) g_{i-1}(s) \right] ds \right).$$

From the solvability condition of the system (23)

$$P_d^* h_i(\varepsilon, c_{i-1}^r, ..., c_0^r) = 0$$

and (24) we get

$$\overline{D}(\varepsilon)c^r_{i-1} = P^*_d b_i(\varepsilon).$$

Let the conditions (H5), (H6) be satisfied. Then

$$c^r_{i-1} = \overline{D}_0^{\,+} P^*_d b_i(\varepsilon). \tag{25}$$

We substitute (25) in (21) and obtain the resultant expression for $\Pi_{i-1}(\tau)$:

$$\Pi_{i-1}(\tau) = X_r(\tau)\overline{D}_0^{\,+} P^*_d b_i(\varepsilon) + g_{i-1}(\tau), \tag{26}$$

where

$$g_{i-1}(\tau) = X(\tau)D_0^+ h_{i-1}(c^r_{i-2}, \dots, c^r_0) + \int\limits_0^\tau X(\tau)X^{-1}(s)f_{i-1}(s, c^r_{i-2}, \dots, c^r_0)\,ds.$$

Lemma 2.1. *Let the matrix A satisfy the condition* (H1), *and let the vector function $f(t) \in C[0,+\infty)$ and satisfy the inequality $\|f(t)\| \leq c^* \exp(-\alpha^* t)$, where $t \geq 0$, $c^* > 0$, $\alpha^* > 0$. Then there exist positive constants c and γ, so that the system $\dfrac{dx}{dt} = Ax + f(t)$ has a particular solution of the form*

$$\overline{x}(t) = \int\limits_0^{+\infty} K(t,s)f(s)\,ds,$$

satisfying the inequality

$$\|\overline{x}(t)\| \leq ce^{(-\gamma t)},\ t \geq 0, \tag{27}$$

where

$$K(t,s) = \begin{cases} X(t)X^{-1}(s), & \text{if } 0 \leq s \leq t < \infty, \\ 0, & \text{if } 0 < t < s \leq \infty. \end{cases}$$

Proof. The fact that $\overline{x}(t)$ is a solution is verified directly. From the condition (H1) it follows that $\|X(t)X^{-1}(s)\| \leq \overline{c}\exp(-\overline{\alpha}(t-s))$ when $\overline{c} > 0$, $\overline{\alpha} > 0$. We have

$$\|\overline{x}(t)\| \leq \int\limits_0^t \|X(t)X^{-1}\|\,\|f(s)\|\,ds \leq c^*\overline{c}e^{-\overline{\alpha}t} \int\limits_0^t e^{(\overline{\alpha}-\alpha^*)s}\,ds.$$

If $\alpha^* < \overline{\alpha}$ (or $\alpha^* > \overline{\alpha}$), then choosing $c = \dfrac{2c^*\overline{c}}{\overline{\alpha}-\alpha^*} > 0$ $\left(c = \dfrac{2c^*\overline{c}}{\alpha - \overline{\alpha}}\right)$ and $\gamma \leq \alpha^*$ ($\gamma \leq \overline{\alpha}$) we get (27).

Let $\alpha^* = \overline{\alpha}$. Then $\|x(t)\| \leq c^*\,\overline{c}\,t\,e^{-\overline{\alpha}t}$. But $\lim\limits_{t\to\infty} te^{-\overline{\alpha}t} = 0$. Consequently, there exist constants $c > 0$, $\gamma > 0$, so that (27) is fulfilled for $t \geq t_1 > 0$. □

Theorem 2.1. *Let* $D(\varepsilon) = D_0 + O\left(\varepsilon^s \exp\left(-\frac{\alpha}{\varepsilon}\right)\right)$ *and the conditions* (H1)–(H6) *be satisfied. If* $\varphi(t) \in C^\infty[a,b]$ *and* $h \in \mathbb{R}^m$ *satisfies the condition* $P_d^*(h - l(x_0)) = 0$, *the boundary-value problem* (1), (2) *has a unique formal expansion of the form* (3). *The coefficients of the expansion* $x_i(t)$ *and* $\Pi_{i-1}(\tau)$ *have the representations* (5), (20), (26), *respectively, and the boundary functions* $\Pi_i(\tau)$ *decrease exponentially.*

Proof. From the above conclusions and the conditions of the theorem it follows that really the coefficients of the expansion (3) for the boundary-value problem (1), (2) have the representations (5), (20), (26). It will be proved that the functions $\Pi_i(\tau)\,(i = 0, 1, \ldots)$ decrease exponentially. This we have done for $\Pi_0(\tau)$. Let the inequalities (4) be satisfied, that is $\|\Pi_k(\tau)\| \le \gamma_k \exp(-\alpha_k \tau)$ for $\tau \ge 0$ and $k = \overline{1, i-2}$. It is known that for $\beta_{i-1} < \alpha = \max\limits_k \alpha_k$, $c_{i-1}^* = \max\limits_k \gamma_k$ we have

$$(c_{i-1}^* \tau^{i-2} + \cdots + c_{i-1}^* \tau + c_{i-1}^*) \exp(-\alpha\tau) \le c_{i-1}^* \exp(-\beta_{i-1}\tau).$$

Thus $\|f_{i-1}(\tau)\| \le c_{i-1}^* \exp(-\beta_{i-1}\tau)$ for $\tau \ge 0$, where $f_{i-1}(\tau)$ are the functions from (7). Using this inequality, Lemma 2.1 and the estimates $\|b_i(\varepsilon)\| \le c_i$ at $\varepsilon \in (0, \varepsilon_0]$, from (26) we obtain

$$\|\Pi_{i-1}(\tau)\| \le \gamma_{i-1} \exp(-\alpha_{i-1}\tau), \quad \tau \ge 0,$$

that is the boundary functions decrease exponentially. □

Corollary 1. *Let the conditions* (H1)–(H3) *be satisfied and rank*$D_0 = n_1 = n$. *Then for any function* $\varphi(t) \in C^\infty[a,b]$ *and for any* $h \in \mathbb{R}^m$, *satisfying* $P_d^* h_i(\varepsilon) = 0$, $i = 0, 1, \ldots$, *the boundary-value problem* (1), (2) *has an unique formally asymptotic expansion in the form* (3). *The coefficients* $x_i(t)$ *have the form* (5), *and the boundary functions* $\Pi_i(\tau)$ *have the representations*

$$\Pi_i(\tau) = X(\tau) D_0^+ h_i + \int\limits_0^\tau X(\tau) X^{-1}(s) f_i(s)\, ds.$$

In this case $P = 0$, $c_i = D_0^+ h_i(\varepsilon)$ $(i = 0, 1, \ldots)$, where $h_i(\varepsilon) = b_i(\varepsilon)$, $h_0 = h - l(x_0)$.

Remark 1. If $m = n$ and $\det D_0 \ne 0$, then it is sufficient to replace D_0^+ with D_0^{-1} in Corollary 1. If $m = n$ and rank$D_0 < m = n$, then all considerations in this case coincide with the mentioned above ones.

Remark 2. If $m \ne n$, rank$D_0 = n_1 = m$, then $P^* = 0$ and all systems $D_0 c_i = h_i(\varepsilon)$, $i = 0, 1, \ldots$, are always solvable. In this case we get the family of boundary functions.

2.2. Let $D(\varepsilon) = D_0 + D_1\varepsilon + D_2\varepsilon^2 + \cdots + D_s\varepsilon^s + O(\varepsilon^q \exp(-\alpha\varepsilon))$, where D_i are $(m \times n)$-constant matrices, $\alpha > 0$, $q \in \mathbb{N}$.

We reject the exponentially small elements in $D(\varepsilon)$ and introduce the $(2s+1)m \times (s+1)n$-matrix

$$Q = \left[\begin{array}{ccccc} D_0 & & & & \\ D_1 & D_0 & & \mathbf{0} & \\ D_2 & D_1 & D_0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ D_s & D_{s-1} & D_{s-2} & \dots & D_0 \\ & D_s & D_{s-1} & \dots & D_1 \\ & & D_s & \dots & D_2 \\ & \mathbf{0} & & \ddots & \vdots \\ & & & & D_s \end{array}\right].$$

We also introduce the $(s+1)n$-vector $c_i = [c_{i0}\; c_{i1}\; \cdots\; c_{is}]^T$, where c_{ij} are n-vectors and the $(2s+1)m$-vector $b_i = [\underbrace{b_{i0}\; \cdots\; b_{is}}_{(s+1)m}\underbrace{0\; \cdots\; 0}_{sm}]^T$.

Let the following condition be fulfilled:
(H7) $\operatorname{rank} Q = (s+1)n$, $((2s+1)m > (s+1)n)$.
Then $\operatorname{rank} P_1 = 0$, $\operatorname{rank} P_1^* = d_1 = (2s+1)m - (s+1)n$, where

$$P_1 : R^{(s+1)n} \to \ker(Q), \quad P_1^* : R^{(2s+1)m} \to \ker(Q^*), \quad Q^* = Q^T.$$

The algebraic system

$$Qc_i = b_i \tag{28}$$

has the solution

$$c_i = Q^+ b_i \;\text{ or }\; c_{ij} = [Q^+ b_i]_{n_j}, \quad j = \overline{0, s} \tag{29}$$

if and only if $P_1^* b_i = 0$. So we obtain the conditions
(H8) $P^*_{1d_1} b_i = 0$, $i = 0, 1, \dots,$
where $P^*_{1d_1}$ is a $(d_1 \times (2s+1)m)$-matrix, and $[Q^+ b_i]_{n_0}, [Q^+ b_i]_{n_1}, \dots, [Q^+ b_i]_{n_s}$ are the first n elements, the second n elements, ..., the last n elements of the $(s+1)n$-vector $Q^+ b_i$, respectively.

In this case we shall seek the solution of the system $D(\varepsilon)c_0 = h_0$ in the form

$$c_0 = c_{00} + \varepsilon c_{01} + \cdots + \varepsilon^s c_{0s},$$

where $c_{0j} \in \mathbb{R}^n$, $j = \overline{1,s}$. We find the vectors c_{0j} from the system (28). From the conditions (H5), (H6) and the equality (29) for $i = 0$ and $\Pi_0(\tau)$ we obtain

$$\Pi_0(\tau) = X(\tau) \sum_{j=0}^{s} \varepsilon^j [Q^+ b_0]_{n_j}, \tag{30}$$

where $b_0 = [h_0 \; 0 \; \cdots \; 0]^T$. Obviously, the boundary function fulfills the requirement $\lim\limits_{\tau \to \infty} \Pi_0(\tau) = 0$.

Analogously, we find $\Pi_1(\tau)$ from (14) and the system $D(\varepsilon) c_1 = h(\varepsilon)$, where

$$h_1(\varepsilon) = -l(x_2) - l\left(\int_0^{(\cdot)} X\left(\frac{(\cdot) - a}{\varepsilon} \right) X^{-1}(s) f_1(s)\, ds \right), \quad f_1(\tau) = A_1(a) \Pi_0(\tau).$$

We seek c_1 in the form $c_1 = c_{10} + \varepsilon c_{11} + \cdots + \varepsilon^s c_{is}$.

Assume that after ignoring the exponentially small elements, $h_1(\varepsilon) = h_{10} + \varepsilon h_{11} + \cdots + \varepsilon^s h_{1s}$. Then we obtain

$$\Pi_1(\tau) = X(\tau) \sum_{j=0}^{s} \varepsilon^j [Q^+ b_1]_{n_j} + \int_0^{\tau} X(\tau) X^{-1}(s) f_1(s)\, ds,$$

where $b_1 = [h_{10} \; \cdots \; h_{1s} \; 0 \; \cdots \; 0]^T$ and $\lim\limits_{\tau \to \infty} \Pi_1(\tau) = 0$.

It is possible to prove (inductively) that the solution of the systems (6)–(8) for an arbitrary i and

$$\begin{aligned} h_i(\varepsilon) &= -l(x_i) - l\left(\int_0^{(\cdot)} X(\cdot) X^{-1}(s) f_i(s)\, ds \right) \\ &= h_{i0} + \varepsilon h_{i1} + \cdots + \varepsilon^s h_{is} + O(\varepsilon^q \exp(-\alpha\varepsilon)) \end{aligned}$$

has the form

$$\Pi_i(\tau) = X(\tau) \sum_{j=0}^{s} \varepsilon^j [Q^+ b_i]_{nj} + \int_0^{\tau} X(\tau) X^{-1}(s) f_i(s)\, ds, \tag{31}$$

where $b_i = [h_{i0} \; \cdots \; h_{is} \; 0 \; \cdots \; 0]^T$.

For $\Pi_i(\tau)$ the bound (4) is fulfilled.

So we have proved the following theorem:

Theorem 2.2. *Let $D(\varepsilon) = D_0 + D_1\varepsilon + D_2\varepsilon^2 + \cdots + D_s\varepsilon^s + O(\varepsilon^q \exp(-\alpha\varepsilon))$ and the conditions (H1)–(H3), (H7), (H8) be satisfied. Then the solution of the boundary-value problem (1), (2) has an unique representation in the form (3). The coefficients of the expansion are defined by the equalities (5), (30), (31).*

Remark 3. If rank $Q < (s+1)n$, then we obtain c_i with determination of the boundary function $\Pi_{i+1}(\tau)$.

Remark 4. If $D(\varepsilon) = l(X) = l(e^{A\tau})$

$$= l(E) + \varepsilon^{-1} l\left(A\frac{((\cdot)-a)}{1!}\right) + \varepsilon^{-2} l\left(A^2\frac{((\cdot)-a)^2}{2!}\right) + \cdots$$

$$= D_0 + \varepsilon^{-1} D_{-1} + \varepsilon^{-2} D_{-2} + \cdots,$$

then we seek c_i in the form $c_i = c_{i0} + \varepsilon^{-1} c_{i1} + \cdots$ From the structure of the matrix $X(\tau)$ it follows that $\lim_{\tau\to\infty} \Pi_i(\tau) = 0$.

3. A BOUND OF THE REMAINDER TERM OF THE ASYMPTOTIC SERIES

The solution of the boundary-value problem (1), (2) we seek in the form

$$x(t,\varepsilon) = X_n(t,\varepsilon) + \varepsilon^{n+1}\xi(t,\varepsilon), \tag{32}$$

where

$$X_n(t,\varepsilon) = \sum_{i=0}^{n} [x_i(t) + \Pi_i(\tau)]\varepsilon_i.$$

We shall proof that in [a,b], when $\varepsilon \to 0$, the function $\xi(t,\varepsilon)$ fulfills the inequality $\|\xi(t,\varepsilon)\| \le K$, where K is a positive constant.

We substitute (32) in (1), (2), where $x_i(t)$ and $\Pi_i(\tau)$ are defined in Section 2. After some transformations we obtain that the function $\xi(t,\varepsilon)$ satisfies the boundary-value problem

$$\varepsilon\,\dot{\xi}(t,\varepsilon) = A\xi(t,\varepsilon) + \varepsilon A_1(t)\xi(t,\varepsilon) + H(t,\varepsilon), \;\; l(\xi(\cdot,\varepsilon)) = 0, \tag{33}$$

where

$$H(t,\varepsilon) = \frac{1}{\varepsilon^{n+1}}(H_1(t,\varepsilon) + H_2(t,\varepsilon)),$$

$$H_1(t,\varepsilon) = -\varepsilon^{n+1} A x_{n+1}(t), \;\; H_2(t,\varepsilon) = \varepsilon^{n+1} F_1(t,\varepsilon), \tag{34}$$

$$F_1(t,\varepsilon) = \sum_{k=0}^{n} \frac{1}{(n-k)!} A_1^{(n-k)}(a)\tau^{n-k}\Pi_k(\tau)$$

$$+ \sum_{i=1}^{n} \varepsilon^i \sum_{k=0}^{n-i} \frac{1}{(n-k)!} A_1^{(n-k)}(a)\tau^{n-k}\Pi_{k+i}(\tau)$$

$$+ \frac{1}{(n+1)} A_1^{(n+1)}(a+\theta\tau\varepsilon)\tau^{n+1} \sum_{i=1}^{n+1} \varepsilon^i \Pi_{i-1}(\tau), \quad 0 < \theta < 1.$$

Since $x_i(t)$, $i = 0, 1, \ldots$, are continuous functions in $[a,b]$, then $\|x_i(t)\| \le \eta_i$, where η_i are positive constants.

So we have

$$\left\|\frac{1}{\varepsilon^{n+1}}H_1(t,\varepsilon)\right\| \le \|A\|\,\|x_{n+1}(t)\| \le \|A\|\eta_{n+1}. \tag{35}$$

Let

$$K_1 = \max_{k=\overline{0,n}}\left(\frac{1}{(n-k)!}\|A_1^{(n-k)}(a)\|,\ \frac{1}{(n+1)!}\|A_1^{(n+1)}(a+\theta\tau\varepsilon)\|\right),$$

when $0<\theta<1$ and $t\in[a,b]$, $\|\Pi_i(\tau)\|\le p_i e^{-\alpha_i\tau}$, $p_i>0$, $\alpha_i>0$ $(i=\overline{0,n})$ and $\alpha=\min_i(\alpha_i)$, $p=\max_i(p_i)$.

When $\varepsilon\in(0,\varepsilon_0]$, let denote $c=\max_{i=\overline{0,n+1}}(c_i)$, where $c_0=1$, $c_j=1+\sum_{k=1}^{j}\varepsilon^k$, $j=\overline{1,n}$, $c_{n+1}=\sum_{k=1}^{n+1}\varepsilon^k$.

By (34) we obtain

$$\|F_1(t,\varepsilon)\| \le K_1[c_{n+1}\tau^{n+1}+c_n\tau^n+\cdots+c_1\tau+c_0]\,p\,e^{-\alpha\tau}$$
$$\le K_1\,c\,p[\tau^{n+1}+\cdots+\tau+1]e^{-\alpha\tau}.$$

Let $K_2=K_1cp$. There exists $\overline{\alpha}$, $0<\overline{\alpha}<\alpha$, such that $(\tau^{n+1}+\cdots+\tau+1)e^{-\alpha\tau}\le e^{-\overline{\alpha}\tau}$.

Consequently,

$$\left\|\frac{1}{\varepsilon^{n+1}}H_2(t,\varepsilon)\right\| = \|F_1(t,\varepsilon)\| \le K_2e^{-\overline{\alpha}\tau} \le K_3=\text{const}.$$

Keeping in mind (35) and the last inequality, we have

$$\|H(t,\varepsilon)\| \le \left\|\frac{1}{\varepsilon^{n+1}}H_1(t,\varepsilon)\right\| + \left\|\frac{1}{\varepsilon^{n+1}}H_2(t,\varepsilon)\right\| \le \|A\|\eta_{n+1}+K_3=\eta,$$

that is $\|H(t,\varepsilon)\|\le\eta$, $\eta>0$.

Let $W(t,s,\varepsilon)$ be a fundamental matrix for the homogeneous system

$$\varepsilon\frac{d\xi}{dt}=A\xi,\quad W(t,s,\varepsilon)=E_n,\quad E_n \text{ — } (n\times n)\text{-unit matrix.}$$

Lemma 3.1 [18, 19]. *For the matrix $W(t,s,\varepsilon)$, when $a\le s\le t\le b$, $0<\varepsilon\le\varepsilon_0$ the exponential bound*

$$\|W(t,s,\varepsilon)\| \le \beta\exp\left(-\frac{\alpha(t-s)}{\varepsilon}\right) \tag{36}$$

is fulfilled, where $\alpha>0$, $\beta>0$ are any constants.

Lemma 3.2 [18, 19]. *Any continuous solution of the system* (33) *is a solution of the system of integral equations*

$$\xi(t,\varepsilon) = W(t,a,\varepsilon)\xi(a,\varepsilon) + \int\limits_a^t W(t,s,\varepsilon)\frac{1}{\varepsilon}[\varepsilon A_1(s)\xi(s,\varepsilon) + H(s,\varepsilon)]\,ds, \qquad (37)$$

and conversely.

Lemma 3.3 [18, 19]. *When* $\varepsilon \to 0$, *the integral* $\int\limits_a^t \left\| \frac{1}{\varepsilon} W(t,s,\varepsilon) \right\| ds$ *is uniformly bounded in the segment* $[a,b]$.

Lemma 3.3 reveals that there exists a constant $M > 0$ such that for $\varepsilon \to 0$ and $t \in [a,b]$ the inequality

$$\int\limits_a^t \left\| \frac{1}{\varepsilon} W(t,s,\varepsilon) \right\| ds \leq M$$

holds.

The system (37) will be solved by the method of successive approximations. Let

$$\xi_0(t,\varepsilon) = 0,$$

$$\xi_j(t,\varepsilon) = F(t,\varepsilon) + \int\limits_a^t W(t,s,\varepsilon)\frac{1}{\varepsilon}[\varepsilon A_1(s)\xi_{j-1}(s,\varepsilon) + H(s,\varepsilon)]\,ds \qquad (38)$$

be the Picard successive approximations, where $F(t,\varepsilon) = W(t,a,\varepsilon)\xi(a,\varepsilon)$.

Theorem 3.1. *Let the conditions of Theorem* 2.1 *(or Theorem* 2.2*) be fulfilled. Let* $\overline{\beta}$, h, h_1, h_2, h_3, h_4, ε_0 *be positive constants such that*

$$\|W(t,a,\varepsilon)\| \leq \overline{\beta}; \quad \|F(t,\varepsilon)\| \leq h_1, \quad \textit{where } h_1 = 2\overline{\beta}h,\ 0 < 2\overline{\beta} < 1;$$

$$\|A_1(t)\| \leq h_2, \quad \textit{where } t \in [a,b]; \quad \left\| \overline{\overline{D}}_0^{\,+} + \right\| \leq h_3;$$

$$\|l(\psi)\| \leq h_4\|\psi\|, \quad h_3h_4 < 2, \quad \varepsilon_0 \leq \frac{1}{2Mh_2}.$$

If $\dfrac{M\eta}{1-2\overline{\beta}} \leq h \leq \dfrac{2-(1-2\overline{\beta})h_3h_4}{h_3h_4 2\overline{\beta}} h_2$, *then the asymptotic solution of the boundary-value problem* (1), (2) *has the representation* (32), *where* $\xi(t,\varepsilon)$ *satisfies the inequality* $\|\xi(t,\varepsilon)\| \leq 2h$. *The vector* $\xi(a,\varepsilon)$ *is defined by the algebraic system* $\overline{\overline{D}}(\varepsilon)\xi(a,\varepsilon) = g(\varepsilon)$, *where* $\overline{\overline{D}}(\varepsilon) = l(W(\cdot,a,\varepsilon))$ *is an* $(m \times n)$*-matrix,*

$$g(\varepsilon) = -l\left(\int\limits_a^{(\cdot)} W(\cdot,s,\varepsilon)\frac{1}{\varepsilon}[\varepsilon A_1(s)\xi(s,\varepsilon) + H(s,\varepsilon)]\,ds\right). \qquad (39)$$

Besides, $x(t,\varepsilon)$ approaches the degenerating system at $\varepsilon \to 0$ and $t \in (a,b]$.

Proof. Using (38), we shall prove that the system (37) has an unique continuous solution, which does not leave the domain

$$\Omega = \{(t,\xi) \mid a \le t \le b,\ \|\xi\| \le 2h\},$$

depending on an arbitrary vector $\xi(a,\varepsilon)$.

By the equalities (38), for the first approximation we have

$$\|\xi_1 - \xi_0\| \le \|F(t,\varepsilon)\| + \int_a^t \left\|W(t,s,\varepsilon)\frac{1}{\varepsilon}\right\| \|H(t,s)\|\, ds \le h_1 + M\eta \le h.$$

If $0 < \varepsilon \le \varepsilon_0$ and $\varepsilon_0 \le \dfrac{1}{2Mh_2}$, we obtain

$$\|\xi_j - \xi_{j-1}\| \le \varepsilon \int_a^t \left\|W(t,s,\varepsilon)\frac{1}{\varepsilon}\right\| ds\, \|A_1(t)\|\, \|\xi_{j-1}(t,\varepsilon) - \xi_{j-2}(t,\varepsilon)\|$$

$$\le \varepsilon M h_2\, \|\xi_{j-1}(t,\varepsilon) - \xi_{j-2}(t,\varepsilon)\| \le \frac{1}{2}\, \|\xi_{j-1} - \xi_{j-2}\|,\ j = 2,3,\ldots$$

This reveals that in the segment $[a,b]$, when ε is sufficiently small, the successive approximations (38) are absolutely and uniformly convergent. We shall show that the successive approximations do not leave the domain Ω. We have

$$\|\xi_k(t,\varepsilon)\| \le \sum_{j=1}^{k} \|\xi_j(t,\varepsilon) - \xi_{j-1}(t,\varepsilon)\| \le h + \frac{h}{2} + \frac{h}{2^2} + \cdots + \frac{h}{2^{n-1}} \le 2h.$$

Let $\lim\limits_{k\to\infty} \xi_k(t,\varepsilon) = \xi(t,\varepsilon)$ satisfy (37) identically. Then in the interval $[a,b]$ for $\varepsilon \to 0$ the inequality $\|\xi(t,\varepsilon)\| \le 2h$ is fulfilled.

Consequently, the system (37) has an unique continuos solution, which does not leave the domain Ω and depends on an arbitrary vector $\xi(a,\varepsilon)$.

We define $\xi(a,\varepsilon)$ by the algebraic system

$$\overline{\overline{\overline{D}}}(\varepsilon)\xi(a,\varepsilon) = g(\varepsilon), \tag{40}$$

where $\overline{\overline{\overline{D}}}(\varepsilon)$ and $g(\varepsilon)$ are the expressions from (39). The system (40) is obtained substituting $\xi(t,\varepsilon)$ in the boundary condition $l(\xi) = 0$ of (33).

Let $\overline{\overline{\overline{D}}}(\varepsilon) = \overline{\overline{\overline{D}}}_0 + O\left(\varepsilon^s \exp\left(-\dfrac{\gamma}{\varepsilon}\right)\right)$, $\gamma > 0$, $s \in \mathbb{N}$, where $\overline{\overline{\overline{D}}}_0$ is $(m \times n)$-constant matrix. Then if $\operatorname{rank}\overline{\overline{\overline{D}}}_0 = n$, for $\varepsilon \in (0,\varepsilon_0]$ the system (40) has an unique solution

$$\xi(a,\varepsilon) = \overline{\overline{\overline{D}}}_0^{+} g(\varepsilon)$$

if and only if

$$P_3^* g(\varepsilon) = 0 \quad \text{and} \quad P_3^* : R^n \to \ker(\overline{\overline{D}}_0^{\,*}).$$

The inequality $\|\xi(a,\varepsilon)\| \le 2h$ is fulfilled for $\xi(a,\varepsilon)$. Really,

$$\|\xi(a,\varepsilon)\| = \|\overline{\overline{D}}_0^{\,+}\|\, \|g(\varepsilon)\|$$

$$\le h_3h_4 \int_a^t \left\| W(t,s,\varepsilon)\frac{1}{\varepsilon} \right\| \left[\varepsilon \|A_1(s)\|\, \|\xi(s,\varepsilon)\| + \|H(s,\varepsilon)\|\right] ds$$

$$\le h_3h_4M(2\varepsilon h_1 h + \eta) \le h_3h_4M\left(2\frac{1}{2Mh_2}2\overline{\beta}h^2 + \frac{h(1-2\overline{\beta})}{M}\right)$$

$$\le hh_3h_4\left(2\overline{\beta}\frac{h}{h_2} + 1 - 2\overline{\beta}\right)$$

$$\le hh_3h_4\left(\frac{2\overline{\beta}}{h_2}\frac{2-(1-2\overline{\beta})h_3h_4}{h_3h_4 2\overline{\beta}}h_2 + 1 - 2\overline{\beta}\right) = 2h. \quad \square$$

4. EXAMPLE

We consider the two-point boundary-value problem

$$\varepsilon\dot{x} = Ax + \varphi(t), \quad t \in [0,1], \quad l(x) = Mx(0) + Nx(1) = h,$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad A = \begin{bmatrix} -3 & 4 \\ -1 & 1 \end{bmatrix}, \quad \varphi(t) = \begin{bmatrix} t-1 \\ t \end{bmatrix},$$

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 1 & 0 \\ 0 & 10 \\ 6 & 0 \end{bmatrix}, \quad h = \begin{bmatrix} 6 \\ 31 \\ 25 \end{bmatrix}.$$

If $\varepsilon = 0$, then $x_0(t) = -A^{-1}\varphi(t) = \begin{bmatrix} 3t+1 \\ 2t+1 \end{bmatrix}$. It is obvious that $l(x_0) = [5 \;\; 31 \;\; 25]^T \ne h$. Since $\lambda_{1,2} = -1$ and the normal fundamental matrix has the form $X(t) = \begin{bmatrix} 1-2t & 4t \\ -t & 1+2t \end{bmatrix} e^{-t}$, then $D(\varepsilon) = MX(0) + NX\left(\frac{1}{\varepsilon}\right)$ has the representation

$$D(\varepsilon) = M + N\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} e^{-\frac{1}{\varepsilon}} + N\begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix} \frac{1}{\varepsilon}e^{-\frac{1}{\varepsilon}}$$

$$= D_0 + O\left(e^{-\frac{1}{\varepsilon}} + \frac{1}{\varepsilon}e^{-\frac{1}{\varepsilon}}\right),$$

where $D_0 = M$ and $\operatorname{rank} D_0 = 2$.

We obtain sequentially

$$D_0^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad P_1 = \frac{1}{2}[0\ 1\ -1], \quad h_0 = [1\ 0\ 0]^T, \quad h_1 = [10\ 33\ 33]^T.$$

In this case the conditions $P_1^* h_i = 0$, $i = 0, 1, \ldots$, are fulfilled.

According to Corollary 1, we obtain

$$c_0 = D_0^+ h_0 = [1\ 0]^T, \quad c_1 = D_0^+ h_1 = [10\ 33]^T,$$

$$\Pi_0(\tau) = X(\tau)c_0 = \begin{bmatrix} 1-2\tau \\ \tau \end{bmatrix} e^{-\tau}, \quad \Pi_1(\tau)c_1 = X(\tau)c_1 = \begin{bmatrix} 10+112\tau \\ 33-23\tau \end{bmatrix} e^{-\tau}.$$

The asymptotic solution of the two-point boundary-value problem has the form

$$x(t,\varepsilon) = \begin{bmatrix} -1 & 4 \\ -1 & 3 \end{bmatrix} \cdot \begin{bmatrix} t-1 \\ t \end{bmatrix} + \begin{bmatrix} 1 - \frac{2t}{\varepsilon} \\ \frac{t}{\varepsilon} \end{bmatrix} e^{-\frac{t}{\varepsilon}}$$

$$+ \varepsilon \left(\begin{bmatrix} -5 \\ -3 \end{bmatrix} + \begin{bmatrix} 10 + 112\frac{t}{\varepsilon} \\ 33 - 23\frac{t}{\varepsilon} \end{bmatrix} e^{-\frac{t}{\varepsilon}} \right) + O(\varepsilon^2).$$

REFERENCES

1. Tikhonov, A. N. On dependence of the solution of differential equations on small parameter. *Mat. sb.*, **22**, 2, 1948, 193–204 (in Russian).
2. Tikhonov, A. N. System of differential equations with small parameters of the derivatives. *Mat. sb.*, **31**, 3, 1952, 575–586 (in Russian).
3. Wasow, W. Asymptotic expansions for ordinary differential equation. John Wiley, New York, 1965.
4. Vasil'eva, A. B. On the differential equations with small parameters of the derivatives. *Mat. sb.*, **31**, 3, 1952, 587–644 (in Russian).
5. Vasil'eva, A. B. Asymptotic behavior of solutions of certain problems for ordinary nonlinear differential equations with small parameters of the higher derivative. *UMN*, **18**, 3, 1963, 15–86 (in Russian).
6. Vasil'eva, A. B., V. F. Butuzov. Asymptotic expansions of solution of singularly perturbed equations. Nauka, M., 1973 (in Russian).
7. Vasil'eva, A. B., V. F. Butuzov. Singularly perturbed equations in the critical case. Moscow State University, 1978 (in Russian).
8. Mishchenko, E. F., N. H. Rozov. Differential equations with small parameter and relaxational oscillations. Nauka, M., 1975 (in Russian).

9. Mishchenko, E. F., U. S. Kolesov, A. U. Kolesov, N. H. Rozov. Periodic movements and bifurcation process at singularly perturbed systems. Physical–mathematical literature, M., 1995 (in Russian).

10. Lomov, S. A. Introduction in the general theory of singular perturbations. Nauka, M., 1981 (in Russian).

11. Feshchenko, S. F., M. I. Shkil', L. D. Nicolenko. Asymptotic methods in the theory of linear differential equations. Nauk. Dumka, Kiev, 1966 (in Russian).

12. Shkil', M. I. Asymptotic methods in differential equations. Visha Shkola, Kiev, 1971 (in Ukrainian).

13. Boichuk, A. A., V. F. Zhjuravliov, A. M. Samoilenko. Generalized inverse operators and Noether's boundary-value problems. IM NAN Ukraina, Kiev, 1995 (in Russian).

14. Karandjulov, L. I., A. A. Boichuk, V. A. Bozhko. Asymptotic expansions of solution of singularly perturbed linear boundary-value problem. *Dokl. AN Ukraina*, **1**, 1994, 7–10 (in Russian).

15. Penrose, R. A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.*, **51**, 1955, 406–413.

16. Penrose, R. On best approximate solution of linear matrix equations. *Proc. Cambridge Philos.Soc.*, **52**, 1956, 17–19.

17. Generalized invers and applications. M. Z. Nashed, ed., Acad. Press, New York, San Francisco, London, 1967.

18. Haber, S., N. Levinson. A boundary-value problem for a singularly perturbed differential equation. *Proc. Amer. Math. Soc.*, **6**, 1955, 866–872.

19. Levinson, N. A. A boundary-value problem for a singularly perturbed differential equation. *Duke math. journ.*, **2**, 1958, 331–342.

Received March 7, 1998
Revised November 19, 1998

Ljudmil Ivanov Karandjulov
Technical University-Sofia
Institute of Applied Math. and Informatics
P.O. Box 384, Sofia-1000
E-mail: `likar@vmei.acad.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

ABOUT THE FIRST CROSSING OF THE POISSON PROCESS WITH A CURVED UPPER BOUNDARY

AVGUSTIN MARINOV, JIVKO JELIAZKOV, TZVETAN IGNATOV*

The paper is concerned with the distribution of the first crossing of a simple Poisson process trajectory with an upper boundary. Exact formula is derived when the upper boundary has a vertical asymptote.

Keywords: risk theory, Poisson process, first crossing time, upper boundary

1991/95 Math. Subject Classification: primary 60J75, secondary 60G40

1. INTRODUCTION

Many problems in risk, queuing and storage theories can be reduced to the study of the first crossing time or level of a given boundary with a trajectory of a certain stochastic process. Such problems have been mainly investigated for continuous time Gaussian or similar to Gaussian processes. In case of non-linear boundaries and of discrete-state space, the literature is rather sparse, and it treats only the ordinary or compound Poisson process. The reader is referred to Lundberg (1903), Cramér (1955), Whittle (1961), Daniels (1963), Gallot (1966, 1993), Zacks (1991), Stadje (1994), Schäl (1993), Picard and Lefèvre (1997), Kalashnikov (1996).

In the present work, the interest will be focused on the classical continuous time model of an insurance company, i.e. the Poisson model.

Suppose that $\xi_1, \xi_2, \xi_3, \dots$ are independent and exponentially distributed with

* The author was partially supported by the SRF of the Sofia University "St. Kliment Ohridski" under Contract No 221/1998.

parameter λ, so that

$$k(t) = \max\{n : \xi_1 + \xi_2 + \cdots + \xi_n \leq t\} \tag{1}$$

defines an ordinary Poisson process $k(t)$, $t \geq 0$. We shall interpret

$$S_n := \xi_1 + \xi_2 + \cdots + \xi_n \tag{2}$$

as the moment of the n-th insurance claim. If η_n represents the amount of the n-th claim, then

$$Z_t := \sum_{i \leq k(t)} \eta_i \tag{3}$$

represents the total amount of claims to time t. The stochastic process Z_t, $t \geq 0$, coincides with $k(t)$, $t \geq 0$, when $\eta_i \equiv 1$, $i = 1, 2, \ldots$

Let

$$U_t = f(t) - Z_t, \tag{4}$$

where $f(t)$ is a non-decreasing real function defined on the set $\mathbb{R}_+ = \{x : x \geq 0\}$. In the classical risk model, usually the function $f(t)$ has the form

$$f(t) = u + ct,$$

where c is the premium income per unit time, and $u := f(0)$ is the initial surplus.

Define the ruin time T as

$$T := \inf\{t : U_t \leq 0,\ t > 0\}, \tag{5}$$

i.e. T is the time of the first crossing of the trajectory $t \to Z_t$ with the boundary $t \to f(t)$ (disregarding the origin when $f(0) = 0$).

Recently, Picard and Lefèvre (1997) have investigated the compound Poisson risk model when the integer valued random variables η_1, η_2, ... are independent identically distributed and the sequences ξ_1, ξ_2, ... and η_1, η_2, ... are independent. They derived the expressing for the ruin probability $\mathrm{P}(T \leq x)$ in terms of generalized Appell's polynomials under the assumption

$$\mathrm{P}(\eta_i \geq 1) = 1 \quad \text{and} \quad \lim_{t \to \infty} f(t) = +\infty.$$

Our purpose is to find the ruin probability $\mathrm{P}(T \leq x)$ in the particular case when $f(t)$ has a vertical asymptote (i.e. $\lim_{t \uparrow v} f(t) = \infty$ for some $v > 0$) and $\eta_i \equiv 1$.

2. THE PROBABILITY OF RUIN IN FINITE TIME

It is worth noting that the distribution of T is defective ($\mathrm{P}(T = \infty) > 0$). Further we shall use the quantities

$$v_n : f^{-1}(n), \quad n = 0, 1, 2, \ldots, \tag{6}$$

where the inverse function $f^{-1}(x)$ is defined by

$$f^{-1}(x) := \inf\{y : f(y) \geq x\}.$$

Obviously, $v_0 = 0 \le v_1 \le v_2 \le \dots$ and

$$\lim_{n\to\infty} v_n = v. \tag{7}$$

We shall use the formula for the non-ruin probability $\mathrm{P}(T > x)$ derived by Ignatov and Kaishev (1997) in the form

$$\mathrm{P}(T > x) = \sum_{n\ge 0} e^{-x}\left[\sum_{i=0}^{n}(-1)^i\delta(v_1,\dots,v_i)\sum_{j=0}^{n-i}\frac{x^j}{j!}\right].I_{[v_n,v_{n+1})}(x), \tag{8}$$

where $\delta(v_1,\dots,v_i) = 1$ for $i = 0$ and

$$\delta(v_1,\dots,v_i) = \det\begin{pmatrix} v_1 & 1 & 0 & \dots & 0 & 0 \\ \frac{v_2^2}{2!} & v_2 & 1 & \dots & 0 & 0 \\ \frac{v_3^3}{3!} & \frac{v_3^2}{2!} & v_3 & \dots & 0 & 0 \\ \dots\dots & \dots\dots & \dots\dots & \dots & & \\ \frac{v_{i-1}^{i-1}}{(i-1)!} & \frac{v_{i-1}^{i-2}}{(i-2)!} & \frac{v_{i-1}^{i-3}}{(i-3)!} & \dots & v_{i-1} & 1 \\ \frac{v_i^i}{i!} & \frac{v_i^{i-1}}{(i-1)!} & \frac{v_i^{i-2}}{(i-2)!} & \dots & \frac{v_i^2}{2!} & v_i \end{pmatrix} \tag{9}$$

for $i = 2, 3, \dots,$ and $I_{[v_n,v_{n+1})}(x)$ is the indicator function of the interval $[v_n, v_{n+1})$.

The formula (8) is obtained under the assumptions $1 \equiv \eta_1 \equiv \eta_2 \equiv \cdots$ and the parameter $\lambda \equiv 1$ of the sequence $\xi_1, \xi_2, \dots$ We shall assume now that the last assumptions are fulfilled.

The main result in this section is the following

Theorem. *If $f(t)$ is such that $v_n \uparrow v$, then*

$$\mathrm{P}(T > x) = \begin{cases} \sum\limits_{i=0}^{\infty}(-1)^i\delta(v_1,\dots,v_i), & x \ge v, \\ \\ \sum\limits_{n\ge 0} e^{-x}\left[\sum\limits_{i=0}^{n}(-1)^i\delta(v_1,\dots,v_i)\sum\limits_{j=0}^{n-i}\frac{x^j}{j!}\right].I_{[v_n,v_{n+1})}(x), & 0 \le x < v. \end{cases} \tag{10}$$

Proof. To prove the theorem, we shall use the next two lemmas.

Lemma 1. *For the determinants $\delta(v_1,\dots,v_n)$ we have the identities*

$$\delta(cv_1,\dots,cv_n) = c^n\delta(v_1,\dots,v_n), \tag{11}$$

$$\delta(v_1 + c,\dots,v_n + c) = \delta(v_1,\dots,v_n) - \delta(v_1,\dots,v_{n-1},-c), \tag{12}$$

$$\delta(v_1,\dots,v_n) = (-1)^{n-1}\delta(v_n - v_1,\dots,v_n - v_{n-1},v_n). \tag{13}$$

Proof. The identity (11) follows immediately from the definition of a determinant as a sum of certain products, in our case of the form $(-1)^h v_1^{j_1} v_2^{j_2} \dots v_n^{j_n}$, where h is a suitable integer and $j_1 + j_2 + \cdots + j_n = n$.

Let us introduce the matrix

$$\Delta(v_1+c,\dots,v_n+c) := \begin{pmatrix} \frac{(v_1+c)^1}{1!} & 1 & 0 & \dots & 0 \\ \frac{(v_2+c)^2}{2!} & \frac{(v_2+c)^1}{1!} & 1 & \dots & 0 \\ \frac{(v_3+c)^3}{3!} & \frac{(v_3+c)^2}{2!} & \frac{(v_3+c)^1}{1!} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \frac{(v_n+c)^n}{n!} & \frac{(v_n+c)^{n-1}}{(n-1)!} & \frac{(v_n+c)^{n-2}}{(n-2)!} & \dots & \frac{(v_n+c)^1}{1!} \end{pmatrix},$$

then $\det(\Delta(v_1+c,\dots,v_n+c)) = \delta(v_1+c,\dots,v_n+c)$.

If we add elements of the $(j+1)$-th column multiplied by $\frac{(-c)^j}{j!}$ to the elements of the first column for $j = 1, \dots, n-1$, we get

$$\delta(v_1+c,\dots,v_n+c) = \det\begin{pmatrix} \frac{v_1^1}{1!} & 1 & \dots & 0 \\ \frac{v_2^2}{2!} & \frac{v_2+c}{1!} & \dots & 0 \\ \dots & \dots & \dots & \\ \frac{v_{n-1}^{n-1}}{(n-1)!} & \frac{(v_{n-1}+c)^{n-2}}{(n-2)!} & \dots & 1 \\ \frac{v_n^n}{n!} - \frac{(-c)^n}{n!} & \frac{(v_n+c)^1}{(n-1)!} & \dots & \frac{(v_n+c)^1}{1!} \end{pmatrix}. \tag{14}$$

Indeed, for the element in the first column and the i-th row we have

$$\frac{(v_i+c)^i}{i!} + \frac{(-c)^1(v_i+c)^{i-1}}{1!(i-1)!} + \frac{(-c)^2(v_i+c)^{i-2}}{2!(i-2)!} + \cdots + \frac{(-c)^i(v_i+c)^0}{i!0!}$$

$$\equiv \frac{1}{i!}\left(\frac{i!}{0!i!}(v_i+c)^0(v_i+c)^i + \frac{i!}{1!(i-1)!}(-c)^1(v_i+c)^{i-1}\right.$$

$$\left. + \frac{i!}{2!(i-2)!}(-c)^2(v_i+c)^{i-2} + \cdots + \frac{i!}{i!0!}(-c)^i(v_i+c)^0\right)$$

$$\equiv \frac{1}{i!}(-c+v_i+c)^i = \frac{v_i^i}{i!} \tag{15}$$

for $i = 1, \dots, n-1$.

For $i = n$ it is easy to find the identity

$$\frac{(v_n+c)^n}{n!} + \frac{(-c)^1(v_n+c)^{n-1}}{1!(n-1)!} + \frac{(-c)^2(v_n+c)^{n-2}}{2!(n-2)!} + \cdots + \frac{(-c)^{n-1}(v_n+c)^1}{(n-1)!1!}$$
$$\equiv \frac{v_n^n}{n!} - \frac{(-c)^n}{n!}.$$

A similar construction will be used to the elements of the second column and so on.

Finally, we get

$$\delta(v_1+c,\ldots,v_n+c)$$
$$= \det\begin{pmatrix} \frac{v_1^1}{1!} & 1 & \ldots & 0 \\ \frac{v_2^2}{2!} & \frac{v_2^1}{1!} & \ldots & 0 \\ \hdotsfor{3} & \\ \frac{v_{n-1}^{n-1}}{(n-1)!} & \frac{v_{n-1}^{n-2}}{(n-2)!} & \ldots & 1 \\ \frac{v_n^n}{n!} - \frac{(-c)^n}{n!} & \frac{v_n^{n-1}}{(n-1)!} - \frac{(-c)^{n-1}}{(n-1)!} & \ldots & \frac{v_n^1}{1!} - \frac{(-c)^1}{1!} \end{pmatrix}. \tag{16}$$

From the well-known property of determinants and the form of the elements of the last row in (16) we obtain the identity (12).

The identity (13) follows from (11) and (12). Indeed, using (12) with $c = v_n$, we have

$$\delta(v_n - v_1,\ldots,v_n - v_{n-1}, v_n + 0) = \delta(-v_1,\ldots,-v_{n-1},0) - \delta(-v_1,\ldots,-v_{n-1},-v_n)$$
$$= -\delta(-v_1,\ldots,-v_{n-1},-v_n) = (-1)^{n+1}\delta(v_1,\ldots,v_n).$$

In the second equality we use the fact that $\delta(-v_1,\ldots,-v_{n-1},0) \equiv 0$. In the third equality we have used (11).

The proof of Lemma 1 is complete.

Lemma 2. *If $v_n \uparrow v$, then the series $\sum\limits_{n=0}^{\infty} |\delta(v_1,\ldots,v_n)|$ are convergent, i.e.*

$$\sum_{n=0}^{\infty} |\delta(v_1,\ldots,v_n)| < +\infty. \tag{17}$$

Proof. Let s be chosen such that

$$0 \le v - v_{s+n} \le \frac{1}{3} \tag{18}$$

for each $n \ge 0$. We shall use the Laplace's expansion of a determinant and for this purpose we introduce the notation $\delta_{j_1,\ldots,j_s}^{r_1,\ldots,r_s}(v_1,\ldots,v_s)$ for the determinant formed from $\delta(v_1,\ldots,v_s)$ by using the elements in the rows $r_1, \ldots, r_s$ and the columns

$j_1, \dots, j_s$. Let us expand the determinant $\delta(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n})$ applying the Laplace's formula

$$\begin{aligned} &\delta(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &= \sum_{(j_1,\dots,j_s)\in C_s^{s+n}} (-1)^h \delta^{1,\dots,s}_{j_1,\dots,j_s}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &\qquad \times \delta^{s+1,\dots,s+n}_{j_{s+1},\dots,j_{s+n}}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}), \end{aligned} \tag{19}$$

where the summation is taken over the set C_s^{n+s}, the set of $\binom{n+s}{s}$ subsets $(j_1, \dots, j_s)$ from the integers $(1, 2, \dots, s+n)$. It is easy to find that

$$\delta^{1,2,\dots,s}_{j_1,\dots,j_s}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) = 0$$

when at least one of the indices $j_1, \dots, j_s$ is greater than $s+1$. Therefore in this case we obtain

$$\begin{aligned} &\delta(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &= \sum_{(j_1,\dots,j_s)\in C_s^{s+1}} (-1)^h \delta^{1,\dots,s}_{j_1,\dots,j_s}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &\qquad \times \delta^{s+1,\dots,s+n}_{j_{s+1},\dots,j_{s+n}}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}). \end{aligned} \tag{20}$$

For the sake of simplicity we shall denote the subset $(j_1, \dots, j_s) \in C_s^{s+1}$ by $(1, 2, \dots, i-1, \hat{i}, i+1, \dots, s+1)$ if $j_k \neq i$ for $k = 1, \dots, s$, and also we shall choose $j_{s+1} = i$, $j_{s+2} = s+2, \dots, j_{s+n} = s+n$.

Now we can rewrite (20) as

$$\begin{aligned} &\delta(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &= \sum_{i=1}^{s+1} (-1)^h \delta^{1,2,\dots,s}_{1,\dots,i-1,\hat{i},i+1,\dots,s+1}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}) \\ &\qquad \times \delta^{s+1,\dots,s+n}_{i,s+2,\dots,s+n}(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n}). \end{aligned} \tag{21}$$

To prove the inequality (17), it is enough to prove that

$$\sum_{n=0}^{\infty} |\delta(v_1, \dots, v_{s+n})| < \infty \tag{22}$$

for some positive integer s.

From (19) and (21) we get

$$\sum_{n=0}^{\infty} |\delta(v_1, \dots, v_{s+n})| = \sum_{n=0}^{\infty} |\delta(v_{s+n} - v_1, \dots, v_{s+n} - v_{s+n-1}, v_{s+n})|$$

$$\leq \sum_{n=0}^{\infty}\sum_{i=1}^{s+1}\left|\delta^{1,2,\ldots,s}_{1,\ldots,i-1,\hat{i},i+1,\ldots,s+1}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|$$

$$\times\left|\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|$$

$$=\sum_{i=1}^{s+1}\sum_{n=0}^{\infty}\left|\delta^{1,2,\ldots,s}_{1,\ldots,i-1,\hat{i},i+1,\ldots,s+1}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|$$

$$\times\left|\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|. \tag{23}$$

Let us recall the Hadamard's inequality for a determinant of order n and value D with real or complex elements a_{ij}:

$$|D|^2 \leq \prod_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|^2\right), \tag{24}$$

and obviously,

$$|D| \leq \prod_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|\right). \tag{25}$$

From (25) and $v_{s+n}-v_1 \leq v$, $\ldots$, $v_{s+n}-v_{s+n-1} \leq v$, $v_{s+n} \leq v$ for the determinant $\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})$ we have

$$\left|\delta^{1,2,\ldots,s}_{1,\ldots,i-1,\hat{i},i+1,\ldots,s+1}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right| \leq e^{sv} \tag{26}$$

for $i=1,\ldots,s+1$.

For the determinant $\left|\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|$ it is possible to prove that for each constant c

$$\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(c(v_{s+n}-v_1),\ldots,c(v_{s+n}-v_{s+n-1}),cv_{s+n})$$

$$=c^{n+s-i+1}\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n}). \tag{27}$$

The determinant $\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})$ depends only on $v_{s+n}-v_{s+1},\ldots,v_{s+n}-v_{s+n-1},v_{s+n}$. Consequently, using inequalities (21), we have

$$v_{s+n}-v_{s+i} \leq v-v_{s+i} \leq \frac{1}{3}. \tag{28}$$

From (18), (27) and (28) we obtain

$$\left|\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(v_{s+n}-v_1,\ldots,v_{s+n}-v_{s+n-1},v_{s+n})\right|$$

$$=\left(\frac{1}{3}\right)^{n+s-i+1}\delta^{s+1,\ldots,s+n}_{i,s+2,\ldots,s+n}(3(v_{s+n}-v_1),\ldots,3(v_{s+n}-v_{s+n-1}),3v_{s+n})$$

$$<\left(\frac{1}{3}\right)^{n+s-i+1}e^{n-i}e^{3v}. \tag{29}$$

Replacing the determinants in (23) with the upper bounds in (26) and (29), we get

$$\sum_{i=1}^{s+1}\sum_{n=0}^{\infty} e^{sv}\left(\frac{1}{3}\right)^{n+s-i+1} e^{n-1}e^{3v} = e^{(s+3)v}\sum_{i=1}^{s+1}\left(\frac{1}{3}\right)^{s-i+2}\sum_{n=0}^{\infty}\left(\frac{e}{3}\right)^{n-1}$$

$$= e^{(s+3)v}\left(\frac{1}{3}\right)^{s+2}\sum_{i=1}^{s+1}\left(\frac{1}{3}\right)^{-i}\left(\frac{3}{e}\right)\frac{1}{1-\frac{e}{3}} < \infty.$$

The proof of Lemma 2 is complete.

Proof of the theorem. The expression for the probability $\mathrm{P}(T > x)$ when $0 \le x < v$ follows immediately from formula (8). Indeed, the probability $\mathrm{P}(T > x)$ depends only on the form of the upper boundary $f(t)$ in the interval $[0, x)$, therefore we can imagine that the condition $\lim_{t\to\infty} f(t) = \infty$ is true and in this case we can use the formula (8). It is clear that

$$\mathrm{P}(T \ge x) = \mathrm{P}(T \ge v) = \lim_{y\uparrow v}\mathrm{P}(T \ge y) \quad \text{for } x \ge v.$$

Now we can see that $\lim_{y\uparrow v}\mathrm{P}(T \ge x) = \sum_{i=0}^{\infty}(-1)^i\delta(v_1,\dots,v_i)$. The formula (8) may be expressed as

$$\mathrm{P}(T \ge x) = \sum_{n=0}^{\infty}\left(\sum_{i=0}^{n}(-1)^i\delta(v_1,\dots,v_i)e^{-x}\sum_{j=0}^{n-i}\frac{x^j}{j!}\right).I_{[v_n,v_{n+1})}(x)$$

$$= \sum_{i=0}^{\infty}(-1)^i\delta(v_1,\dots,v_i).g_i(x), \tag{30}$$

where $g_0(x) = 1$ for $x \in [0, v)$ and for $i \ge 1$

$$g_i(x) = \begin{cases} 0, & x \in [0, v_n), \\ e^{-x}\sum_{j=0}^{n-i}\frac{x^j}{j!}, & x \in [v_n, v_{n+1}). \end{cases}$$

When $x \to v$, we have $n \to \infty$, so we get

$$\lim_{x\uparrow v} g_i(x) = 1, \quad i = 0, 1, \dots \tag{31}$$

Since $|g_i(x)| \le 1$, $x \in [0, v)$, we have

$$\sum_{i=0}^{\infty}|(-1)^i\delta(v_1,\dots,v_i)g_i(x)| \le \sum_{i=0}^{\infty}|\delta(v_1,\dots,v_i)|. \tag{32}$$

Combining (32) and Lemma 2, we get that the series $\sum_{i=0}^{\infty}(-1)^i\delta(v_1,\ldots,v_i)g_i(x)$ is uniformly convergent. Consequently, taking into account (31), we have

$$\lim_{x\uparrow v}\sum_{i=0}^{\infty}(-1)^i\delta(v_1,\ldots,v_i)g_i(x)=\sum_{i=0}^{\infty}(-1)^i\delta(v_1,\ldots,v_i)\lim_{x\uparrow v}g_i(x)$$

$$=\sum_{i=0}^{\infty}(-1)^i\delta(v_1,\ldots,v_i).$$

The proof of the theorem is complete.

REFERENCES

1. Lundberg, F. I. Approximerad fromställning af sannolifhetstunktionen, II. Återförsäkring av kollektivisker. Almquist & Wiksell, Uppsala, 1903.
2. Cramér, H. Collective risk theory. Jubilee volume of Försäkringsaktieboleget Skandia, Stockholm, 1955.
3. Whittle, P. Some exact results for one-sided distribution tests of the Kolmogorov-Smirnov type. *Ann. Math. Statist.*, **32**, 1961, 499–505.
4. Daniels, H. E. The Poisson process with a curved absorbing boundary. *Bull. Inter. Statist. Inst.*, 34th session, **40**, *1963, 994–1008.*
5. Gallot, S. F. L. Asymptotic absorption probabilities for a Poisson process. *J. Appl. Prob.*, **3**, 1966, 445–452.
6. Gallot, S. F. L. Absorption and first passage times for a compound Poisson process in a general upper boundary. *J. Appl. Prob.*, **30**, 1993, 835–850.
7. Thorin, O. Probabilities of ruin. *Scand. Actuarial J.*, 1982, 65–102.
8. Zacks, S. Distributions of stopping times for Poisson process with linear boundaries. *Commum. Statist. Stoch. Models*, **7**, 1991, 233–242.
9. Stadje, W. Distribution of first-exit times for empirical counting and Poisson processes with moving boundaries. *Commun. Statist. Stoch. Models*, **9**, 1991, 91–103.
10. Schäl, M. On hitting times for jump-diffusion process with past depend local characteristics. *Stoch. Proc. Appl.*, **47**, 1993, 131–142.
11. Picard, Ph., C. Lefèvre. The Probability of Ruin in Finite Time with Discrete Claim Size Distribution. *Scand. Actuarial J.*, **1**, 1997, 58–69.
12. Picard, Ph., C. Lefèvre. First Crossing of Basic Counting Processes with Lower Non-Linear Boundaries. A Unified Approach through Pseudo-Polynomials (I). *Adv. Appl. Prob.*, **28**, 1996, 853–876.
13. Kalashnikov, V. V. Two-Sided Bounds of Ruin Probabilities. *Scand. Actuarial J.*, **1**, 1996, 1–18.
14. Ignatov, Tz., V. Kaishev. Two-Sided Bounds for the Finite Time Probability of Ruin. *Scand. Actuarial J.*, 1998 (submitted).

Received March 9, 1998
Revised April 13, 1998

Faculty of Mathematics and Informatics
Sofia University
5 Blvd James Bourchier
BG-1164 Sofia, Bulgaria
E-mail address: ignatov@feb.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

AN ESTIMATE OF THE PERIOD OF PERIODICAL SOLUTION OF AN AUTONOMOUS SYSTEM OF DIFFERENTIAL EQUATIONS

ANA D. MIHAILOVA, NEKO N. GEORGIEV

In this paper we obtain an estimate for the period of the periodical solution of an autonomous system of differential equations from above.

Keywords: autonomous systems of differential equations, periodical solution, estimates for the period

1991/1995 Math. Subject Classification: 34C25

We consider the autonomous system

$$\left|\begin{aligned} \dot{x} &= a\left[f\left(x\right)-\left(1+b\right)x-z\right], \\ \dot{y} &= -c\left[f\left(x\right)-x-z\right], \\ \dot{z} &= -d\left[y+z\right], \end{aligned}\right. \tag{1}$$

which is found in studying of the oscillations of electrical circuits. In this system $f(x)$ is a twice continuous differentiable function in R such that

$$xf(x)>0 \text{ for } x\neq 0, \tag{2}$$

$$|f(x)|<M \text{ for } x\in R, \tag{3}$$

M is a positive constant. The positive constants a, b, c, d are subordinate to the conditions

$$g > \frac{c}{2a} + \frac{d}{2a} + (1+b) - \sqrt{\left(\frac{c}{2a} - \frac{d}{2a}\right)^2 + \frac{bc}{a}}, \quad g = f'(0), \tag{4}$$

$$d > 4.6c \sup_{x \in R} \frac{f(x)}{x} + 9.7c + 5a + 2.4ab. \tag{5}$$

Under these assumptions all solutions of (1) are defined in R and through every point $(t_0, x_0, y_0, z_0) \in R \times R^3$ goes an unique integral curve (see [2] and [3]).

In [1] it has been proved that the system (1) possesses a closed phase curve, different from the degenerate curve, consisting of its unique equilibrium position — the origin of coordinates.

This curve lies in the solid homeomorphic torus V bounded by two cone surfaces

$$\frac{y+z}{\sqrt{2}\sqrt{x^2+y^2+z^2}} = \frac{1}{2}, \tag{6}$$

$$\frac{y+z}{\sqrt{2}\sqrt{x^2+y^2+z^2}} = -\frac{1}{2}, \tag{6'}$$

by the ellipsoid

$$\frac{1}{c^2}y^2 + \frac{1}{b}\left(\frac{x}{a} + \frac{y}{c}\right)^2 + \frac{z^2}{cd} = L, \tag{7}$$

and by the cylindrical surface

$$q_2^2 + q_3^2 = K. \tag{8}$$

Here

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = S \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix},$$

L is the smallest and K is the greatest positive constant for which the orbits of (1) cross the contour of V from the outside; S is the matrix, reducing the matrix to the corresponding to (1) linear system

$$\left| \begin{aligned} \dot{x} &= a\left[(g-1-b)\,x - z\right], \\ \dot{y} &= -c\left[(g-1)\,x - z\right], \\ \dot{z} &= -d\left[y+z\right], \end{aligned} \right. \tag{1'}$$

of Jordan's normal form.

To each of the existing two possibilities for the roots of the characteristic equation of (1′)

$$\lambda^3 + \left[d + a(1+b) - ag\right]\lambda^2 + d\left[c + a(1+b) - ag\right]\lambda + abcd = 0, \tag{9}$$

namely,

$$\lambda_1 < 0, \quad \lambda_2 = \mu + i\kappa, \quad \lambda_3 = \mu - i\kappa\mu > 0, \tag{10}$$

or

$$\lambda_1 < 0, \quad \lambda_2 = \mu_1 > 0, \quad \lambda_3 = \mu_2 > 0, \quad \mu_1 \neq \mu_2, \tag{11}$$

corresponds a different matrix S.

In both cases (10) and (11) the following estimate is valid:

$$-d < \lambda_1 < -0.876d. \tag{12}$$

We know from [1] that except for the two orbits, lying in the domain $x^2 < y^2 + z^2 + 4yz$, all phase curves of (1) enter V and remain in it with the growth of time t. Furthermore, all orbits, starting from

$$D_1 = V \cap \{x = 0,\ z < 0\},$$

intersect

$$D_2 = V \cap \{x = 0,\ z > 0\}$$

passing from $x > 0$ to $x < 0$, and all orbits, starting from D_2, intersect D_1 passing from $x < 0$ to $x > 0$.

We denote by T the length of the interval of time for which the orbits of (1) with initial points in D_1 intersect D_1 again after their passing through D_2. In the present paper an estimate from above for T is obtained. The estimate does not depend on the initial point of the orbits if this point belongs to D_1.

A key role at the estimate of T will be played by the following

Lemma 1. *Suppose that the constants B and γ are defined by*

$$B = \begin{cases} \dfrac{1.2\sqrt{K}\,|\kappa|\,ac}{g\,(d+1)^4} & \textit{for } (10), \\[2ex] \dfrac{\sqrt{K}\,|\mu_2 - \mu_1|\,bc}{g^4\,(d+1)^7} & \textit{for } (11), \end{cases} \qquad \gamma = \begin{cases} \dfrac{0.6\sqrt{K}\,|\kappa|\,ac}{g\,(d+1)^3} & \textit{for } (10), \\[2ex] \dfrac{0.4\sqrt{K}\,|\mu_2 - \mu_1|\,bc}{g^4\,(d+1)^7} & \textit{for } (11). \end{cases}$$

Then for the third coordinates of all points from $V \cap \{|x| \leq \gamma\}$ the estimate $|z| \geq B$ is valid.

Proof. The intersection of the conical surface (6) with the plane $y + z = 1$ is defined by the system

$$\left| \begin{array}{l} x = \pm\sqrt{1 + 2z - 2z^2}, \\ y = 1 - z, \quad (1 - \sqrt{3})/2 \leq z \leq (1 + \sqrt{3})/2. \end{array} \right. \tag{13}$$

Let $(\tilde{x}, \tilde{y}, \tilde{z})$ be a point from this intersection. The generatrix of (6), passing through $(\tilde{x}, \tilde{y}, \tilde{z})$, intersects (8) in the point $A(x_1, y_1, z_1)$, where

$$x_1 = \pm\sqrt{1 + 2\tilde{z} - 2\tilde{z}^2}\,(S_1)_\pm, \quad y_1 = (1 - \tilde{z})\,(S_1)_\pm, \quad z_1 = \tilde{z}\,(S_1)_\pm. \tag{14}$$

Here

$$(S_1)_\pm = \frac{\sqrt{K}|\det S|}{(W_1)_\pm},$$

$$(W_1)_\pm = \Big\{[\pm A_{12}\sqrt{1+2\tilde{z}-2\tilde{z}^2} + A_{22}(1-\tilde{z}) + A_{32}\tilde{z}]^2 \tag{15}$$

$$\times [\pm A_{13}\sqrt{1+2\tilde{z}-2\tilde{z}^2} + A_{23}(1-\tilde{z}) + A_{33}\tilde{z}]^2\Big\}^{1/2},$$

A_{ij} are the algebraic adjuncts to the elements of the matrix S, $1 \le i,j \le 3$.

In (13)–(15) there is a correspondence between the signs ±, written before the radicals, and those in the symbol $(S_1)_\pm$.

Investigating $x_1(\tilde{z})$, $y_1(\tilde{z})$ and $z_1(\tilde{z})$ for $\tilde{z} \in \big[(1-\sqrt{3})/2, (1+\sqrt{3})/2\big]$ allows us to conclude that the intersection of (6) and (8) represents a simple closed curve. By analogy we come to the same conclusion also for the intersection of (6′) and (8).

Calculating A_{ij} and $\det S$ and estimating them by the following inequalities resulting from (5), (12) and from the relations between the roots of (9) and its coefficients:

$$a < \frac{d}{5}, \quad ab < \frac{d}{2.4}, \quad c < \frac{d}{9.7}, \tag{16}$$

$$abc < \mu^2 + \kappa^2 < 0,05d^2 \quad \text{for the case (10)}, \tag{17}$$

$$abc < \mu_1^2 + \mu_2^2 < a^2g^2 \quad \text{for the case (11)}, \tag{18}$$

we get that for $(1-\sqrt{3})/2 \le \tilde{z} \le (1-\sqrt{3})/4$ the following estimate is valid:

$$|z_1| \ge B. \tag{19}$$

The points

$$A_1\left(\frac{1}{2}\sqrt{4-\sqrt{3}}, \frac{3+\sqrt{3}}{4}, \frac{1-\sqrt{3}}{4}\right) \quad \text{and} \quad A_2\left(-\frac{1}{2}\sqrt{4-\sqrt{3}}, \frac{3+\sqrt{3}}{4}, \frac{1-\sqrt{3}}{4}\right)$$

are obtained from (13) for $z = 1-\sqrt{3}$.

We denote by α_j the plane containing the axis of the cone (6) and passing through the point A_j, $j = 1, 2$:

$$\alpha_1 : x = \frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z),$$

$$\alpha_2 : x = -\frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z).$$

From the position of V in the space it follows that for the third coordinates of all points from

$$V_1 = V \cap \left\{-\frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z) \le x \le \frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z)\right\}$$

the estimate (19) is valid as well.

By analogy, considering (6'), one proves similarly that the estimate (19) is valid also for the third coordinates of all points from

$$V_2 = V \cap \left\{ \frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z) \le x \le -\frac{\sqrt{4-\sqrt{3}}}{1+\sqrt{3}}(y-z) \right\}.$$

We consider the generatrix of the cones (6) and (6'), lying in α_1 and α_2, and their intersection points with the surface (8).

Estimating A_{ij}, $\det S$ and using the inequalities (12), (16)–(18), we obtain for the first coordinates of these points the inequality

$$|x| \ge \gamma.$$

Then $V \cap \{|x| \le \gamma\} \subset V_1 \cup V_2$. The lemma is proved. □

Let $x(t, x_0, y_0, z_0)$ denote the x-component of the orbit of (1), corresponding to the initial condition $x(0) = x_0$, $y(0) = y_0$, $z(0) = z_0$.

We introduce now the constants

$$M_1 = \max_{|x| \le a\sqrt{L}\left(\sqrt{b}+1\right)} |f'(x)|, \quad M_2 = \max_{|x| \le a\sqrt{L}\left(\sqrt{b}+1\right)} |f''(x)|.$$

Lemma 2. *Let δ_1, δ_2, t_0 be defined as follows:*

$$\begin{aligned} &\delta_1 = \min\left(\gamma, \frac{\sqrt{g^2+BM_2}-g}{M_2}\right), \quad \delta_2 = \frac{a}{4} t_0 B \left(1 - \frac{1}{4.10^n}\right), \\ &t_0 = \frac{1}{10^n} \frac{Ba}{2d^2\left\{M\left(M_1+1\right) + d\sqrt{L}\left[d\left(1+M_1\right)\left(1+\sqrt{b}\right)+1\right]\right\}}. \end{aligned} \tag{20}$$

Then:

a) *$x\left(t, x_0, y_0, z_0\right)$ is an increasing function in the interval $|t| \le t_0$, $\forall\left(x_0, y_0, z_0\right) \in V_1 \cap \{|x| \le \delta_1\}$;*

b) *$x(t_0/2) - x_0 \ge \delta_2$ and $x_0 - x(-t_0/2) \ge \delta_2$, $\forall(x_0, y_0, z_0) \in V_1 \cap \{|x| \le \delta_1\}$.*

Proof. We develop $f(x)$ by Taylor's formula about $x = 0$:

$$f(x) = gx + \frac{x^2}{2} f''(\theta x), \quad \theta \in (0,1), \; g = f'(0).$$

Let (x_0, y_0, z_0) be an arbitrary point from $V_1 \cap \{|x| \le \gamma\}$. Then $-z_0 > B$. We develop $\dot{x}\left(t, x_0, y_0, z_0\right)$ about $t = 0$:

$$\dot{x}\left(t, x_0, y_0, z_0\right) = a\left[\left(g-1-b\right)x_0 - z_0 + \frac{x_0^2}{2} f''\left(\theta x_0\right)\right] + t\ddot{x}\left(\theta_1 t\right), \quad \theta_1 \in (0,1).$$

Suppose that $|x_0| \le \delta_1$. Then

$$\left|(g-1-b)\,x_0+\frac{x_0^2}{2}f''(\theta x_0)\right| \le \frac{B}{2}$$

and

$$(g-1-b)\,x_0-z_0+\frac{x_0^2}{2}f''(\theta x_0) \ge \frac{B}{2}. \qquad (21)$$

The orbit with an origin (x_0, y_0, z_0) lies in V and the following estimates are valid:

$$|x| \le a\sqrt{L}\left(\sqrt{b}+1\right), \quad |y| \le c\sqrt{L}, \quad |z| \le \sqrt{cd}\sqrt{L}, \quad \forall\,(x,y,z)\in V. \qquad (22)$$

Then

$$|\ddot{x}(t,x_0,y_0,z_0)| \le d^2\left\{M(M_1+1)+d\sqrt{L}\left[d(1+M_1)\left(1+\sqrt{b}\right)+1\right]\right\}$$

for any $t \in R$ and any $(x_0, y_0, z_0) \in V$. Therefore, for every $t \in R$ for which $|t| \le t_0$ and for every point $(x_0, y_0, z_0) \in V_1 \cap \{|x| \le \delta_1\}$ the following two inequalities are valid:

$$|t\ddot{x}(\theta_1 t, x_0, y_0, z_0)| \le \frac{1}{10^n}\frac{Ba}{2}, \quad \dot{x}(t, x_0, y_0, z_0) > 0,$$

where n is a suitable positive integer.

To prove the second part of the lemma, we develop $x(t, x_0, y_0, z_0)$ about $t = 0$:

$$x(t,x_0,y_0,z_0) = x_0 + a\left[f(x_0)-(1+b)\,x_0-z_0\right]t+\frac{t^2}{2}\ddot{x}(\theta_2 t,x_0,y_0,z_0), \quad \theta_2\in(0,1).$$

On the one hand, it follows from (20) that

$$\left|\frac{t_0^2}{8}\ddot{x}(\theta_2 t,x_0,y_0,z_0)\right| \le \frac{t_0}{16}\frac{Ba}{10^n}$$

and, from the other hand, (21) implies

$$\left[f(x_0)-(1+b)\,x_0-z_0\right]\frac{t_0}{2} \ge \frac{B}{4}t_0.$$

These two estimates together with (21) yield

$$x\left(\frac{t_0}{2}\right)-x_0 \ge \delta_2, \quad \forall\,(x_0,y_0,z_0)\in V_1\cap\{|x_0|\le\delta_1\}.$$

In a similar way the estimate for $x_0 - x(-t_0/2)$ is obtained. The lemma is proved. □

Theorem 1. *For $\delta = \min(\delta_1, \delta_2)$ we have:*

a) *All orbits of (1) going from points $(x_0, y_0, z_0) \in V_1 \cap \{-\delta \leq x \leq 0\}$ intersect the plane $x = 0$ for an interval of time not greater than $t_0/2$;*

b) *All orbits of (1) going from points $(x_0, y_0, z_0) \in V_1 \cap \{0 \leq x \leq \delta\}$ intersect the plane $x = \delta$ for an interval of time not greater than $t_0/2$;*

c) *All orbits of (1) going from points $V_2 \cap \{0 \leq x \leq \delta\}$ intersect the plane $x = 0$ for an interval of time with length not greater than $t_0/2$;*

d) *All orbits of (1) going from points $V_2 \cap \{-\delta \leq x \leq 0\}$ intersect the plane $x = -\delta$ for an interval of time not greater than $t_0/2$.*

The proof follows from Lemma 2 and the repetition of reasoning for the points from $V_2 \cap \{|x| \leq \gamma\}$.

We are already prepared to prove the following

Theorem 2. *All orbits of equation* (1), *starting from D_1, intersect again D_1 for the interval of time not greater than $T = 2t_0 + 4\sqrt{L}/(b\delta)$.*

Proof. Let $(0, y_0, z_0)$ be an arbitrary point from D_1. Denote with t_1 the first moment when the orbit beginning at this point intersects $x = \delta$, and with T_1 the second moment. Then

$$x\,(t, 0, y_0, z_0) \geq \delta, \quad \forall t \in [t_1, T_1]\,. \tag{23}$$

Multiply the first equation of (1) by c, the second by a and sum the results:

$$c\frac{dx}{dt} + a\frac{dy}{dt} = -abcx\,(t)\,.$$

Integrate this equation on the given orbit from t_1 to T_1. We obtain

$$bc\int_{t_1}^{T_1} x\,(t, 0, y_0, z_0)\; dt = y\,(t_1) - y(T_1).$$

Estimate the left-hand side of this equation from below with the help of (23), and the right-hand one above with the help of (22). This yields the inequality

$$T_1 - t_1 \leq \frac{2\sqrt{L}}{b\delta}. \tag{24}$$

In a similar way we obtain that if t_2 is the first moment in which the considered orbit intersects $x = -\delta$, and T_2 is the second such moment, then

$$T_2 - t_2 \leq \frac{2\sqrt{L}}{b\delta}. \tag{25}$$

The obtained estimates do not depend of the point $(0, y_0, z_0) \in D_1$.

Finally, we take into account Lemma 2, the results of Theorem 1 and eqs. (24) and (25). The theorem is proved. □

REFERENCES

1. Pliss, V. A. Non-local problems of theory of oscillations. Moscow, 1964 (in Russian).
2. Wintner, A. The non-local existence problems of ordinary differential equations. *Amer. Journ. of Math.*, **67**, 1945.
3. Wintner, A. The infinites of the non-local existence problem of ordinary differential equations. *Amer. Journ. of Math.*, **68**, 1946.

Received March 15, 1998
Revised May 30, 1998

Ana D. MIHAILOVA
Faculty of Mathematics and Informatics
"K. Preslavski" University of Shumen
BG-9700 Shumen, Bulgaria

Neko N. GEORGIEV
Faculty of Mathematics and Informatics
"K. Preslavski" University of Shumen
BG-9700 Shumen, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 91, 1997

K-THEORY OF THE C^*-ALGEBRA OF MULTIVARIABLE WIENER-HOPF OPERATORS ASSOCIATED WITH SOME POLYHEDRAL CONES IN R^n

NIKOLAJ BUYUKLIEV

We consider the C^*-algebra $WH(R^n, P)$ of the multivariable Wiener-Hopf operators associated with a polyhedral cone in R^n and the extension $0 \to \mathcal{K} \to WH(R^n, P) \to WH(R^n, P)/\mathcal{K} \to 0$.

The main theorem states that if P satisfies suitable geometric conditions (satisfied, e.g., for all simplicial cones and the cones in R^n, $n \leq 3$), then $K_*(WH(R^n, P)) = (0,0)$; $K_*(WH(R^n, P)/\mathcal{K}) = (0, Z)$, and that the index map is an isomorphism. In the cource of the proof we construct a Fredholm operator in $WH(R^n, P)$ with an index 1. The proof is inductive and uses the Mayer-Vietoris exact sequence and the standart six term exact sequence in K-theory.

Keywords: K-theory, Wiener-Hopf operators

1991/95 Math. Subject Classification: 47A53

0. INTRODUCTION

Let P be a polyhedral cone in R^n. The Wiener-Hopf operators are obtained by compressing the left convolution operators on $L^2(R^n)$ to the $L^2(P)$:

$$W(f)\xi(t) = \int_P f(t-s)\xi(s)\,ds.$$

The C^*-algebra $WH(R^n, P)$, generated by $W(f)$ when f runs through $C_c(R^n)$, is the C^*-algebra of multivariable Wiener-Hopf operators. It is studied with various techniques in [2, 4, 5].

In [4] P. Muhly and J. Renault prove that $WH(R^n, P)$ contains $\mathcal{K} = \mathcal{K}(P)$ — the ideal of the compact operators in $B(L^2(P))$. They obtain a composition series for $WH(R^n, P)$:

$$0 \subset I_0 \cong \mathcal{K} \subset I_1 \subset \ldots \subset I_n \cong WH(R^n, P), \tag{0.1}$$

where $I_k/I_{k-1} \cong C_0(Z) \otimes \mathcal{K}$ and Z is an appropriate locally compact space. They state a problem to calculate the K-theory of $WH(R^n, P)$. Here are calculated $K_*(WH(R^n, P))$ and $K_*(WH(R^n, P)/\mathcal{K})$ when P satisfies suitable geometric conditions (satisfied, e.g., for all simplicial cones and the cones in R^n, $n \leq 3$).

In the present paper we consider the extension

$$0 \to \mathcal{K} \to WH(R^n, P) \to WH(R^n, P)/\mathcal{K} \to 0. \tag{0.2}$$

Our first observation is that if there exists an index 1 Fredholm operator and if $K_*(WH(R^n, P)/\mathcal{K}) = (0, Z)$ (in order to simplify notations, the K-theory is considered to be Z_2-graded theory: $K_*(A) = K_0(A) \oplus K_1(A)$), then we may apply the fundamental six term exact sequence of K-theory:

$$\begin{tikzcd}
K_0(\mathcal{K}) \arrow[r] & K_0(WH(R^n, P)) \arrow[r] & K_0(WH(R^n, P)/\mathcal{K}) \arrow[d] \\
K_1(WH(R^n, P)/\mathcal{K}) \arrow[u, "\text{ind}"] & K_1(WH(R^n, P)) \arrow[l] & K_1(\mathcal{K}) \arrow[l]
\end{tikzcd} \tag{0.3}$$

Then we obtain that $K_*(WH(R^n, P)) = (0, 0)$ and the index map of the extension (0.2):

$$\text{ind} : K_1(WH(R^n, P)/\mathcal{K}) \to K_0(\mathcal{K}), \tag{0.4}$$

is an isomorphism.

Further, the quotient $WH(R^n, P)/\mathcal{K}$ can be represented as a groupoid C^*-algebra. There are groupoid subalgebras, which are more simple (in a K-theory sense). The basic idea is to construct an increasing sequence of such algebras

$$\mathcal{B}_1 \subset \mathcal{B}_2 \subset \ldots \subset \mathcal{B}_N \cong WH(R^n, P)/\mathcal{K}$$

and to calculate their K-theory applying the Mayer-Vietoris exact sequence in each step.

The groupoid approach gives naturally pullback diagrams of appropriate defined groupoid C^*-algebras:

$$\begin{tikzcd}
\mathcal{B}_k \arrow[r] \arrow[d] & \mathcal{B}_{k-1} \arrow[d] \\
\mathcal{D}_k \arrow[r] & \mathcal{A}_k
\end{tikzcd}$$

Then the corresponding exact Mayer-Vietoris sequence is

$$\begin{tikzcd}
K_0(\mathcal{B}_k) \arrow[r] & K_0(\mathcal{D}_k) \oplus K_0(\mathcal{B}_{k-1}) \arrow[r] & K_0(\mathcal{A}_k) \arrow[d] \\
K_1(\mathcal{A}_k) \arrow[u] & K_1(\mathcal{D}_k) \oplus K_1(\mathcal{B}_{k-1}) \arrow[l] & K_1(\mathcal{B}_k) \arrow[l]
\end{tikzcd}$$

In a general situation it is not sufficient to know only the K-groups of $\mathcal{A}_k$ and $\mathcal{D}_k$. Now we note that when the middle terms in the above exact sequence are trivial, then the maps corresponding to the vertical arrows are isomorphisms. If all these K-groups are trivial, then the same is true for $\mathcal{B}_k$. This fact motivates us to define the class of exhaustible cones — i.e. those cones, for which we can find a sequence of subalgebras as above, but having a trivial K-theory.

The organization of the paper is as follows: In Section 1 we set up the groupoid notations. In Section 2 we prove that there exists a Fredholm operator with index 1 in $WH(R^n, P)^\dagger$ — the algebra with the identity adjoined. As a corollary of the six term exact sequence in the K-theory we show that if $K_*(WH(R^n, P)/\mathcal{K}) = (0, Z)$, then $K_*(WH(R^n, P)) = (0, 0)$ and the index map (0.2) is an isomorphism. Section 3 is concerned with the quotient $WH(R^n, P)/\mathcal{K}$. We define geometrically the property a cone to be exhaustible and we prove the main Theorem 3.5. An example is given.

1. PRELIMINARIES

In this section we collect some facts concerning the groupoid approach to C^*-algebras and the groupoid construction made in [4] of a groupoid whose associated groupoid C^*-algebra is isomorphic to the one generated by the Wiener-Hopf operators.

In the paper P is a polyhedral cone in R^n, i.e. P is generated by its extreme rays. We assume that P contains no line and spans R^n. Let $\mathcal{F}(P)$ denote the set of all faces of P; we count P and $\{0\}$ among the faces of P. For $F \in \mathcal{F}(P)$, $\langle F \rangle$ is the linear subspase $F - F$ generated by F and $St(F)$ is the collection of all faces containing F.

In [4] P. Muhly and J. Renault prove that in a general context (G is a locally compact group and P is its subsemigroup) $\mathcal{B} = WH(G, P)$ is isomorphic with an explicitly constructed groupoid C^*-algebra $C^*(\mathcal{G})$. Here we briefly recall their construction in the case $G = R^n$

First step in this construction is the definition of a locally compact space Y. It may be presented as

$$Y = \{(F, t) : F \in \mathcal{F}(P);\ t \in R^n \ominus \langle F \rangle\}.$$

The space R^n is imbedded in Y ($t \mapsto (\{0\}, t)$) as a dense subset and the space X is defined to be the closure of P in Y. There exists a natural action of R^n on Y and the basic for the constructed groupoid $\mathcal{G}$, whose C^*-algebra yields $WH(R^n, P)$, is a reduction of a transformation group $Y \times R^n$ by the closed subspace X of Y. Explicitly, the elements of $\mathcal{G} = Y \times R^n | X$ are the pairs $(x, s) \in Y \times R^n$ such that $x \in X$ and $x + s \in X$. The family of measures on X:

$$\lambda^x(y, s) = \delta_x(y)\ \chi_X(y)\ \chi_X(y + s)\ ds$$

(here χ_X is the characteristic function of X), is called the left Haar system of measures of $\mathcal{G}$.

The family $C_c(\mathcal{G})$ of the finite functions on X becomes a normed C^*-algebra under the operations and the norm defined as follows:

$$f * g\,(x,t) = \int f(x,s)g(x+s,t-s)\chi_X(x)\chi_X(x+s)\,ds,$$

$$f^*(x,t) = \overline{f(x+s,-s)},$$

$$\|f\|_I = \sup\Big\{\int f\,d\lambda^x,\ \int f^*\,d\lambda^x : x \in X\Big\}.$$

The completion of $C_c(\mathcal{G})$ by the norm $\|\cdot\|_I$ is $L_I(\mathcal{G})$ and $C^*(\mathcal{G})$ is defined as their enveloping C^*-algebra.

Let $A \subset \mathcal{F}(P)$ and $X(A)$ consist of those $x = (F,t) \in X$ such that the face F belongs to A. Then $\mathcal{G}(A)$ is defined to be the groupoid obtained by a reduction of $\mathcal{G}$ by $X(A)$ and $C^*(\mathcal{G}(A))$ to be the corresponding C^*-algebra. We denote some often used groupoids as follows: $\mathcal{G}(F) = \mathcal{G}(\{F\})$, $\mathcal{G}_0 = \mathcal{G}(\{0\})$ and $\mathcal{G}_\infty = \mathcal{G}(\mathcal{F}(P)\setminus\{0\})$.

1.1. Proposition ([4, § 4.7]). *There exists an isomorphism between the C^*-algebra $WH(R^n,P)$ and the groupoid C^*-algebra $C^*(\mathcal{G})$. $WH(R^n,P)$ contains $\mathcal{K} = \mathcal{K}(L^2(P))$, which is isomorphic to $C^*(\mathcal{G}_0)$, and the quotient $WH(R^n,P)/\mathcal{K}$ is isomorphic to $C^*(\mathcal{G}_\infty)$.*

Let F be a face of P. The set $P-F$ is a cone containing the linear space $\langle F\rangle$ and $P_F = (P-F)/\langle F\rangle$ denotes the cone in $R^n \ominus \langle F\rangle$ determined by F. More generally, if $F_1 \in St(F)$, then $F - F_1$ contains $\langle F\rangle$ and the map

$$F_1 \mapsto (F_1 - F)/\langle F\rangle$$

is an order preserving bijection between $ST(F)$ and $\mathcal{F}(P_F)$. The next proposition describes the groupoid C^*-algebra $C^*(\mathcal{G}(St(F))) \cong WH(R^n, P-F)$.

1.2. Proposition ([4, § 3.7.1]). *$WH(R^n,P-F)$ is isomorphic to $WH(R^n \ominus \langle F\rangle, P_F) \otimes C^*_{\text{red}}(\langle F\rangle)$, where the tensor product is endowed with the least C^*-cross norm.*

We note that $C^*_{\text{red}}(\langle F\rangle) \cong C_0(\langle F\rangle)$ and that all the algebras considered here are postliminal ([2]) and there exists an unique C^*-cross norm. The above fact and the Bott periodicity say that $K_i(WH(R^n,P-F)) = K_{i+l\,(\text{mod } 2)}(WH(R^n \ominus \langle F\rangle, P_F))$, where $l = \dim(\langle F\rangle)$.

1.3. Observation. Here we describe a construction which allows us to use often the Mayer-Vietoris exact sequence.

Let choose subsets A, B, C and D of $\mathcal{F}(P)$ such that $B = C \cup D$ and $A = C \cap D$. Let denote the corresponding groupoids by $\mathcal{G}(A)$, $\mathcal{G}(B)$, $\mathcal{G}(C)$, $\mathcal{G}(D)$ and their groupoid C^*-algebras by $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$.

When one glues the groupoids $\mathcal{G}(C)$, $\mathcal{G}(D)$ along $\mathcal{G}(A)$, the result is $\mathcal{G}(B)$. Then the following diagram of C^*-algebras is commutative:

$$\begin{tikzcd}
\mathcal{B} \arrow[r, "\psi_2"] \arrow[d, "\psi_1"'] & \mathcal{C} \arrow[d, "\varphi_2"] \\
\mathcal{D} \arrow[r, "\varphi_1"] & \mathcal{A}
\end{tikzcd}$$

The C^*-algebra $\mathcal{B}$ is a pullback of $(\mathcal{C}, \mathcal{D})$ along φ_1, φ_2 (i.e. $\mathcal{B} \cong \{(c,d) : \varphi_1(d) = \varphi_2(c)\} \subseteq \mathcal{C} \oplus \mathcal{D}$ (cf. [1, § 15.3])).

By [1, § 18.12.4], when a pullback diagram of C^*-algebras is given as above, then we may write the corresponding exact Mayer-Vietoris sequence

$$\begin{tikzcd}
K_0(\mathcal{B}) \arrow[r] & K_0(\mathcal{D}) \oplus K_0(\mathcal{C}) \arrow[r] & K_0(\mathcal{A}) \arrow[d] \\
K_1(\mathcal{A}) \arrow[u] & K_1(\mathcal{D}) \oplus K_1(\mathcal{C}) \arrow[l] & K_1(\mathcal{B}) \arrow[l]
\end{tikzcd}$$

2. CONSTRUCTION OF A FREDHOLM OPERATOR WITH INDEX 1

In this section an one-dimensional projector $E(x,s)$ in $WH(R^n, P)$ and an essentially unitary operator S in $WH(R^n, P)^\dagger$ — the algebra with the identity adjoined, are given explicitly.

Let us choose points y_i, $i = 1, 2, \ldots, N$, on the extreme rays of P such that $|y_i| = 1$. We may assume that y_i, $i = 2, 3, \ldots, n$, determine extreme rays of $P_1 = (P - F_1)/\langle F_1 \rangle$ and let P' be the cone spaned on y_i, $i = 1, 2, 3, \ldots, n$. We define

$$E(x,s) = C \prod_{k=1}^{n} e^{-(x,y_k)} e^{-\frac{1}{2}(s,y_k)} \chi_{P'}(x) \chi_{P'}(x+s),$$

$$F(x,s) = C\, e^{\frac{1}{2}(s,y_1)} \chi_{(-\infty,0]}(s, y_1) \prod_{k=2}^{n} e^{-(x,y_k)} e^{-\frac{1}{2}(s,y_k)} \chi_{P'}(x) \chi_{P'}(x+s).$$

2.1. Lemma. (i) *E is an one-dimensional projection in $WH(R^n, P)$.*
(ii) *F is in $WH(R^n, P)$ and satisfies the equalities*

$$F^* * F = F + F^* \quad \textit{and} \quad F * F^* = F + F^* - E.$$

Proof. Let first assume that $P = R^n_+$. Then we rewrite $E^*(x,s)$ and $F(x,s)$:

$$E(x,s) = \prod_{k=1}^{n} e^{-x_k} e^{-\frac{1}{2}s_k} \chi_{R^n_+}(x) \chi_{R^n_+}(x+s),$$

$$F(x,s) = e^{\frac{1}{2}s_1} \chi_{(-\infty,0]}(s) \prod_{k=2}^{n} e^{-x_k} e^{-\frac{1}{2}s_k} \chi_{R^n_+}(x) \chi_{R^n_+}(x+s) \chi_X(x) \chi_X(x+s).$$

The elements of $L_I(\mathcal{G})$ are the measurable functions on $\mathcal{G}$ with a finite norm $\|\cdot\|_I$. We observe that $E(x,s) = \overline{E(x+s,-s)} = E^*(x,s)$. Using the Fubini theorem and the fact that

$$\int e^{-(x+s)}\chi_{[0,\infty)}(x+s)\,ds = 1, \tag{2.1}$$

where $x, s \in R$, we obtain

$$|E|_I = \sup\{\int E(x,s)\,ds,\ \int E(x,s)\,ds : x \in X\} \leq 1$$

and E belongs to $WH(R^n,P)$. Similar estimate proves that F is in $WH(R^n,P)$ and we omit it.

To prove that E is an one-dimensional projector, we have to check the equalities $E = E^*$, $E = E*E$ and $\operatorname{tr}(E) = 1$. The first one is obvious. Using again the Fubini theorem and (2.1), we get

$$\begin{aligned} E*E(x,t) &= \int E(x,s)E(x+s,t-s)\chi_X(x+s)\,ds \\ &= E(x,t)\int \prod_{k=1}^{n} e^{-(x_k+s_k)}\chi_{[0,\infty)}(x+s)\,ds = E(x,t). \end{aligned}$$

By [4] $E(x,x-s)$, where $x \in P$, may be considered as a kernel of a selfadjoint integral operator in $L^2(R_+)$. Using the well-known formula for the trace of a selfadjoint integral operator with a continuous kernel, we obtain

$$\operatorname{tr}(E) = \int E(x,0)dx = \int \prod_{k=1}^{n} e^{-x_k}\chi_{[0,\infty)}(x)\,dx = 1$$

and hence E is an one-dimensional projector.

We rewrite F as follows:

$$F(x,s) = e^{\frac{1}{2}s_1}\chi_{(-\infty,0])}(s)E_{n-1},$$

$$F*(x,s) = e^{-\frac{1}{2}s_1}\chi_{[0,\infty))}(s)E_{n-1},$$

and then easy but tedious calculations prove the equalities of (ii).

Further, let Φ be the linear map determined by the matrice $(y_{i,j})$. Then the map $(x,t) \mapsto (\Phi(x),\Phi(t))$ may be extended to a topological isomorphism between $\mathcal{G}(R^n,R^n_+)$ and $\mathcal{G}(R^n,P')$. The measures in the left-hand Haar systems differ with a constant $C = |\det(y_{i,j})|$ and the statement is true if $P = P'$.

Finally, in the general case for P, the supports of $E(x,s)$ and $F(x,s)$ are in the reduction of $\mathcal{G}(R^n,P')$ by the $X(\{0\}) \cup X(F_1)$, which is a subgroupoid of $\mathcal{G}$. Thus $E(x,s)$ and $F(x,s)$ are in $C^*(\mathcal{G})$ and the above equalities are satisfied.

2.2. Theorem. (i) *There exists a Fredholm operator* $S \in WH(R^n,P)$ *such that* $\operatorname{ind}S = 1$.

Proof. Let $S = 1-F$. Then by Lemma 2.1 we have $S^*S = 1$ and $SS^* = 1-E$.

2.3. Corollary. *If* $K_*(WH(R^n,P)/\mathcal{K}) = (0,Z)$, *then*:

(i) $K_*(WH(R^n, P)) = (0,0)$ *and*
(ii) *the index map of the extension* (2)

$$\mathrm{ind} : K_1(WH(R^n, P)/\mathcal{K}) \to K_0(\mathcal{K}) \tag{2.2}$$

is an isomorphism.

Proof. Let us consider the fundamental six-term exact sequence of K-theory corresponding to the extension (2):

$$\begin{tikzcd}
K_0(\mathcal{K}) \arrow[r] & K_0(WH(R^n, P)) \arrow[r] & K_0(WH(R^n, P)/\mathcal{K}) \arrow[d] \\
K_1(WH(R^n, P)/\mathcal{K}) \arrow[u, "\mathrm{ind}"] & K_1(WH(R^n, P)) \arrow[l] & K_1(\mathcal{K}) \arrow[l]
\end{tikzcd}$$

Let $[S]$ be the generator of $K_1(WH(R^n, P)/\mathcal{K}) = Z$ and let $[E]$ be the generator of $K_0(\mathcal{K}) = Z$. If $\mathrm{ind}([S]) = m[E]$, then the image of the morphism "ind" is $mZ \subset Z = K_0(\mathcal{K})$. But by Theorem 2.2 $[E]$ belongs to the image of ind. Thus $|m| = 1$ and ind is an isomorphism. We know that the right-hand groups of (6) are equal to 0, thus $K_*(WH(R^n, P)) = (0,0)$.

3. K-THEORY OF THE QUOTIENT ALGEBRA

3.1. Proposition. *Let $n \leq 3$. Then $K_*(WH(R^n, P)/\mathcal{K}) = (0, Z)$.*

Proof. Let $n = 1$ and $P = R_+$. We write the extension (2) as follows:

$$0 \to \mathcal{K} \to WH(R, R_+) \to C_0(R) \to 0.$$

The fact that $K_*(WH(R^n, P)/\mathcal{K}) = K_*(C_0(R)) = (0, Z)$ is well known.

Let $n = 2$ and P be a polyhedral cone in R^2 (the quarter plane-case). The faces of P are $\{0\}$, P and two one-dimentional faces F_1 and F_2.

Let us denote:

$$\mathcal{G}_1 = \mathcal{G}|X(F_1) \cup X(P) \quad \text{and} \quad \mathcal{B}_1 = C^*(\mathcal{G}_1);$$
$$\mathcal{G}_2 = \mathcal{G}|X(F_2) \cup X(P) \quad \text{and} \quad \mathcal{B}_2 = C^*(\mathcal{G}_2);$$
$$\mathcal{G}_{1,2} = \mathcal{G}|X(P) \quad \text{and} \quad \mathcal{B}_{1,2} = C^*(\mathcal{G}_{1,2}).$$

We recall that $\mathcal{G}_\infty = \mathcal{G}|X(F_1) \cup X(F_2) \cup X(P)$ and by Proposition 1.1 $WH((R^2, P)/\mathcal{K} \cong C^*(\mathcal{G}_\infty)$.

There exists an isomorphism of groupoids $\mathcal{G}_1 \cong R \times \mathcal{G}(R, R_+)$, where $\mathcal{G}(R, R_+)$ is the groupoid corresponding to the Wiener-Hopf algebra with $P = R_+ \subset R$. Then $\mathcal{B}_1 = C^*(\mathcal{G}_1) \cong C_0(R) \otimes WH((R, R_+)$ and therefore

$$\begin{aligned} K_*(\mathcal{B}_1) = K_*(C^*(\mathcal{G}_1)) &= K_*(C_0(R)) \times K_*(WH((R, R_+)) \\ &= (0, Z) \times (0,0) = (0,0). \end{aligned}$$

Analogously, $K_*(\mathcal{B}_2) = K_*(C^*(\mathcal{G}_2)) = (0,0)$.

Further, $\mathcal{B}_{1,2} = C^*(\mathcal{G}_{1,2}) \cong C_0(R^2)$ and $K_*(\mathcal{B}_{1,2}) = K_*(C_0(R^2)) = (Z, 0)$.

There is a pullback diagram of C^*-algebras by (1.3):

$$\begin{tikzcd}
C^*(\mathcal{G}_\infty) \arrow[r] \arrow[d] & \mathcal{B}_1 \arrow[d] \\
\mathcal{B}_2 \arrow[r] & \mathcal{B}_0
\end{tikzcd}$$

Further, the corresponding Mayer-Vietoris exact sequence is

$$\begin{tikzcd}
K_0(WH((R^2,P)/\mathcal{K}) \arrow[r] & K_0(\mathcal{B}_1)\oplus K_0(\mathcal{B}_2) \arrow[r] & K_0(C_0(R^2)) \arrow[d] \\
K_1(C_0(R^2)) \arrow[u] & K_1(\mathcal{B}_1)\oplus K_1(\mathcal{B}_2) \arrow[l] & K_1(WH((R^2,P)/\mathcal{K}) \arrow[l]
\end{tikzcd}$$

The middle terms equal $\{0\}$ and hence the vertical maps are isomorphisms:

$$K_0(WH(R^2,P)/\mathcal{K}) = K_1(C_0(R^2)) = 0,$$
$$K_1(WH(R^2,P)/\mathcal{K}) = K_0(C_0(R^2)) = Z.$$

Let us recall that the set $St(F_l)$ of faces of P containing F_l is bijective to the set of faces of $P - F_l$. Therefore each subset A_l of $St(F_l)$ determines a subset $\widetilde{A_l}$ of $\mathcal{F}(P_l)$, where P_l is the lower-dimensional cone

$$(P - F_l)/(\langle F_l\rangle) \subset R^n \ominus \langle F_l\rangle\rangle.$$

The next definition is recursive and outlines the cones with which we deal.

3.2. Definition. Let P be a polyhedral cone in R^n, $n \geq 2$. We say that $L \subset \mathcal{F}(P)$ satisfies the condition (C) iff:

(i) there exists an one-dimensional face which does not belong to L;

(ii) L is an union of stars of some one-dimensional faces of P;

(iii) there is an ordering $F_1, \dots, F_k$ of these one-dimensional faces such that for each $l = 2, \dots, k$

$$A_l = St(F_l) \cap [St(F_1) \cup \ldots \cup St(F_{l-1})]$$

determines a subset $\widetilde{A_l} \subset \mathcal{F}(P_l)$ which satisfies the condition (C).

If $n = 2$, we count $St(F_1)$ and $St(F_2)$ among the sets satisfying the condition (C).

3.3 Definition. We say that P is exhaustible iff there exists an one-dimensional face F of P such that $L = \mathcal{F}(P)\backslash\{\{0\}, F\}$ satisfies the condition (C).

3.4. Lemma. *The cones in R^2 and R^3 and the simplicial cones in R^n are exhaustible.*

Proof. When $n = 2$, by the definitions P is exhaustible.

Let $n = 3$ and P be a polyhedral cone in R^3. Let us choose the custumary ordering $F_1, F_2, \dots, F_N$ of the one-dimensional faces of P (i.e. the extreme rays of P) . Two neighbouring one-dimensional faces F_k and F_{k+1} of P (the calculations with the indices are mod N) span the two-dimensional face $F_{k,k+1}$. The rest faces of P are $\{0\}$ and P.

It is sufficient to prove that $F_1, F_2, \ldots, F_{N-1}$ satisfy the condition (C). It is evident that

$$A_l = St(F_l) \cap [St(F_1) \cup \ldots \cup St(F_{l-1})] = St(F_{l-1,l})$$

for $l = 2, \ldots, N-1$. The associated with $St(F_{l-1,l}) \subset \mathcal{F}_P$ family of faces of the cone $P_l = (P - F_l)/\langle F_l \rangle \subset R^2$ satisfies the condition (C) by the definition and this proves the case $n = 3$.

Finally, let P be a simplicial cone in R^n. Note that each collection of extreme rays uniquely determines a face of P and for each one-dimensional faces F_k and F_l of P follows that $St(F_k) \cap St(F_l) = St(F_{k,l})$. A trivial induction on the dimension n proves that for each ordering $F_1, F_2, \ldots, F_n$ of the one-dimensional faces of P and for each $l < n$ the subset $F_1, F_2, \ldots, F_l$ satisfies the condition (C).

3.5. Theorem. *Let P be an exhaustible polyhedral cone in R^n, $n \geq 2$. Then:*
(i) $K_*(WH(R^n, P)) = (0,0)$;
(ii) $K_*(WH(R^n, P)/\mathcal{K}) = (0, Z)$;
(iii) *the index map of the extension* (0.2)

$$\text{ind} : K_1(WH(R^n, P)/\mathcal{K}) \to K_0(\mathcal{K}) \tag{3.1}$$

is an isomorphism;
(iv) *if $A \subset \mathcal{F}(P)$ satisfies the condition* (C), *then* $K_*(C^*(\mathcal{G}(A))) = (0,0)$.

Proof. We shall prove the theorem by induction on the dimension n. If $n = 2$, Lemma 3.1 and Theorem 2.3 prove the statements (i)–(iv). Now suppose that they are true for $2, \ldots, n-1$.

Let P be an exhaustible polyhedral cone in R^n. By Definition 3.3 there exists an ordering $F_1, \ldots, F_N$ of the one-dimensional faces of P such that $B_{N-1} = St(F_1) \cup \ldots \cup St(F_{N-1}) \subset \mathcal{F}(P)$ satisfies the condition (C) given in Definition 3.2.

Now let us consider some subsets of $\mathcal{F}(P)$ and the corresponding C^*-algebras:
$D_k = St(F_k)$ and $\mathcal{D}_k = C^*(\mathcal{G}(D_k))$ for $k = 1, 2, \ldots, N$;
$B_k = St(F_k) \cup \ldots \cup St(F_k)$ and $\mathcal{B}_k = C^*(\mathcal{G}(B_k))$ for $k = 1, 2, \ldots, N$;
$A_k = D_k \cap B_k$ and $\mathcal{A}_k = C^*(\mathcal{G}(A_k))$ for $k = 2, 3, \ldots, N$.
We note that $\mathcal{B}_1 = \mathcal{D}_1$ and $\mathcal{B}_N = WH(R^n, P)/\mathcal{K}$ by Proposition 1.1.

Our first aim is to compute the K-theory of these algebras, in particular to prove that $K_*(\mathcal{B}_k) = (0,0)$ for $k = 1, 2, \ldots, N-1$.

By Proposition 1.2 there is an isomorphism $\mathcal{D}_k \cong C_0(\langle F_k \rangle) \otimes WH(R^n \ominus \langle F_k \rangle, P_k)$. Since by the condition (i) of the inductive supposition $K_*(WH(R^n \ominus \langle F_k \rangle, P_k)) = (0,0)$, then for $k = 1, 2, \ldots, N$

$$K_*(\mathcal{D}_k) = (0,0).$$

Further, $A_k \subset St(F_k)$ and by Proposition 1.2 it determines a family $\widetilde{A_k} \subset \mathcal{F}(P_k)$ of faces of $\mathcal{F}(P_k)$ and associated with it groupoid C^*-algebra $\widetilde{\mathcal{A}_k}$ such that $\mathcal{A}_k \cong C_0(\langle F_k \rangle) \otimes \widetilde{\mathcal{A}_k}$.

If $1 < k < N$, then by Definition 3.2 (iii) A_k satisfies the condition (C), and therefore by the condition (iv) of the inductive supposition it follows $K_*(\widetilde{\mathcal{A}_k}) = (0,0)$ and hence

$$K_*(\mathcal{A}_k) = (0,0), \quad k = 2,3,\dots,N-1. \tag{3.2}$$

Now we shall show that $\mathcal{A}_N$ has a non-trivial K-theory. Indeed, $A_N = St(F_N)\setminus\{F_N\}$ and by Proposition 1.2

$$\mathcal{A}_N \cong C_0(\langle F_N\rangle) \otimes [WH(R^n \ominus \langle F_N\rangle, P_N)/\mathcal{K}].$$

By the condition (ii) of the inductive supposition $K_*(WH(R^n) \ominus \langle F_N\rangle, P_N)/\mathcal{K}) = (0,Z)$ and hence

$$K_*(\mathcal{A}_N) \cong K_*(C_0(R)) \otimes K_*(WH(R^n \ominus \langle F_N\rangle, P_N)/\mathcal{K}) = (0,Z)\times(0,Z) = (Z,0).$$

The equalities $B_k = B_{k-1} \cup D_k$, $A_k = B_{k-1} \cap D_k$ and Proposition 1.4 imply that there are pullbacks of the corresponding C^*-algebras for $k = 1,2,\dots,N$:

$$\begin{tikzcd}
\mathcal{B}_k \arrow[r] \arrow[d] & \mathcal{B}_{k-1} \arrow[d] \\
\mathcal{D}_k \arrow[r] & \mathcal{A}_k
\end{tikzcd}$$

Now we shall prove that

$$K_*(\mathcal{B}_k) = (0,0); \; k = 1,2,\dots,N-1. \tag{3.3}$$

Indeed, $\mathcal{B}_1 = \mathcal{D}_1$ and $K_*(\mathcal{B}_1) = K_*(\mathcal{D}_1) = (0,0)$. Suppose that the above holds for $1,\dots,k-1$ and we write the Mayer-Vietoris exact sequence

$$\begin{tikzcd}
K_0(\mathcal{B}_k) \arrow[r] & K_0(\mathcal{B}_{k-1}) \oplus K_0(\mathcal{D}_k) \arrow[r] & K_0(\mathcal{A}_k) \arrow[d] \\
K_1(\mathcal{A}_k) \arrow[u] & K_1(\mathcal{B}_{k-1}) \oplus K_1(\mathcal{D}_k) \arrow[l] & K_1(\mathcal{B}_k) \arrow[l]
\end{tikzcd}$$

The middle terms in this exact sequence are the groups $\{0\}$, hence the vertical arrows maps are isomorphisms. For $k = 1,2,\dots,N$ it follows that $K_0(\mathcal{B}_k) = K_1(\mathcal{A}_k)$ and $K_1(\mathcal{B}_k) = K_0(\mathcal{A}_k)$. Using $(N-2)$ times the Mayer-Vietoris exact sequence, we obtain that $K_*(\mathcal{B}_k) = (0,0)$ for $k = 1,2,\dots,N-1$. Here we note that the proof of the condition (iv) is the same as the above fragment and we omit it. Further, the final Mayer-Vietoris exact sequence gives

$$K_*(\mathcal{B}_N) = (0,Z). \tag{3.4}$$

So, the condition (ii) is verified for n. The left standing for n conditions (i) and (iii) follow from Theorem 2.3.

It is attractive to conjecture that all the polyhedral cones in R^n are exhaustible. However, we are unable to prove it. The next example shows that the ordering of the one-dimensional faces in Definition 3.2 (iii) is essential. We construct $L \subset \mathcal{F}(P)$ which is an union of stars of some one-dimensional faces, but which does not satisfy the condition (C), because some of the corresponding C^*-algebras have non-trivial K-groups.

3.6. Example. Let P be a cone in R^4 such that the cut Q through P determined of a hyperplane α is a cube. We denote the extreme points of Q (ordered in the custumary way) by $A_1, \ldots, A_8$ and the corresponding one-dimensional faces of P by $F_1, \ldots, F_8$:

$L_1 = St(F_1)$, $K_*(C^*(\mathcal{G}(L_1))) = (0,0)$,
$L_2 = St(F_2) \cup L_1$, $K_*(C^*(\mathcal{G}(L_2))) = (0,0)$,
$L_3 = St(F_6) \cup L_2$, $K_*(C^*(\mathcal{G}(L_3))) = (0,0)$,
$L_4 = St(F_7) \cup L_3$, $K_*(C^*(\mathcal{G}(L_4))) = (0,0)$,
$L_5 = St(F_8) \cup L_4$, $K_*(C^*(\mathcal{G}(L_5))) = (Z,0)$,
$L_6 = St(F_4) \cup L_5$, $K_*(C^*(\mathcal{G}(L_6))) = (0,Z)$,
$L_7 = St(F_3) \cup L_6$, $K_*(C^*(\mathcal{G}(L_7))) = (0,0)$.

Clearly, L_7 with the above order of the one-dimensional faces is not exhaustible. It can be verified that the customary order of the extreme points of the cube determines an order of the one-dimensional faces of P such that L_7 is exhaustible.

ACKNOWLEDGEMENTS. I wish to express my indebtedness to R. Levi for the numerous and fruitful comments at the various stages of this paper.

REFERENCES

1. Blacadar, B. K-theory for operator algebras. *Math. Sci. Res. Inst. Publ.*, **5**, Springer-Verlag, New-York, 1986.
2. Dynin, A. Multivariable Wiener-Hopf operators. I: Integral Equation Operator Theory. 1986, 537–556.
3. Klee, V. Some characterizations of convex polyhedra. *Acta Math.*, **102**, 1959, 79–107.
4. Muhly, P., J. Renault. C^*-algebras of multivariable Wiener-Hopf operators. *Trans. Amer. Math. Soc.*, **274**, 1982, 1–44.
5. Park, E. Index theory and Toeplitz algebras on certain cones in Z^2. *J. Operator Theory*, **23**, 1990, 125–146.
6. Renault, J. A groupoid approach to C^*-algebras. *Lecture Notes in Math.*, **793**, Springer-Verlag, New Jork, 1980.

Received March 17, 1998

Section of Functional and Real Analysis
Department of Mathematics
Sofia University
5 James Bourchier Blvd.
BG-1164 Sofia, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Meçanique
Tome 91, 1997

EACH 11-VERTEX GRAPH WITHOUT 4-CLIQUES HAS A TRIANGLE-FREE 2-PARTITION OF VERTICES

EVGENI NEDIALKOV, NEDYALKO NENOV

Let G be a graph, $\mathrm{cl}(G)$ denotes the clique number of the graph G. By $G \to (3,3)$ we denote that in any 2-partition $V_1 \cup V_2$ of the set $V(G)$ of his vertices either V_1 or V_2 contains 3-clique (triangle) of the graph G; $\alpha = \min\{|V(G)|,\ G \to (3,3) \text{ and } \mathrm{cl}(G) = 4\}$, $\beta = \min\{|V(G)|,\ G \to (3,3) \text{ and } \mathrm{cl}(G) = 3\}$. In the current article, we consider graphs G with the property $G \to (3,3)$. As a consequence from proven results it follows that $\alpha = 8$ and $\beta \geq 12$.

Keywords: chromatic number, triangle free partition of vertices of graph

1991/95 Math. Subject Classification: 05C55, 05C35

1. INTRODUCTION

We consider only finite, non-oriented graphs without loops and multiple edges. $V(G)$ and $E(G)$ denote respectively the set of the vertices and the set of the edges of the graph G. We say that G is an n-vertex graph when $|V(G)| = n$. If $v, w \in V(G)$ and $[v, w] \in E(G)$, then v and w are called adjacent vertices of the graph G, and the edge $[v, w]$ is called incidental to the vertices v and w. For $v \in V(G)$ we denote by $\mathrm{Ad}(v)$ the set of all vertices adjacent to v, and by $d(v)$ the number of such vertices, i.e. $d(v) = |\mathrm{Ad}(v)|$. For the graph G we put $\delta(G) = \min\{d(v) \mid v \in V(G)\}$ and $\Delta(G) = \max\{d(v) \mid v \in V(G)\}$. The set of vertices of a given graph is called *clique* if arbitrary two of its elements are adjacent vertices. If the number of vertices in a given clique is p, then we call it *p-clique*. The biggest natural number p, such that the graph G contains a p-clique, is called *clique-number* of G and is denoted by $\mathrm{cl}(G)$.

Let $u \in V(G)$ and $[v,w] \in E(G)$. We say that the vertex u is adjacent to the edge $[v,w]$ if $\{u,v,w\}$ is a 3-clique of G.

The set of vertices of a given graph is called *anticlique* if each two of them are not adjacent. The anticlique consisting of p vertices is called *p-anticlique.* The biggest natural number p, for which the graph G has p-anticlique, is called the number of independence of G and is denoted by $\alpha(G)$.

The graph G_1 is called a subgraph of the graph G if $V(G_1) \subset V(G)$ and $E(G_1) \subset E(G)$. Let $M \subset V(G)$. We denote by $\langle M \rangle$ the subgraph generated by M, i.e. $V(\langle M \rangle) = M$, and two vertices of M are adjacent in $\langle M \rangle$ if and only if they. are adjacent in G. We denote by $G - M$ the subgraph of G that is produced by taking off the vertices of M and all the edges incidental to the vertices of M.

The partition of $V(G)$ into r pairwise disjoint subsets, $V(G) = V_1 \cup V_2 \cup \ldots \cup V_r$, is called r-partition of vertices. If all of V_i, $i = 1, \ldots, r$, are anticliques, then this partition is called *r-chromatic partition.* The smallest natural number r, for which G has an r-chromatic partition, is called *chromatic number* of G and is denoted by $\chi(G)$. The graph G is called *k-chromatic* if $\chi(G) = k$. The graph G is called *vertex-critical k-chromatic* graph if $\chi(G) = k$ and $\chi(G - v) < k$ for arbitrary $v \in V(G)$. We need the following obvious

Proposition 1. *If G is a vertex-critical k-chromatic graph, then $\delta(G) \geq k-1$.*

The *supplement* $\overline{G}$ of a given graph G is defined by setting $V(G) = V(\overline{G})$; two vertices are adjacent in $\overline{G}$ if and only if they are not adjacent in G. It is clear that $\alpha(G) = \mathrm{cl}(\overline{G})$.

Let p and q be given natural numbers. The number $R(p,q)$ is the minimum of all natural numbers n, such that for arbitrary n-vertex graph G either $\mathrm{cl}(G) \geq p$ or $\alpha(G) \geq q$. The existence of the numbers $R(p,q)$ is proved by F. Ramsey in [14]. Therefore they are refered as *Ramsey numbers.* We need the identity $R(4,3) = R(3,4) = 9$, see [3], and more precisely, its obvious consequence:

Proposition 2. *If $|V(G) \geq 9$ and $\mathrm{cl}(G) \leq 3$, then $\alpha(G) \geq 3$.*

If arbitrary two vertices of the given n-vertex graph are adjacent, then it is called *complete* n-vertex graph and is denoted by K_n. The simple cycle of length n is denoted by C_n. Let G_1 and G_2 be two graphs without common vertices, i.e. $V(G_1) \cap V(G_2) = \varnothing$. We denote by $G_1 + G_2$ the graph G, for which $V(G) = V(G_1) \cup V(G_2)$ and $E(G) = E(G_1) \cup E(G_2) \cup E'$, where $E' = \{[v_1, v_2] \mid v_i \in V(G_i),\ i = 1,2\}$.

2. MAIN RESULTS

Definition. *The 2-partition $V(G) = V_1 \cup V_2$ of the verteces of the graph G is free of 3-cliques if each of the sets V_1 and V_2 does not contain a 3-clique of the graph G. We write $G \to (3,3)$ when there is no 3-cliques free 2-partition of the vertices of G.*

It is obvious that if $\chi(G) \leq 4$, then G has a 3-cliques free 2-partition of vertices. Therefore we have the following

Proposition 3. *If $G \to (3,3)$, then $\chi(G) \geq 5$.*

It is clear that $K_5 \to (3,3)$ and, conversely, if $\mathrm{cl}(G) \geq 5$, then $G \to (3,3)$. The opposite direction is false since it is easy to check that $\overline{C}_9 \to (3,3)$, but $\mathrm{cl}(\overline{C}_9) = 4$.

Definition. *We denote by α the minimum of all natural numbers n such that there exists an n-vertex graph $G \to (3,3)$ with $\mathrm{cl}(G) = 4$. We denote by β the smallest natural n such that there is an n-vertex graph $G \to (3,3)$ with $\mathrm{cl}(G) = 3$.*

We prove in this paper that $\alpha = 8$ and the unique 8-vertex $G \to (3,3)$ with $\mathrm{cl}(G) = 4$ is the graph $K_1 + \overline{C}_7$ (Theorem 1). The existence of the number β is proved by P. Erdös and C. Rogers in [1]. R. Irving shows in [5] that $\beta \leq 17$. N. Nenov constructs in [9] a 14-vertex graph $\Gamma_1 \to (3,3)$ with $\mathrm{cl}(\Gamma_1) = 3$ (see Fig. 1), showing that $\beta \leq 14$. In the paper [10] N. Nenov proves that $\beta \geq 11$. In the present work we prove that $\beta \geq 12$ (Theorem 2).

Theorem 1. *Let the graph G be such that $G \to (3,3)$ and $\mathrm{cl}(G) = 4$. Then $|V(G)| \geq 8$ and $|V(G)| = 8$ only if $G = K_1 + \overline{C}_7$.*

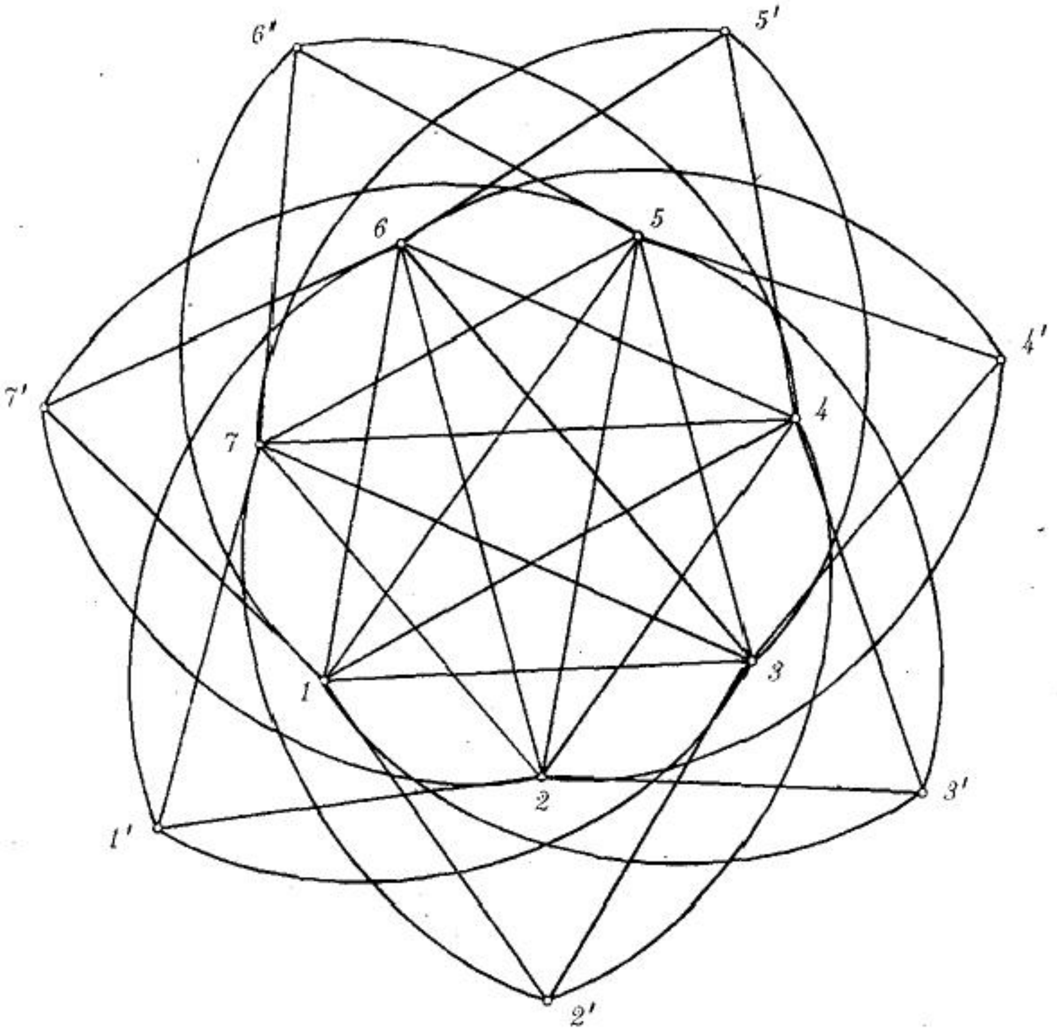

Fig. 1. Graph Γ_1

Theorem 2. *Let the 11-vertex graph G be such that $\mathrm{cl}(G) = 3$. Then G has a 3-cliques free 2-partition of vertices.*

Definition. *We say that the graph G is 3-saturated, if for an arbitrary anticlique A of G, the subgraph $G - A$ contains a 3-clique.*

To prove the Theorems 1 and 2 we need also the next assertions.

Theorem 3. *Let G be a 3-saturated graph and $\mathrm{cl}(G) = 3$. Then $|V(G)| \geq 7$ and $|V(G)| = 7$ only if $G = \overline{C}_7$.*

Theorem 4. *Let G be a 3-saturated graph, $|V(G)| = 8$ and $\mathrm{cl}(G) = 3$. Then either G is isomorphic to one of the graphs L_i, $i = 1, \ldots, 14$, shown at* Fig. 2–15, *or there is $v \in V(G)$ such that $G - v = \overline{C}_7$.*

Theorem 5. *The graphs L_i, $i = 1, \ldots, 14$, are 3-saturated, L_i is not isomorphic to L_j for $i \neq j$ and for arbitrary $v \in V(L_i)$ the graph $L_i - v$ is not isomorphic to $\overline{C}_7$.*

The connection between the 3-saturated graphs and the graphs satisfying $G \to (3,3)$ is given by the following

Proposition 4. *Let $G \to (3,3)$ and B be an anticlique in G. Then the subgraph $G_1 = G - B$ is 3-saturated.*

Proof. Assume that in fact G_1 is not 3-saturated and let A be such anticlique of G_1 that $G_2 = G_1 - A$ contains no 3-cliques. In such case $V(G) = V(G_2) \cup (A \cup B)$ is a 3-clique free 2-partition, which is a contradiction. ■

If a given graph has a 3-chromatic partition, then obviously it is not 3-saturated. That is why we have

Proposition 5. *If G is 3-saturated, then $\chi(G) \geq 4$.*

We state also the following obvious

Proposition 6. *If $\mathrm{cl}(G) = 3$, then $\mathrm{Ad}(v)$ does not contain 3-cliques for arbitrary $v \in V(G)$.*

We are going to use the next results.

Theorem A ([6]). *Let G be an 8-vertex graph with $\mathrm{cl}(G) = 3$ and $\alpha(G) = 2$. Then G is isomorphic to one of the graphs L_1, L_2, L_3 from* Fig. 2–4.

Different proofs of the above theorem could be found in [8], [12] and [13].

Theorem B ([10]). *Let the graph G be such that $\mathrm{cl}(G) \leq r$ and $\chi(G) \geq r + 1$ for some $r \geq 3$. If $|V(G)| = r + 4$, then one of the following two assertions is satisfied:*

(i) *there is a vertex $v \in V(G)$ such that $G - v = K_{r-2} + C_5$;*

(ii) *the graph G is isomorphic to one of the graphs $K_{r-3} + F_i$, $i = 1, \ldots, 7$, where the graphs $F_1, \ldots, F_7$ are shown at the* Fig. 16–22.

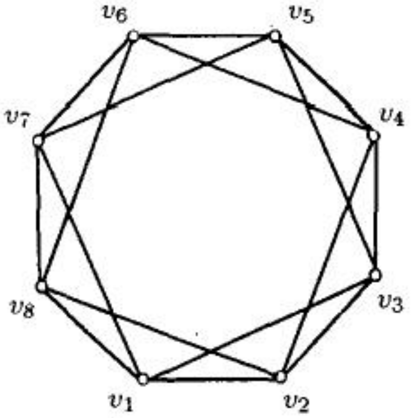

Fig. 2. Graph L_1

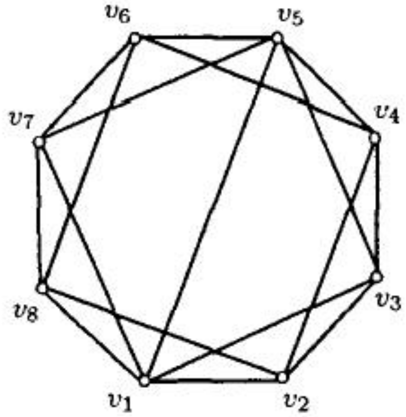

Fig. 3. Graph L_2

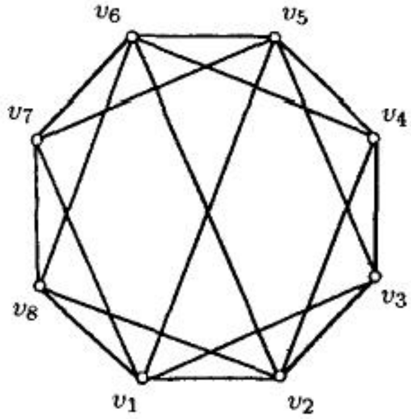

Fig. 4. Graph L_3

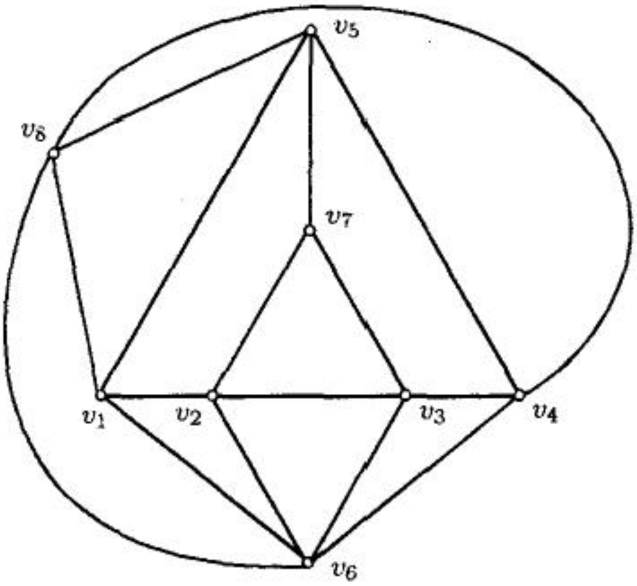

Fig. 5. Graph L_4

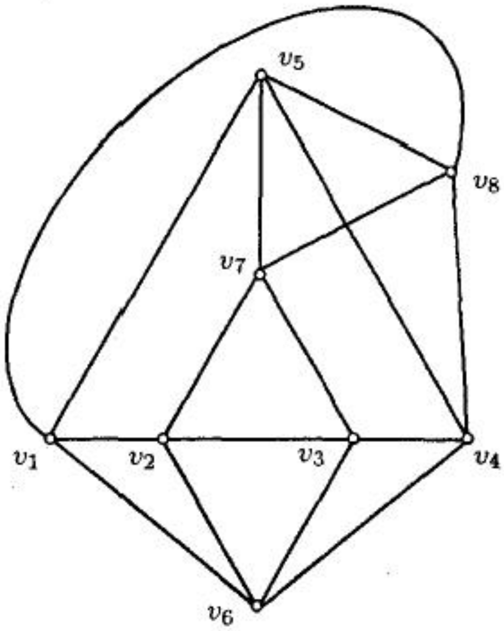

Fig. 6. Graph L_5

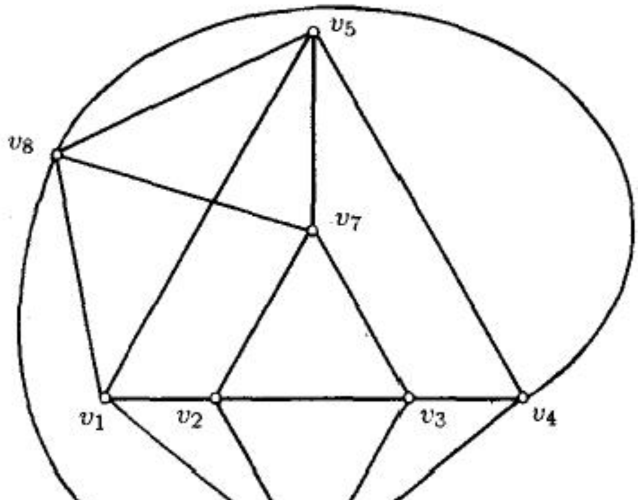

Fig. 7. Graph L_6

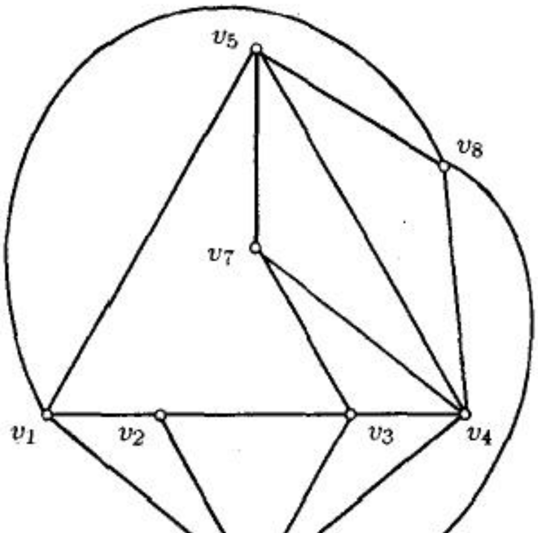

Fig. 8. Graph L_7

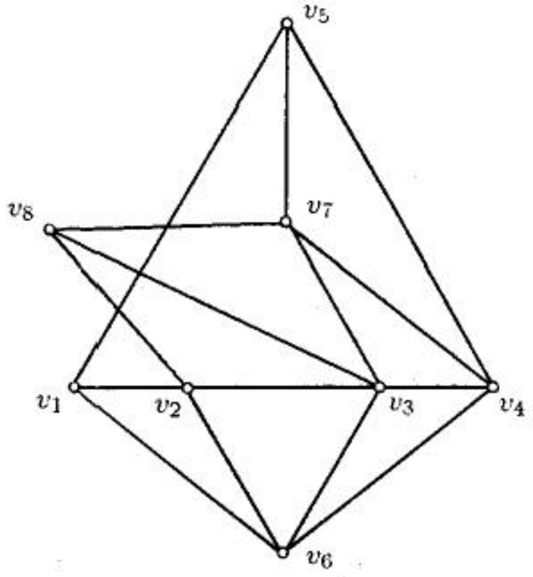

Fig. 9. Graph L_8

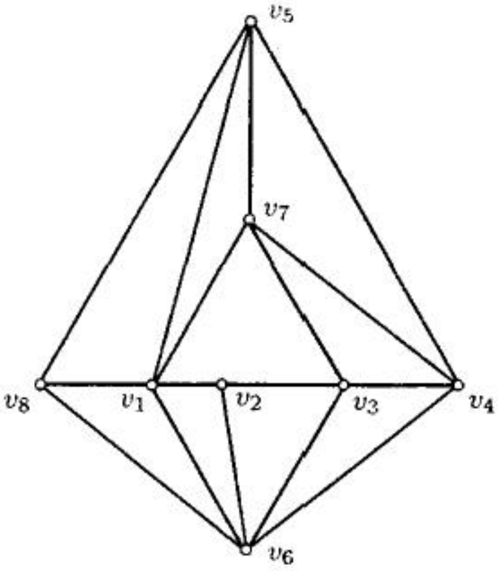

Fig. 10. Graph L_9

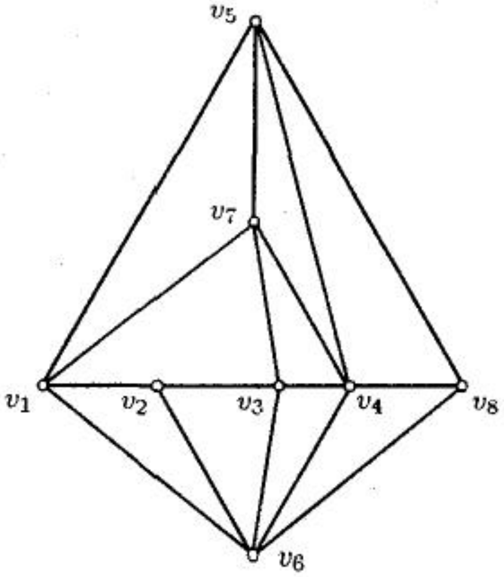

Fig. 11. Graph L_{10}

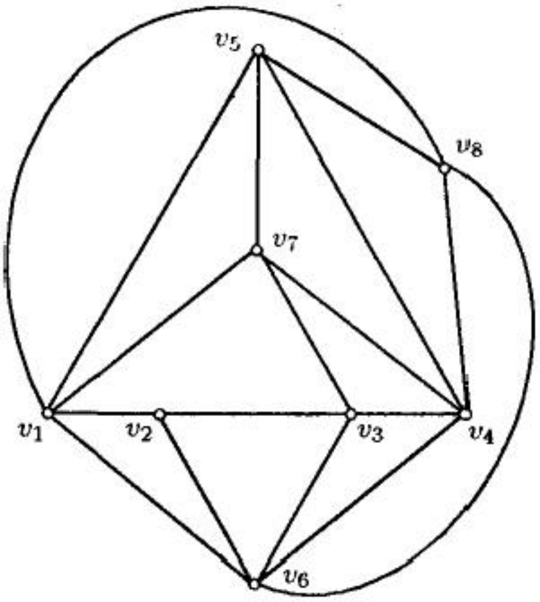

Fig. 12. Graph L_{11}

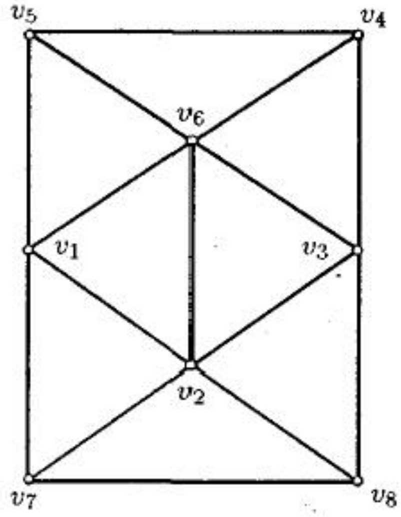

Fig. 13. Graph L_{12}

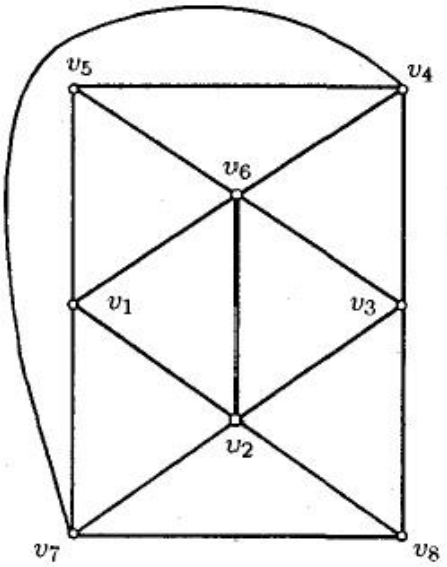

Fig. 14. Graph L_{13}

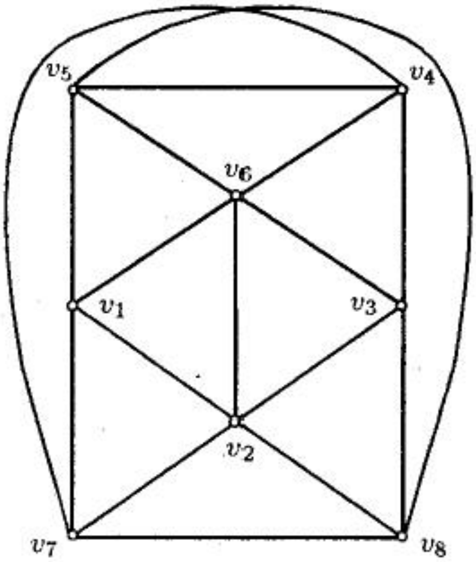

Fig. 15. Graph L_{14}

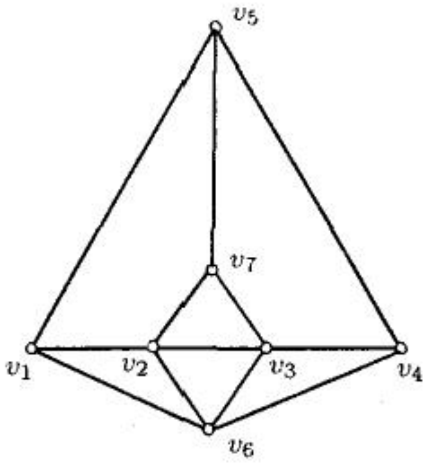

Fig. 16. Graph F_1

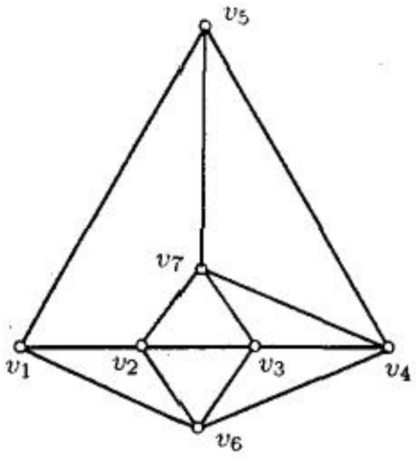

Fig. 17. Graph F_2

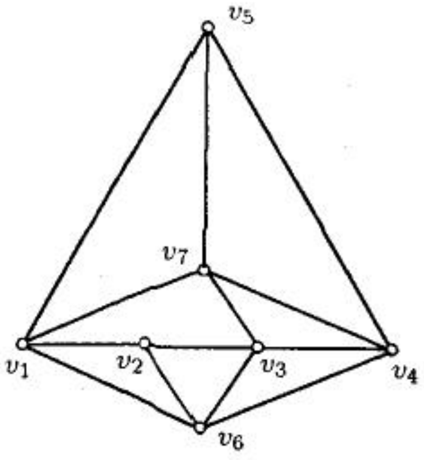

Fig. 18. Graph F_3

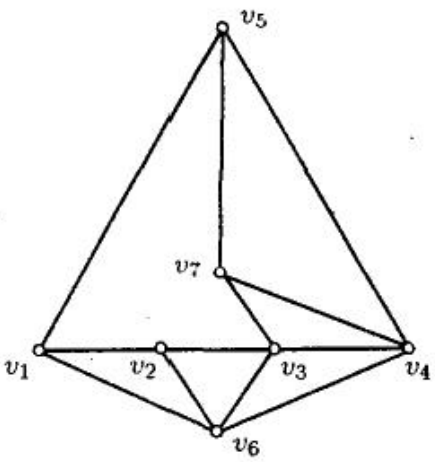

Fig. 19. Graph F_4

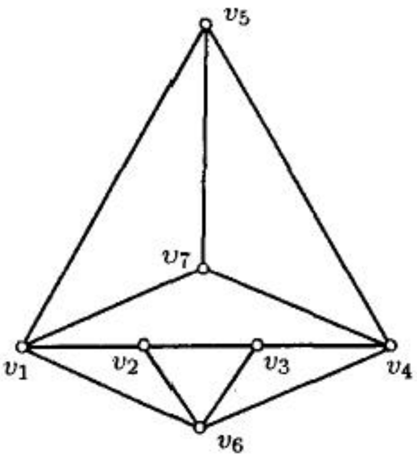

Fig. 20. Graph F_5

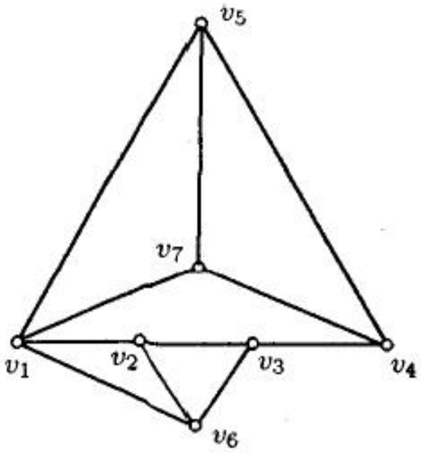

Fig. 21. Graph F_6

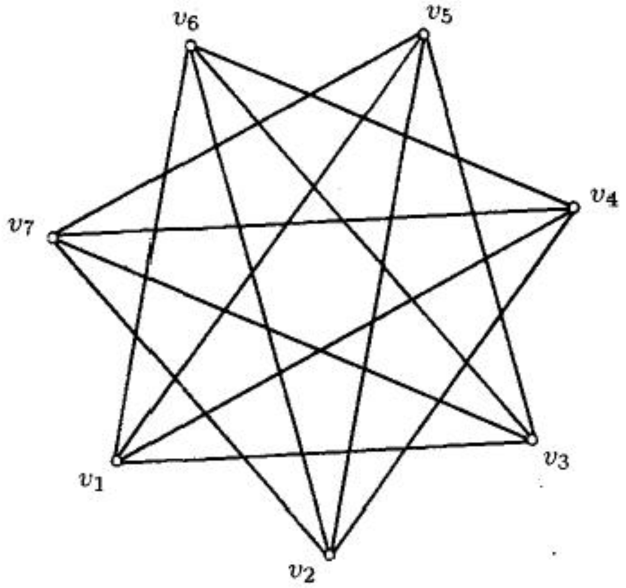

Fig. 22. Graph $F_7 = \overline{C}_7$

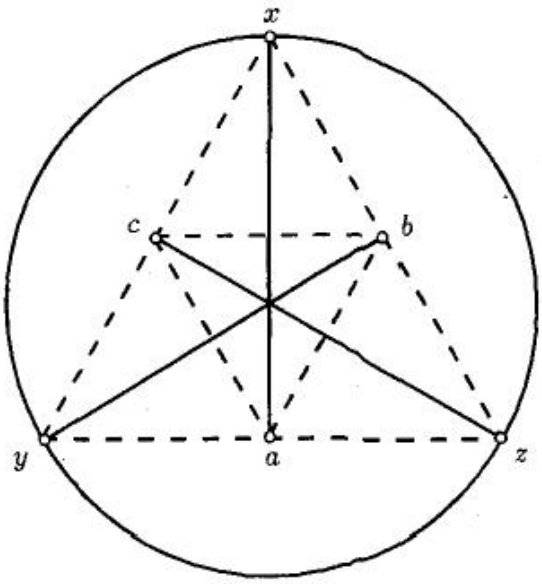

Fig. 23

Theorem C ([11]). *Let the graph G be such that $|V(G)| \leq 10$ and $\mathrm{cl}(G) = 3$. Then $\chi(G) \leq 4$.*

3. PROOFS OF THEOREMS 3, 4 AND 5

Proof of Theorem 3. Assume that G is 3-saturated and $|V(G)| \leq 7$. By adding if necessary few isolated vertices, we may assume that $|V(G)| = 7$. According to Proposition 5, $\chi(G) \geq 4$. As $\mathrm{cl}(G) = 3$, we see that G satisfies the conditions of Theorem B with $r = 3$ and we conclude that there are only two possible cases:

Case 1. $G - v = K_1 + C_5$ for some vertex $v \in V(G)$. Let $V(K_1) = \{u\}$. If u and v are not adjacent, then $G - \{u, v\} = C_5$ and consequently the graph G is not 3-saturated. If u and v are adjacent, then $G - u = \langle \mathrm{Ad}(u) \rangle$. According to Proposition 6, $\mathrm{Ad}(u)$ does not contain 3-cliques and therefore G is not 3-saturated.

Case 2. G coincides with some of the graphs F_i, $i = 1, \ldots, 7$ (Fig. 16–22). Each of the graphs F_i, $i = 1, \ldots, 6$, satisfies $F_i - \{v_6, v_7\} = C_5$, so these graphs are not 3-saturated. Then the assumption $|V(G)| \leq 7$ leads to $G = \overline{C}_7$. Obviously, $\overline{C}_7$ is 3-saturated, which finishes the proof. ∎

To prove Theorem 4, we need some preparation.

Lemma 1. *Let the graph G be such that $|V(G)| = 8$, $\mathrm{cl}(G) = 3$, and $\alpha(G) > 3$. Then G is not 3-saturated.*

Proof. Let $\{v_1, v_2, v_3, v_4\}$ be a 4-anticlique in G and v_5, v_6, v_7, v_8 be the other vertices of G. If $G - \{v_1, v_2, v_3, v_4\}$ contains no 3-cliques, we are done. In the other case, let for example $\{v_5, v_6, v_7\}$ be a 3-clique in G. From $\mathrm{cl}(G) = 3$ it follows that v_8 is non-adjacent to some of the vertices v_5, v_6, v_7. We may assume without a loss of generality that $[v_7, v_8] \notin E(G)$.

Case 1. The vertex v_8 is adjacent to some of v_5 and v_6, for example v_8 is adjacent to v_5. We denote by A the set consisting of the vertex v_5 and these of the vertices v_1, v_2, v_3, v_4, which are not adjacent to v_5. It is clear that A is an anticlique in G. As $G - A = \langle \mathrm{Ad}(v_5) \rangle$, according to Proposition 6 $G - A$ does not contain 3-cliques and the assertion of the lemma is shown to be true in this case.

Case 2. The vertex v_8 is not adjacent neither to v_5 nor to v_6. If A is the anticlique defined in Case 1, then $G - A = \langle \mathrm{Ad}(v_5) \cup \{v_8\} \rangle$. As the vertex v_8 is not adjacent to v_6 and v_7 and $\mathrm{Ad}(v_5)$ does not contain 3-cliques, $G - A$ does not contain 3-cliques, too. ∎

Lemma 2. *Let G be a 3-saturated 8-vertex graph and $\mathrm{cl}(G) = 3$. Then $\Delta(G) \leq 5$. Moreover, if v is a vertex of a 3-anticlique of G, then $d(v) \leq 4$.*

Proof. Assume that for some $v \in V(G)$ we have $d(v) = 7$. Then $G - v = \langle \mathrm{Ad}(v) \rangle$. According to Proposition 6, $\mathrm{Ad}(v)$ does not contain 3-cliques of G, which contradicts the fact that G is a 3-saturated graph. If we assume that $d(v) = 6$ and denote by w the vertex of G non-adjacent to v, then $G - \{v, w\} = \langle \mathrm{Ad}(v) \rangle$. Once again the last equality contradicts the 3-saturatedness of G. So, by now we have proved that $\Delta(G) \leq 5$.

Assume now that the second part of the lemma is false and let for example $\{v, u, w\}$ be a 3-anticlique of G and $d(v) > 4$. It follows that $G - \{v, u, w\} = \langle \mathrm{Ad}(v) \rangle$. Again an application of Proposition 6 gets a contradiction to the fact that G is a 3-saturated graph. ∎

Lemma 3. *Let G be a vertex-critical 4-chromatic graph, $|V(G)| = 8$, and G contain two 3-anticliques without common vertices. Then G is not a 3-saturated fraph.*

Proof. As $\chi(K_4) = 4$ and G is vertex-critical, $\mathrm{cl}(G) < 4$. Let $\{v_1, v_2, v_3\}$ and $\{v_4, v_5, v_6\}$ be the two 3-anticliques given by the condition, and v_7 and v_8 be the other vertices. If $G - \{v_4, v_5, v_6\}$ contains no 3-cliques, then the assertion is proved. Assume that $G - \{v_4, v_5, v_6\}$ contains a 3-clique and let for example $\{v_1, v_7, v_8\}$ be such 3-clique. By a similar argument we may assume that the graph $G - \{v_1, v_2, v_3\}$ contains a 3-clique, say $\{v_4, v_7, v_8\}$. From $\mathrm{cl}(G) < 4$ it follows that $[v_1, v_4] \notin E(G)$.

Assume that v_7 is not adjacent to v_2 and v_3. If v_7 is not adjacent also to v_5 and v_6, then $G - v_8$ does not contain 3-cliques and the lemma is proved. If v_7 is adjacent to both v_5 and v_6, then v_7 is adjacent to each of the vertices of the subgraph $G - \{v_2, v_3, v_7\}$, and from Proposition 6 it follows that $G - \{v_2, v_3, v_7\}$ contains no 3-cliques. If the vertex v_7 is adjacent to only one of v_5 and v_6, for example $[v_5, v_7] \in E(G)$ and $[v_6, v_7] \notin E(G)$, we consider the following two situations:

1. $[v_5, v_8] \notin E(G)$. It is clear that $G - \{v_5, v_8\}$ does not contain 3-cliques and consequently G is not 3-saturated.

2. $[v_5, v_8] \in E(G)$. From $\mathrm{cl}(G) < 4$ it follows that $[v_1, v_5] \notin E(G)$. The subgraph $G - v_8$ does not contain 3-cliques and consequently G is not 3-saturated.

So, in the case when v_7 is not adjacent to the vertices v_2 and v_3 the assertion is proved. Therefore we assume that v_7 is adjacent to some of the vertices v_2 and v_3. Similarly, we may assume also that v_7 is adjacent to some of the vertices v_5 and v_6. We put then without a loss of generality $[v_2, v_7] \in E(G)$ and $[v_5, v_7] \in E(G)$. If the vertex v_7 is adjacent to some of v_3 and v_6, then our assertion is a consequence of Lemma 2, because $d(v_7) \geq 6$. That is why we may and do assume that $[v_3, v_7], [v_6, v_7] \notin E(G)$.

Consider the subgraph $G - \{v_3, v_7\}$. If it does not contain 3-cliques, we are done. Let $G - \{v_3, v_7\}$ contain 3-cliques. Because $G - \{v_3, v_7\} = \langle \mathrm{Ad}(v_7) \cup \{v_6\} \rangle$ and $\mathrm{Ad}(v_7) = \{v_1, v_2, v_4, v_5, v_8\}$ does not contain 3-cliques, certainly $[v_6, v_8] \in E(G)$. By similar argument we conclude that $[v_3, v_8] \in E(G)$. If the vertex v_8 is adjacent also to some of v_2, v_5, then $d(v_8) \geq 6$ and we may apply Lemma 2 to get the conclusion. Therefore we assume that v_8 is not adjacent neither to v_2 nor to v_5.

Let us mention that at least one of the pairs $\{v_2, v_5\}$, $\{v_1, v_5\}$, $\{v_2, v_4\}$ is not adjacent in the graph G, because otherwise we would have $\langle \mathrm{Ad}(v_7) \rangle = C_5$ and $K_1 + C_5 \subset G$, which contradicts to the fact that G is a vertex-critical 4-chromatic graph, since $\chi(K_1 + C_5) = 4$. To conclude, let see that:

If $[v_2, v_5] \notin E(G)$, then $\{v_2, v_5, v_8\}$ is an anticlique and $G - \{v_2, v_5, v_8\}$ does not contain 3-cliques.

If $[v_1, v_5] \notin E(G)$, then $G - \{v_2, v_8\}$ does not contain 3-cliques.

If $[v_2, v_4] \notin E(G)$, then $G - \{v_5, v_8\}$ does not contain 3-cliques. ∎

Lemma 4. *Let G be a vertex-critical 4-chromatic graph and $|V(G)| = 8$. Then $\alpha(G - v) \geq 3$ for arbitrary $v \in V(G)$.*

Proof. It is obvious that if a 7-vertex graph has no 3-anticliques, then its chromatic number is bigger than 3. Therefore $\alpha(G - v) < 3$ implies $\chi(G - v) > 3$, which contradicts the fact that G is a vertex-critical 4-chromatic graph. ∎

Lemma 5. *Let G be a vertex-critical 4-chromatic graph and $|V(G)| = 8$. Then G is not a 3-saturated graph.*

Proof. If $\alpha(G) > 3$, then the assertion follows from Lemma 1. So, we assume that $\alpha(G) < 4$. Taking into account Lemma 3, we may assume that each two 3-anticliques in G have a common vertex.

Case 1. There are two 3-anticliques in G that have exactly one common vertex. We put them to be the 3-anticliques $A = \{a, c, y\}$ and $B = \{a, b, z\}$. Consider the subgraph $G - a$. According to Lemma 4, this subgraph has a 3-anticlique $C = \{u, v, x\}$. Because the sets A and C could not be disjoint, as well as B and C, we may assume that $u = c$ and $v = b$, i.e. $C = \{c, b, x\}$. From the assumption $\alpha(G) < 4$ it follows that $x \neq z$, $x \neq y$, $[a, x] \in E(G)$, $[b, y] \in E(G)$ and $[c, z] \in E(G)$. From the assumption that there are not two disjoint 3-anticliques in G it follows that $[x, y], [x, z], [z, y] \in E(G)$. So we may see that in fact the subgraph generated by the vertices a, b, c, x, y, z coincides with the graph shown at Fig. 23 (the bold lines denote the edges of G and the thin lines — the ones of $\overline{G}$). Let u and v be the last two vertices of G. According to Lemma 2, $\max\{d(x), d(y), d(z)\} \leq 4$. From this inequality we conclude that none of x, y, z can be adjacent to both u and v, hence one of u and v is not adjacent to at least two of x, y, z. We assume without a loss of generality that $[u, x]$, $[u, y] \notin E(G)$.

Subcase 1.a. The vertex u is not adjacent to the vertex z. From $\text{cl}(G) = 3$ it follows that v is not adjacent at least to one of x, y, z. Because of the obvious symmetry we may assume that $[v, x] \notin E(G)$. In the subgraph $G - \{v, x\}$ there are no 3-cliques and consequently the graph G is not 3-saturated.

Subcase 1.b. The vertex u is adjacent to the vertex z. Because $d(z) \leq 4$ (Lemma 2), we have $[z, v] \notin E(G)$. In the subgraph $G - \{z, v\}$ there are no 3-cliques, which shows that G is not 3-saturated.

Case 2. Each two different 3-anticliques in G have two common vertices. Let $A = \{u, v, w\}$ be a 3-anticlique in G. According to Lemma 4, the subgraph $G - w$ contains a 3-anticlique B. Then $B = \{u, v, z\}$, since $|A \cap B| = 2$. Similarly, the subgraph $G - u$ contains 3-anticlique C that has two common vertices with A as well as with B. Then $C = \{z, v, w\}$ and $\{u, v, z, w\}$ is a 4-anticlique and the graph G is not 3-saturated according to Lemma 1. ∎

Lemma 6. *Let G be a 7-vertex graph, $\text{cl}(G) = 3$, $\alpha(G) = 2$ and $\Delta(G) \leq 4$. Then G is isomorphic to one of the graphs F_i, $i = 1, \ldots, 7$* (Fig. 16–22).

Proof. From $\alpha(G) = 2$ it follows that $\chi(G) \geq 4$. Because $\mathrm{cl}(G) = 3$, we may apply Theorem B with $r = 3$. From $\Delta(G) \leq 4$ it follows that the graph G contains no subgraph isomorphic to $K_1 + C_5$. The only possibility remaining is G to be isomorphic to one of F_i. ∎

Proof of Theorem 4. Theorem C implies that $\chi(G) \leq 4$ and from Proposition 5 we know that $\chi(G) \geq 4$. Consequently, $\chi(G) = 4$. According to Lemma 5, G is not a vertex-critical 4-chromatic graph, i.e. there is a vertex, say $v_8 \in V(G)$, such that $\chi(G - v_8) = 4$. We apply Theorem B with $r = 3$ to the subgraph $G - v_8$ to conclude that either $G - v_8$ is isomorphic to some of F_i, $i = 1, \ldots, 7$ (Fig. 16–22) or there is a $v_7 \in V(G)$ such that $G - \{v_7, v_8\} = K_1 + C_5$. Assume that there is no $v \in V(G)$ such that $G - v \neq \overline{C}_7 = F_7$. The above considerations show that there are the following possibilities:

Case 1. $G - v_8 = F_1$ (Fig. 16). We shall use the following automorphisms of the graph F_1:

$$\varphi(v_2) = v_6, \quad \varphi(v_4) = v_7, \quad \varphi(v_6) = v_2, \quad \varphi(v_7) = v_4, \quad \varphi(v_i) = v_i, \quad i = 1, 3, 5,$$
$$\psi(v_1) = v_7, \quad \psi(v_3) = v_6, \quad \psi(v_6) = v_3, \quad \psi(v_7) = v_1, \quad \psi(v_i) = v_i, \quad i = 2, 4, 5.$$

Subcase 1.a. The vertex v_8 is adjacent to at least one of the vertices v_2, v_3, v_6. Because $\varphi(v_6) = v_2$, $\psi(v_6) = v_3$, we may do assume without a loss of generality that v_8 is adjacent to v_6. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to at least one of v_2 and v_3. Because of the symmetry it is enough to consider the case $[v_3, v_8] \notin E(G)$. Certainly, $v_1 \in \mathrm{Ad}(v_8)$, since otherwise $\{v_1, v_3, v_8\}$ would be an anticlique and $G - \{v_1, v_3, v_8\}$ would not contain 3-cliques. From $\mathrm{cl}(G) = 3$ and $v_1, v_6 \in \mathrm{Ad}(v_8)$ it follows that $v_2 \notin \mathrm{Ad}(v_8)$. If we assume that $v_4 \notin \mathrm{Ad}(v_8)$, then $\{v_2, v_4, v_8\}$ is an anticlique and $G - \{v_2, v_4, v_8\}$ does not contain 3-cliques; and if we assume that $v_5 \notin \mathrm{Ad}(v_8)$, then $G - \{v_6, v_7\}$ does not contain 3-cliques. We have got a contradiction in both cases, which means that $v_4, v_5 \in \mathrm{Ad}(v_8)$. Consequently, either $\mathrm{Ad}(v_8) = \{v_1, v_4, v_6, v_5\}$ and G is isomorphic to L_4 (Fig. 5) or $\mathrm{Ad}(v_8) = \{v_1, v_4, v_7, v_5, v_6\}$ and G is isomorphic to L_6 (Fig. 7).

Subcase 1.b. The vertex v_8 is adjacent to none of v_2, v_3, v_6. If we assume that $v_1 \notin \mathrm{Ad}(v_8)$, then $\{v_1, v_3, v_8\}$ is an anticlique and $G - \{v_1, v_3, v_8\}$ does not contain 3-cliques; if $v_4 \notin \mathrm{Ad}(v_8)$, then $\{v_2, v_4, v_8\}$ is an anticlique and $G - \{v_2, v_4, v_8\}$ does not contain 3-cliques; if we assume that $v_5 \notin \mathrm{Ad}(v_8)$, $G - \{v_6, v_7\}$ does not contain 3-cliques, and if $v_7 \notin \mathrm{Ad}(v_8)$, then the subgraph $G - \{v_6, v_7, v_8\}$ does not contain 3-cliques. Thus we have proved that $\{v_1, v_4, v_5, v_7\} \subset \mathrm{Ad}(v_8)$. Because $v_2, v_3, v_6 \notin \mathrm{Ad}(v_8)$, we compute $\mathrm{Ad}(v_8) = \{v_1, v_4, v_5, v_7\}$, and G is isomorphic to the graph L_5 (Fig. 6).

Case 2. $G - v_8 = F_2$ (Fig. 17). We shall use the following automorphism of the graph F_2:

$$\varphi(v_1) = v_5, \quad \varphi(v_2) = v_4, \quad \varphi(v_3) = v_3, \quad \varphi(v_4) = v_2,$$
$$\varphi(v_5) = v_1, \quad \varphi(v_6) = v_7, \quad \varphi(v_7) = v_6.$$

The vertex v_8 is adjacent to at least one of the vertices v_6, v_7, since otherwise $\{v_6, v_7, v_8\}$ would be an anticlique and $G - \{v_6, v_7, v_8\}$ would contain no 3-clique.

Because of the certain symmetry ($\varphi(v_6) = v_7$) we may assume that $v_6 \in \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to the edges $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$. Because $G - \{v_6, v_7\}$ contains 3-cliques, we have two possibilities:

Subcase 2.a. The vertex v_8 is adjacent to the edge $[v_1, v_5]$. From $\mathrm{cl}(G) = 3$ it follows that $v_2 \notin \mathrm{Ad}(v_8)$. Certainly, $v_4 \in \mathrm{Ad}(v_8)$, since otherwise $\{v_2, v_4, v_8\}$ would be an anticlique and $G - \{v_2, v_4, v_8\}$ would contain no 3-clique. So, $\{v_1, v_4, v_5, v_6\} \subset \mathrm{Ad}(v_8)$. Because $\mathrm{cl}(G) = 3$, we have $\mathrm{Ad}(v_8) = \{v_1, v_4, v_5, v_6\}$. We see that $\alpha(G) = 2$ and then by Theorem A the graph G is isomorphic to the graph L_2 (Fig. 3).

Subcase 2.b. The vertex v_8 is adjacent to the edge $[v_4, v_5]$. From $\mathrm{cl}(G) = 3$ it follows that $v_3, v_7 \notin \mathrm{Ad}(v_8)$. If $v_1 \notin \mathrm{Ad}(v_8)$, then $G - \{v_2, v_4\}$ contains no 3-clique. If $v_1 \in \mathrm{Ad}(v_8)$, then as in subcase 2.a we conclude that the graph G is isomorphic to the graph L_2 (Fig. 3).

Case 3. $G - v_8 = F_3$ (Fig. 18). If $v_6, v_7 \notin \mathrm{Ad}(v_8)$, then $\{v_6, v_7, v_8\}$ is an anticlique and $G - \{v_6, v_7, v_8\}$ contains no 3-clique, which is a contradiction. Thus the vertex v_8 is adjacent to at least one of v_6, v_7. Because of the symmetry we may assume that $v_6 \in \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to the edges $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$. Because $G - \{v_6, v_7\}$ contains 3-cliques, v_8 is adjacent to at least one of the edges $[v_1, v_5]$, $[v_4, v_5]$.

Subcase 3.a. The vertex v_8 is adjacent to the edge $[v_1, v_5]$ and is not adjacent to the edge $[v_4, v_5]$, i.e. $v_1, v_5 \in \mathrm{Ad}(v_8)$ and $v_4 \notin \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ it follows that $v_2, v_7 \notin \mathrm{Ad}(v_8)$. So, $\{v_1, v_5, v_6\} \subset \mathrm{Ad}(v_8)$ and $v_2, v_4, v_7 \notin \mathrm{Ad}(v_8)$. That is why either $\mathrm{Ad}(v_8) = \{v_1, v_5, v_6\}$ or $\mathrm{Ad}(v_8) = \{v_1, v_5, v_6, v_3\}$. If $\mathrm{Ad}(v_8) = \{v_1, v_5, v_6\}$, then the graph G is isomorphic to the graph L_9 (Fig. 10). If $\mathrm{Ad}(v_8) = \{v_1, v_5, v_6, v_3\}$, then $\alpha(G - v_2) = 2$ and $\Delta(G - v_2) = \delta(G - v_2) = 4$. From Lemma 6 it follows that $G - v_2$ is isomorphic to some of the graphs F_i, $i = 1, \ldots, 7$. Because $\delta(F_i) = 3$ for $i = 1, \ldots, 6$, we have that $G - v_2 = \overline{C}_7 = F_7$, which contradicts the assumption at the top of the proof.

Subcase 3.b. The vertex v_8 is adjacent to the edge $[v_4, v_5]$ and is not adjacent to the edge $[v_1, v_5]$, i.e. $v_4, v_5 \in \mathrm{Ad}(v_8)$ and $v_1 \notin \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ it follows that $v_3, v_7 \notin \mathrm{Ad}(v_8)$. If $v_2 \in \mathrm{Ad}(v_8)$, then $G - v_6$ is isomorphic to the graph F_1 and we are back to the case 1. If $v_2 \notin \mathrm{Ad}(v_8)$, then G is isomorphic to the graph L_{10} (Fig. 11).

Subcase 3.c. The vertex v_8 is adjacent to the both edges $[v_1, v_5]$ and $[v_4, v_5]$. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to any of v_2, v_3, v_7. We take the conclusion that G is isomorphic to the graph L_{11} (Fig. 12).

Case 4. $G - v_8 = F_4$ (Fig. 19). We use the following automorphism of the graph F_4:

$$\varphi(v_1) = v_5, \quad \varphi(v_2) = v_7, \quad \varphi(v_3) = v_3, \quad \varphi(v_4) = v_6,$$
$$\varphi(v_5) = v_1, \quad \varphi(v_6) = v_4, \quad \varphi(v_7) = v_2.$$

We consider three subcases:

Subcase 4.a. The vertices $v_4, v_6 \in \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ and $v_6 \in \mathrm{Ad}(v_8)$ it follows that the vertex v_8 is not adjacent to the edges $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_4]$. From this fact we conclude that $v_5 \in \mathrm{Ad}(v_8)$ (otherwise v_8 is not adjacent to any of the edges of the 5-cycle v_1, v_2, v_3, v_4, v_5, v_1 and $G - \{v_6, v_7\}$ contains no 3-cliques). From $\mathrm{cl}(G) = 3$ and $v_4 \in \mathrm{Ad}(v_8)$ it follows that the vertex v_8 is not adjacent to the edges $[v_3, v_6]$, $[v_3, v_7]$, $[v_7, v_5]$. Hence $v_1 \in \mathrm{Ad}(v_8)$ (otherwise v_8 is not adjacent to any of the edges of the 5-cycle v_1, v_6, v_3, v_7, v_5, v_1 and $G - \{v_2, v_4\}$ contains no 3-cliques). So, $\{v_1, v_4, v_5, v_6\} \subset \mathrm{Ad}(v_8)$. Because $\mathrm{cl}(G) = 3$, we compute $\mathrm{Ad}(v_8) = \{v_1, v_4, v_5, v_6\}$, and G is isomorphic to the graph L_7 (Fig. 8).

Subcase 4.b. The vertex v_8 is adjacent to only one of the vertices v_4, v_6. Because of the certain symmetry ($\varphi(v_6) = v_4$) we may assume that $v_6 \in \mathrm{Ad}(v_8)$ and $v_4 \notin \mathrm{Ad}(v_8)$. If $v_2 \notin \mathrm{Ad}(v_8)$, then $\{v_2, v_4, v_8\}$ is an anticlique and $G - \{v_2, v_4, v_8\} = C_5$ contains no 3-cliques — a contradiction. If $v_2 \in \mathrm{Ad}(v_8)$, then from $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to v_1 and v_3. Since v_8 is not adjacent also to v_4, we have that v_8 is not adjacent to any of the edges of the 5-cycle v_1, v_2, v_3, v_4, v_5, v_1. This is a contradiction, because $G - \{v_6, v_7\}$ does not contain 3-cliques.

Subcase 4.c. The vertices $v_4, v_6 \notin \mathrm{Ad}(v_8)$. Certainly, $v_2, v_7 \in \mathrm{Ad}(v_8)$: if $v_2 \notin \mathrm{Ad}(v_8)$, then $G - \{v_2, v_4, v_8\}$ contains no 3-cliques; if $v_7 \notin \mathrm{Ad}(v_8)$, then $G - \{v_6, v_7, v_8\}$ contains no 3-cliques. If v_8 is adjacent to the edge $[v_1, v_5]$, then $\alpha(G) = 2$ and from Theorem A it follows that the graph G is isomorphic to the graph L_1 (Fig. 2). Assume now that the vertex v_8 is not adjacent to the edge $[v_1, v_5]$. Then either $v_1 \notin \mathrm{Ad}(v_8)$ or $v_5 \notin \mathrm{Ad}(v_8)$. From the reasons of symmetry ($\varphi(v_1) = v_5$) we may assume that $v_1 \notin \mathrm{Ad}(v_8)$. The subgraph $G - \{v_6, v_7\}$ contains a 3-clique and thus $v_3 \in \mathrm{Ad}(v_8)$. If $v_5 \notin \mathrm{Ad}(v_8)$, then $\mathrm{Ad}(v_8) = \{v_2, v_3, v_7\}$, and G is isomorphic to the graph L_8 (Fig. 9). If $v_5 \in \mathrm{Ad}(v_8)$, then the subgraph $G - v_7$ is isomorphic to F_1, which is the case 1.

Case 5. $G - v_8 = F_5$ (Fig. 20). We consider the following two possibilities:

Subcase 5.a. The vertex $v_6 \notin \mathrm{Ad}(v_8)$. Here we surely have $v_5, v_7 \in \mathrm{Ad}(v_8)$: if $v_5 \notin \mathrm{Ad}(v_8)$, then $\{v_5, v_6, v_8\}$ is an anticlique and $G - \{v_5, v_6, v_8\}$ contains no 3-cliques; if $v_7 \notin \mathrm{Ad}(v_8)$, then $\{v_6, v_7, v_8\}$ is an anticlique and $G - \{v_6, v_7, v_8\}$ contains no 3-cliques. From $\mathrm{cl}(G) = 3$ it follows that $v_1, v_4 \notin \mathrm{Ad}(v_8)$. Because $G - \{v_6, v_7\}$ contains a 3-clique, the vertex v_8 is adjacent to the edge $[v_2, v_3]$. Thus the subgraph $G - v_7$ is isomorphic to F_1, which is the case 1.

Subcase 5.b. The vertex $v_6 \in \mathrm{Ad}(v_8)$. From $\mathrm{cl}(G) = 3$ it follows that the vertex v_8 is not adjacent to the edges $[v_1, v_2]$, $[v_2, v_3]$ and $[v_3, v_4]$. Because $G - \{v_6, v_7\}$ contains a 3-clique, the vertex v_8 is adjacent to at least one of the edges $[v_1, v_5]$ and $[v_4, v_5]$. For the symmetry we may assume that v_8 is adjacent to the edge $[v_1, v_5]$. From $\mathrm{cl}(G) = 3$ we have that $v_7 \notin \mathrm{Ad}(v_8)$ and thus v_8 is not adjacent to the edges $[v_1, v_7]$ and $[v_4, v_7]$. But v_8 is not adjacent also to the edges $[v_1, v_2]$, $[v_2, v_3]$ and $[v_3, v_4]$ and the subgraph $G - \{v_5, v_6\}$ contains no 3-cliques, a contradiction.

Case 6. $G - v_8 = F_6$ (Fig. 21). We shall use the following automorphisms of the graph F_6:

$$\varphi(v_2) = v_6, \quad \varphi(v_6) = v_2, \quad \varphi(v_5) = v_7, \quad \varphi(v_7) = v_5, \quad \varphi(v_i) = v_i \quad i = 1, 3, 4,$$

$$\psi(v_1) = v_1, \quad \psi(v_2) = v_5, \quad \psi(v_3) = v_4, \quad \psi(v_4) = v_3,$$

$$\psi(v_5) = v_2, \quad \psi(v_6) = v_7, \quad \psi(v_7) = v_6,$$

$$\nu(v_1) = v_1, \quad \nu(v_2) = v_7, \quad \nu(v_3) = v_4, \quad \nu(v_4) = v_3,$$

$$\nu(v_5) = v_6, \quad \nu(v_6) = v_5, \quad \nu(v_7) = v_2.$$

Subcase 6.a. The vertex v_8 is not adjacent to some of the vertices v_2, v_5, v_6, v_7. Because of the symmetry ($\varphi(v_2) = v_6$, $\psi(v_2) = v_5$, $\nu(v_2) = v_7$) it is enough to consider only the situation when $v_2 \notin \mathrm{Ad}(v_8)$. In this situation certainly $v_5, v_7 \in \mathrm{Ad}(v_8)$ (if $v_5 \notin \mathrm{Ad}(v_8)$, then $\{v_2, v_5, v_8\}$ is an anticlique and $G - \{v_2, v_5, v_8\}$ contains no 3-clique; if $v_7 \notin \mathrm{Ad}(v_8)$, then $\{v_2, v_7, v_8\}$ is an anticlique and $G - \{v_2, v_7, v_8\}$ contains no 3-clique). From $\mathrm{cl}(G) = 3$ and $v_5, v_7 \in \mathrm{Ad}(v_8)$ it follows that $v_1, v_4 \notin \mathrm{Ad}(v_8)$. The subgraph $G - \{v_6, v_7\}$ contains no 3-cliques, a contradiction.

Subcase 6.b. The vertex v_8 is adjacent to all vertices v_2, v_5, v_6, v_7. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to the vertices v_1, v_3, v_4. We get the conclusion that the subgraph $G - \{v_6, v_7\}$ contains no 3-cliques, a contradiction.

Case 7. There are $v_7, v_8 \in V(G)$ such that $G - \{v_7, v_8\} = K_1 + C_5$. Let $V(K_1) = \{v_6\}$ and $C_5 = v_1, v_2, v_3, v_4, v_5, v_1$. From Lemma 2 and the fact that $d(v_6) \geq 5$ we conclude that $\Delta(G) = 5$ and $v_7, v_8 \notin \mathrm{Ad}(v_6)$. Certainly, $[v_7, v_8] \in E(G)$ (or, otherwise, $\{v_6, v_7, v_8\}$ is an anticlique and $G - \{v_6, v_7, v_8\}$ contains no 3-clique). If we assume that $\alpha(G - v_7) = \alpha(G - v_8) = 2$, then from $[v_7, v_8] \in E(G)$ it follows that $\alpha(G) = 2$ and according to Theorem A the graph G is isomorphic to some of L_1, L_2, L_3.

Let us now assume that at least one of the numbers $\alpha(G - v_7)$, $\alpha(G - v_8)$ is bigger than 2. Without a loss of generality, $\alpha(G - v_7) > 2$. This means that the vertex v_7 together with two non-adjacent vertices of the cycle v_1, v_2, v_3, v_4, v_5, v_1 form a 3-anticlique. Let, for example, $\{v_3, v_5, v_7\}$ be a 3-anticlique. Then $v_1, v_2 \in \mathrm{Ad}(v_7)$, since $G - \{v_6, v_8\}$ contains a 3-clique. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to at least one of the vertices v_1, v_2. Let v_8 be non-adjacent to v_1.

Assume first that v_8 is not adjacent also to v_2. Put $V_5(G) = \{v \in V(G) \mid d(v) = 5\}$. Then $V_5(G) \subset \{v_4, v_6\}$. Because $G - \{v_6, v_7\}$ contains a 3-clique, it follows that the vertex v_8 is adjacent to at least one of the edges $[v_3, v_4]$, $[v_4, v_5]$. For the symmetry we may assume that v_8 is adjacent to the edge $[v_4, v_5]$. Then $\alpha(G - v_3) = 2$ and from $V_5(G) \subset \{v_4, v_6\}$ it folows that $\Delta(G - v_3) = 4$. According to Lemma 6, $G - v_3$ is isomorphic to some of the graphs F_i for $i = 1, \ldots, 7$. By our assumption $G - v_3 \neq F_7$ and thus we turn to one of the cases 1–6.

Assume now that v_8 is adjacent to v_2.

Subcase 7.a. The vertex $v_4 \in \mathrm{Ad}(v_8)$. It is clear that $\alpha(G - v_3) = 2$. Note that $\Delta(G - v_3) = 4$ since $V_5(G) \subset \{v_2, v_4, v_6, v_8\}$. According to Lemma 6, $G - v_3$ is isomorphic to some of the graphs F_i for $i = 1, \ldots, 7$. By our assumption $G - v_3 \neq F_7$ and thus we turn to one of the cases 1–6.

Subcase 7.b. The vertex $v_4 \notin \mathrm{Ad}(v_8)$. Because we have also $v_1 \notin \mathrm{Ad}(v_8)$, the vertex v_8 is adjacent to none of the edges $[v_1, v_5]$, $[v_1, v_2]$, $[v_3, v_4]$, $[v_4, v_5]$. But the subgraph $G - \{v_6, v_7\}$ contains a 3-clique and therefore the vertex v_8 is adjacent to the edge $[v_2, v_3]$. So, we proved that the vertex v_7 is adjacent to the edge $[v_1, v_2]$ and, eventually, to the vertex v_4, and the vertex v_8 is adjacent to the edge $[v_2, v_3]$ and, eventually, to the vertex v_5. Now, if $v_4 \notin \mathrm{Ad}(v_7)$ and $v_5 \notin \mathrm{Ad}(v_8)$, then the graph G is isomorphic to the graph L_{12} (Fig. 13). If $v_4 \in \mathrm{Ad}(v_7)$ and $v_5 \notin \mathrm{Ad}(v_8)$ or $v_4 \notin \mathrm{Ad}(v_7)$ and $v_5 \in \mathrm{Ad}(v_8)$, then the graph G is isomorphic to the graph L_{13}

(Fig. 14). If $v_4 \in \mathrm{Ad}(v_7)$ and $v_5 \in \mathrm{Ad}(v_8)$, then the graph G is isomorphic to the graph L_{14} (Fig. 15). ■

Proof of Theorem 5. We fix some notations:

$e(G) = |E(G)|$,

$t(G)$ is the number of the 3-cliques of the graph G,

$\bar{t}(G)$ is the number of the 3-anticliques of the graph G,

$n(G)$ is the number of the pairs of 3-anticliques that have only one common vertex,

$m(G)$ is the number of the pairs of 3-anticliques that have no common vertex;

$i =$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$e(L_i)$	16	17	18	16	16	17	16	15	16	16	17	15	16	17
$t(L_i)$	8	10	12	8	7	9	9	7	8	8	10	8	8	8
$\bar{t}(L_i)$	0	0	0	1	1	1	1	2	2	2	1	2	2	2
$n(L_i)$	0	0	0	0	0	0	0	1	0	1	0	0	0	0
$m(L_i)$	0	0	0	0	0	0	0	0	0	0	0	1	1	1

From these relations we see that each two of the graphs L_i, $i = 1, \ldots, 14$, are not isomorphic. As $\alpha(L_i) \leq 3$, for proving that the graphs L_i, $i = 1, \ldots, 14$, are 3-saturated, we need to show that:

(1) $t(L_i - v) \geq 1$ for an arbitrary $v \in L_i$, $i = 1, \ldots, 14$;

(2) $t(L_i - \{u, v\}) \geq 1$, $i = 1, \ldots, 14$, for each two non-adjacent vertices u and v from L_i;

(3) $t(L_i - \{u, v, w\}) \geq 1$, $i = 1, \ldots, 14$, for an arbitrary 3-anticlique $\{u, v, w\}$ of L_i.

We need the following assertions:

Proposition 7 ([2], see also [7]). *Let $|V(G)| = 6$. Then $t(G) + \bar{t}(G) \geq 2$.*

Proposition 8 ([4], see also [7]). *Let $|V(G)| = 6$, $\bar{t}(G) = 2$ and the both 3-anticliques of G have only one common vertex. Then $t(G) \geq 1$.*

For arbitrary $i = 1, \ldots, 14$ and for arbitrary vertex of L_i there is non-adjacent vertex of L_i, therefore (2) implies (1). Because $\bar{t}(L_i) \leq 2$, the check of (3) is easy. We only show the 3-anticliques of the graphs L_i. The graphs L_1, L_2 and L_3 have not 3-anticliques. The graphs L_4, L_5 and L_6 have the unique 3-anticlique $\{v_1, v_4, v_7\}$. The graph L_7 has the unique 3-anticlique $\{v_2, v_7, v_8\}$. The graph L_8 has two 3-anticliques — $\{v_1, v_4, v_8\}$ and $\{v_5, v_6, v_8\}$. The graph L_9 has two 3-anticliques — $\{v_2, v_4, v_8\}$ and $\{v_2, v_7, v_8\}$. The graph L_{10} has two 3-anticliques — $\{v_1, v_3, v_8\}$ and $\{v_2, v_7, v_8\}$. The graph L_{11} has the unique 3-anticlique $\{v_2, v_7, v_8\}$. Each of the graphs L_{12}, L_{13} and L_{14} has only these two 3-anticliques — $\{v_3, v_5, v_7\}$ and $\{v_1, v_4, v_8\}$.

We now show the inequalities (2). If $i = 1, 2, 3, 4, 5, 6, 7, 11$, then $\bar{t}(L_i) \leq 1$ and the inequality (2) follows from Proposition 7. Let $i = 8, 10$. If at least one of the vertices u, v is a vertex of a 3-anticlique of the graph L_i, then (2) follows from Proposition 7. If none of the vertices u, v is a vertex of a 3-anticlique of the graph L_i, then the subgraph $L_i - \{u, v\}$ satisfies the conditions of Proposition 8 and hence (2) is satisfied. Let $i = 9$. The graph L_9 has only the 3-anticliques $\{v_2, v_4, v_8\}$ and $\{v_2, v_7, v_8\}$. If $u, v \in V(L_9)$, $[u, v] \notin E(L_9)$, and at least one of the vertices u, v is a vertex of a 3-anticlique, then $\bar{t}(L_9 - \{u, v\}) \leq 1$ and the inequality $t(L_9 - \{u, v\}) \geq 1$ follows from Proposition 7. If none of the vertices u, v is a vertex of a 3-anticlique, then the pair $\{u, v\}$ coincedes with one of the following pairs of non-adjacent vertices of L_9: $\{v_1, v_3\}$, $\{v_5, v_6\}$, $\{v_3, v_5\}$, for which (2) is obvious.

Consider the graphs L_{12}, L_{13} and L_{14}. It is enough to prove (2) for L_{12}, since L_{12} is a subgraph of L_{13} and L_{14}. The only vertices of L_{12} that do not take part in 3-anticliques are v_2 and v_6. Since v_2 and v_6 are adjacent vertices of the graph L_{12}, if the vertices u and v are not adjacent, it follows that one of them is a vertex of a 3-anticlique of L_{12}. Therefore $\bar{t}(L_{12} - \{u, v\}) \leq 1$. From Proposition 7 we get $t(L_{12} - \{u, v\}) \geq 1$.

We can see that $L_i - v \neq \overline{C}_7$, $\forall v \in V(L_i)$, comparing the inequalities $\delta(L_i - v) \leq 3$ and $\delta(\overline{C}_7) = 4$. ■

4. PROOFS OF THEOREMS 1 AND 2

Proof of Theorem 1. Assume that $|V(G)| \leq 8$. By adding if necessary isolated vertices, we may consider only the case $|V(G)| = 8$. According to Proposition 3, we have $\chi(G) \geq 5$. We apply Theorem B ($r = 4$) to conclude that either $G = K_1 + F_i$, $i = 1, \dots, 7$, or there exists $v \in V(G)$ such that $G - v = K_2 + C_5$. We are going to prove that in the second case we can also find a vertex that is adjacent to all other vertices of the graph G. Let $G - v = K_2 + C_5$ and $V(K_2) = \{x, y\}$. If the vertex v is not adjacent to the edge $[x, y]$, then $\{x, y, v\} \cup V(C_5)$ is a 3-cliques free 2-partition of the vertices of G, which is impossible. Hence the vertex v is adjacent to the edge $[x, y]$ and then x is adjacent to all other vertices of the graph G. So, if the graph G satisfies the conditions of Theorem 1, then there is a vertex $v_0 \in V(G)$ which is adjacent to all other vertices of the graph G. Proposition 4 implies that $G - v_0$ is a 3-saturated 7-vertex graph. It is clear that $\mathrm{cl}(G - v_0) = 3$. According to Theorem 3, $G - v_0 = \overline{C}_7$ and since v_0 is adjacent to all vertices of $\overline{C}_7$, it follows that $G = K_1 + \overline{C}_7$. ■

We need the next lemmas.

Lemma 7. *Let A be an anticlique of the graph G, $G_1 = G - A$; and $V(G_1) = B \cup C$ be a 3-cliques free 2-partition of vertices of G_1 such that: each vertex of A, that is adjacent to some edge of the subgraph $\langle B \rangle$, is not adjacent to any edge of the subgraph $\langle C \rangle$. Then G has a 3-cliques free 2-partition of vertices.*

Proof. Let $A_1 = \{v \in A \mid v \text{ is not adjacent to any edge of } \langle B \rangle\}$. Put $V_1 = A_1 \cup B$ and $V_2 = (A \setminus A_1) \cup C$. Consider the 2-partition $V(G) = V_1 \cup V_2$, $V_1 \cap V_2 = \varnothing$. It is clear that V_1 does not contain 3-cliques of the graph G. If $v \in V_2 \cap A$, then v is adjacent to some edge of the subgraph $\langle B \rangle$ and therefore is not adjacent to any edge of the subgraph $\langle C \rangle$. That is why V_2 does not contain 3-cliques, too. ■

Lemma 8. *Let G be a graph, $|V(G)| = n$, $\mathrm{cl}(G) = 3$, and A be an anticlique of G, $|A| = n - 8$. Put $G_1 = G - A$. If $G \to (3,3)$, then either $G_1 = L_{14}$* (Fig. 15) *or there exists $v \in V(G_1)$ such that $G_1 - v = \overline{C}_7$.*

Proof. According to Proposition 4, the subgraph G_1 is a 3-saturated graph. Since $|V(G_1)| = 8$ and $\mathrm{cl}(G) = 3$, we can apply Theorem 4 to the subgraph G_1. If we assume that the assertion of Lemma 8 is false, then G_1 is isomorphic to one of the graphs L_i, $i = 1, \ldots, 13$. We shall consider all these cases:

Case 1. G_1 is some of the graphs L_1, L_2, L_3. We put $B = \{v_3, v_4, v_7, v_8\}$ and $C = \{v_1, v_2, v_5, v_6\}$. For any of L_1, L_2, L_3 we have $E(\langle B\rangle) = \{[v_3, v_4], [v_7, v_8]\}$.

For any of L_1, L_2, L_3 it is true that each edge of $\langle C\rangle$ belongs either to $E(\langle \mathrm{Ad}(v_3)\rangle)$ or to $E(\langle \mathrm{Ad}(v_4)\rangle)$. Therefore, if we assume that some $v \in A$ is adjacent to the edge $[v_3, v_4]$, then $\mathrm{cl}(G) = 3$ implies that v is not adjacent to any of the edges of $\langle C\rangle$. Similarly, if some $v \in A$ is adjacent to $[v_7, v_8]$, then v is not adjacent to any of the edges of $\langle C\rangle$. We see from Lemma 7 that G has a 3-cliques free 2-partition of vertices, which is a contradiction.

Case 2. G_1 is some of the graphs L_4, L_5, L_6. We put $B = \{v_2, v_3, v_5, v_8\}$ and $C = \{v_1, v_4, v_6, v_7\}$. For any of L_4, L_5, L_6 we have $E(\langle B\rangle) = \{[v_2, v_3], [v_5, v_8]\}$ and $E(\langle C\rangle) = \{[v_1, v_6], [v_4, v_6]\}$. If some of the vertices of the anticlique A is adjacent to the edge $[v_2, v_3]$, then $\mathrm{cl}(G) = 3$ implies that this vertex is not adjacent to the edges $[v_1, v_6]$, $[v_4, v_6]$, i.e. it is not adjacent to any of the edges of $\langle C\rangle$. If any of the vertices of the anticlique A is adjacent to the edge $[v_5, v_8]$, then from $\mathrm{cl}(G) = 3$ it follows that this vertex is not adjacent to the vertices v_1 and v_4. Consequently, it is not adjacent to the edges $[v_1, v_6]$ and $[v_4, v_6]$ of the subgraph $\langle C\rangle$. We see then from Lemma 7 that G has a 3-cliques free 2-partition of vertices, which is a contradiction.

Case 3. G_1 is some of the graphs L_7, L_8, L_{10} L_{11}. We put $B = \{v_1, v_3, v_4\}$ and $C = \{v_2, v_5, v_6, v_7, v_8\}$. For any of L_7, L_8, L_{10}, L_{11} we have $E(\langle B\rangle) = \{[v_3, v_4]\}$. Also, for L_7, L_{10}, L_{11} we denote $E_1 = E(\langle C\rangle) = \{[v_2, v_6], [v_6, v_8], [v_8, v_5], [v_5, v_7]\}$, and for L_8 — $E_2 = E(\langle C\rangle) = \{[v_2, v_6], [v_2, v_8], [v_8, v_7], [v_5, v_7]\}$.

Let the vertex $u \in A$ be adjacent to the edge $[v_3, v_4]$. For the graphs L_7, L_{10}, L_{11} we have that $\{v_2, v_6\} \subset \mathrm{Ad}(v_3)$ and $\{v_5, v_6, v_7, v_8\} \subset \mathrm{Ad}(v_4)$. Therefore $\mathrm{cl}(G) = 3$ implies that the vertex u is not adjacent to any of the edges from E_1. For the graph L_8 we have that $\{v_2, v_6, v_7, v_8\} \subset \mathrm{Ad}(v_3)$ and $\{v_5, v_7\} \subset \mathrm{Ad}(v_4)$. Therefore $\mathrm{cl}(G) = 3$ implies that the vertex u is not adjacent to any of the edges from E_2.

So, the conditions of Lemma 7 are satisfied and we conclude that in the considered case the graph G has a 3-cliques free 2-partition of vertices, which is a contradiction.

Case 4. G_1 coincides with the graph L_9. We put $B = \{v_1, v_3, v_4, v_8\}$ and $C = \{v_2, v_5, v_6, v_7\}$. We have that $E(\langle B\rangle) = \{[v_1, v_8], [v_3, v_4]\}$. If some of the vertices of the anticlique A is adjacent to the edge $[v_1, v_8]$, then $\mathrm{cl}(G) = 3$ and $C \subset \mathrm{Ad}(v_1)$ imply that this vertex is not adjacent to the edges $[v_2, v_6]$ and $[v_5, v_7]$, i.e. it is not adjacent to any of the edges of $\langle C\rangle$. If the anticlique A contains a vertex that is adjacent to the edge $[v_3, v_4]$, then from $\mathrm{cl}(G) = 3$ it follows that this vertex is not adjacent to the edges $[v_2, v_6]$, $[v_5, v_7]$, i.e. it is not adjacent to the edges of $\langle C\rangle$. We see from Lemma 7 that G has a 3-cliques free 2-partition of vertices, which is a contradiction.

Case 5. G_1 is some of the graphs L_{12}, L_{13}. We put $B = \{v_1, v_2, v_4\}$ and $C = \{v_3, v_5, v_6, v_7, v_8\}$. We have that $E(\langle B\rangle) = \{[v_1, v_2]\}$ and $E(\langle C\rangle) = \{[v_7, v_8]$, $[v_3, v_8]$, $[v_3, v_6]$, $[v_5, v_6]\}$. Let some of the vertices of the anticlique A be adjacent to the edge $[v_1, v_2]$. From $\text{cl}(G) = 3$ and $\{v_3, v_6, v_7, v_8\} \subset \text{Ad}(v_2)$ it follows that this vertex is not adjacent to the edges $[v_7, v_8]$, $[v_3, v_8]$ and $[v_3, v_6]$; from $\text{cl}(G) = 3$ and $\{v_5, v_6\} \subset \text{Ad}(v_1)$ it follows that this vertex is not adjacent to the edge $[v_5, v_6]$.

The above reasoning shows that the conditions of Lemma 7 are satisfied and we conclude that the graph G has a 3-cliques free 2-partition of vertices, which is a contradiction. ■

Lemma 9. *Let G be an 11-vertex graph, $\text{cl}(G) = 3$, and G have three 3-anticliques, each two of which have an empty intersection. Then the graph G has a 3-cliques free 2-partition of vertices.*

Proof. Let A, B and C be the anticliques given by the condition. Assume the contrary, i.e. $G \to (3,3)$. We put $G_1 = G - A$. Because G_1 has two anticliques B and C with empty intersection and $\alpha(\overline{C}_7) = 2$, we have that $G_1 - v \neq \overline{C}_7$, $\forall v \in V(G_1)$. From Lemma 7 it follows that $G_1 = L_{14}$ (Fig. 15). Let $A = \{v_9, v_{10}, v_{11}\}$. At least one of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_2, v_6]$ (if not, $\{v_2, v_6, v_9, v_{10}, v_{11}\} \cup \{v_1, v_7, v_8, v_3, v_4, v_5\}$ is a 3-cliques free 2-partition of the vertices of G). Thus we assume that v_9 is adjacent to $[v_2, v_6]$. At least one of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_1, v_2]$ (if not, $\{v_1, v_2, v_4, v_9, v_{10}, v_{11}\} \cup \{v_3, v_5, v_6, v_7, v_8\}$ is a 3-cliques free 2-partition of the vertices of G). The vertex v_9 is not adjacent to the edge $[v_1, v_2]$, since otherwise $\{v_1, v_2, v_6, v_9\}$ would be a 4-clique. Hence we may assume that v_{10} is adjacent to the edge $[v_1, v_2]$. Surely, one of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_1, v_6]$ (if not, $\{v_1, v_6, v_8, v_9, v_{10}, v_{11}\} \cup \{v_2, v_3, v_4, v_5, v_7\}$ is a 3-cliques free 2-partition of the vertices of G). $\text{cl}(G) = 3$ implies that both vertices v_9 and v_{10} are not adjacent to the edge $[v_1, v_6]$, thus v_{11} is adjacent to the edge $[v_1, v_6]$.

Consider the 2-partition $V(G) = V_1 \cup V_2$, where $V_1 = \{v_6, v_7, v_8, v_{10}\}$ and $V_2 = \{v_1, v_2, v_3, v_4, v_5, v_9, v_{11}\}$. Since v_{10} is adjacent to the vertex v_2 and $\text{cl}(G) = 3$, the vertex v_{10} is not adjacent to the edge $[v_7, v_8]$. That is why V_1 contains no 3-cliques. From $\text{cl}(G) = 3$ and the fact that v_9 is adjacent to the edge $[v_2, v_6]$ it follows that v_9 is not adjacent neither to the vertices v_1, v_3 nor to the edge $[v_4, v_5]$. Thus v_9 is not adjacent to any of the edges of the 5-cycle v_1, v_2, v_3, v_4, v_5, v_1. From $\text{cl}(G) = 3$ and the fact that v_{11} is adjacent to $[v_1, v_6]$ it follows that v_{11} is not adjacent neither to the vertices v_2 and v_5 nor to the edge $[v_3, v_4]$. This shows that v_{11} is adjacent to none of the edges of the 5-cycle v_1, v_2, v_3, v_4, v_5, v_1. Since v_9 and v_{11} are not adjacent, V_2 does not contain 3-cliques. We have proved that $V(G) = V_1 \cup V_2$ is a 3-cliques free 2-partition of the vertices of G. This contradiction completes the proof. ■

Proof of Theorem 2. Assume the contrary, i.e. $G \to (3,3)$. According to Proposition 2, $\alpha(G) \geq 3$. Let $A = \{v_9, v_{10}, v_{11}\}$ be a 3-anticlique of G. Put $G_1 = G - A$; $V(G_1) = \{v_1, \dots, v_8\}$. Because L_{14} (Fig. 15) has two disjoint 3-anticliques, Lemma 9 implies that $G_1 \neq L_{14}$. From Lemma 8 it follows that there exists $v \in V(G_1)$ such that $G_1 - v = \overline{C}_7$. Let, for example, $G_1 - v_8 = \overline{C}_7 = F_7$ (Fig. 22).

We shall prove first that the vertex v_8 together with some two vertices of $\overline{C}_7$ form a 3-anticlique of the graph G. From $\mathrm{cl}(G) = 3$ it follows that the vertex v_8 is not adjacent to some of the vertices of $\overline{C}_7$. Let, for example, v_8 be not adjacent to v_1 (Fig. 22). If the vertex v_8 is not adjacent to v_2 or v_7, then $\{v_1, v_2, v_8\}$ or, respectively, $\{v_1, v_7, v_8\}$ is a 3-anticlique of G. If v_8 is adjacent to both v_2 and v_7, then $\mathrm{cl}(G) = 3$ implies that $\{v_4, v_5, v_8\}$ is a 3-anticlique of G.

So, we may assume that $\{v_1, v_2, v_8\}$ is a 3-anticlique of the graph G. From $\mathrm{cl}(G) = 3$ it follows that v_8 is not adjacent to one of the vertices of the 3-clique $\{v_3, v_5, v_7\}$. We shall consider the following two cases.

Case 1. The vertex v_8 is not adjacent to v_3 or v_7, for example v_8 is not adjacent to v_3. One of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_1, v_3]$ (if not, $\{v_1, v_2, v_3, v_8, v_9, v_{10}, v_{11}\} \cup \{v_4, v_5, v_6, v_7\}$ is a 3-cliques free 2-partition). Let, for example, v_9 be adjacent to the edge $[v_1, v_3]$. From $\mathrm{cl}(G) = 3$ it follows that $\{v_5, v_6, v_9\}$ is a 3-anticlique. One of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_1, v_6]$ (if not, $\{v_1, v_6, v_7, v_9, v_{10}, v_{11}\} \cup \{v_2, v_3, v_4, v_5, v_8\}$ is a 3-cliques free 2-partition). From $\mathrm{cl}(G) = 3$ it follows that v_9 is not adjacent to the edge $[v_1, v_6]$. Therefore we may assume that v_{10} is adjacent to the edge $[v_1, v_6]$. From $\mathrm{cl}(G) = 3$ it follows that $\{v_3, v_4, v_{10}\}$ is a 3-anticlique. We obtain that G contains the pairwise disjoint 3-anticliques $\{v_1, v_2, v_8\}$, $\{v_5, v_6, v_9\}$ and $\{v_3, v_4, v_{10}\}$, which contradicts Lemma 9.

Case 2. The vertex v_8 is not adjacent to v_5. Surely, one of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_1, v_6]$ (if not, $\{v_1, v_6, v_7, v_9, v_{10}, v_{11}\} \cup \{v_2, v_3, v_4, v_5, v_8\}$ is a 3-cliques free 2-partition). Let, for example, v_9 be adjacent to $[v_1, v_6]$. From $\mathrm{cl}(G) = 3$ it follows that $\{v_3, v_4, v_9\}$ is a 3-anticlique. One of the vertices v_9, v_{10}, v_{11} is adjacent to the edge $[v_2, v_4]$ (if not, $\{v_2, v_3, v_4, v_9, v_{10}, v_{11}\} \cup \{v_1, v_5, v_6, v_7, v_8\}$ is a 3-cliques free 2-partition). Because v_9 is adjacent to v_6 and $\mathrm{cl}(G) = 3$, we know that v_9 is not adjacent to the edge $[v_2, v_4]$. Consequently, we may assume that the vertex v_{10} is adjacent to the edge $[v_2, v_4]$. From $\mathrm{cl}(G) = 3$ it follows that $\{v_6, v_7, v_{10}\}$ is a 3-anticlique. We have obtained that G contains the pairwise disjoint 3-anticliques $\{v_1, v_2, v_8\}$, $\{v_3, v_4, v_9\}$ and $\{v_6, v_7, v_{10}\}$, which contradicts Lemma 9.

The proof of Theorem 2 is completed. ■

5. AN EXAMPLE

We consider the graph L_{14} (Fig. 15) and the following subsets of $V(L_{14})$: $M_1 = \{v_2, v_4, v_6, v_7\}$, $M_2 = \{v_2, v_5, v_6, v_8\}$, $M_3 = \{v_1, v_2, v_5, v_8\}$, $M_4 = \{v_3, v_5, v_6, v_8\}$, $M_5 = \{v_2, v_3, v_4, v_7\}$, $M_6 = \{v_1, v_4, v_6, v_7\}$, $M_7 = \{v_4, v_5, v_7, v_8\}$. We denote by Γ_2

the extension of the graph L_{14} that is obtained by adding to $V(L_{14})$ new 7 vertices $u_1, \ldots, u_7$, none of which are adjacent and such that $\mathrm{Ad}(u_i) = M_i$, $i = 1, \ldots, 7$.

Proposition 9. $\Gamma_2 \to (3,3)$ *and* $\mathrm{cl}(\Gamma_2) = 3$.

Proof. The equality $\mathrm{cl}(\Gamma_2) = 3$ is true, because $\mathrm{cl}(L_{14}) = 3$, $\{u_1, \ldots, u_7\}$ is an anticlique, and $\mathrm{Ad}(u_i)$ does not contain 3-cliques for $i = 1, \ldots, 7$.

Let $V(\Gamma_2) = V_1 \cup V_2$ be an arbitrary 2-partition of the vertices of Γ_2.

Case 1. v_2 and v_6 belong to only one of the sets V_1 and V_2, for example $v_2, v_6 \in V_1$. From $v_2, v_6 \in V_1$ it follows that at least one of the vertices v_7, v_8 belongs to V_2. Let, for example, $v_7 \in V_2$. From $v_2, v_6 \in V_1$ it follows also that at least one of the vertices v_4, v_5 belongs to V_2. Therefore we have only two possibilities:

Subcase 1.a. $v_4 \in V_2$. If $u_1 \in V_1$, then $\{u_1, v_2, v_6\}$ is a 3-clique of Γ_2, contained in V_1. If $u_1 \in V_2$, then $\{u_1, v_4, v_7\}$ is a 3-clique of Γ_2, contained in V_2.

Subcase 1.b. $v_5 \in V_2$. From $v_2, v_6 \in V_1$ it follows also that $v_1 \in V_2$. Let $v_8 \in V_1$. If $u_3 \in V_1$, then $\{u_3, v_2, v_8\}$ is a 3-clique of Γ_2, contained in V_1. If $u_3 \in V_2$, then $\{u_3, v_1, v_5\}$ is a 3-clique of Γ_2, contained in V_2. Assume that $v_8 \in V_2$. If $u_2 \in V_1$, then $\{u_2, v_2, v_6\}$ is a 3-clique of Γ_2, contained in V_1. If $u_2 \in V_2$, then $\{u_2, v_5, v_8\}$ is a 3-clique of Γ_2, contained in V_2.

Case 2. One of the vertices v_2, v_6 belongs to V_1 and the other one belongs to V_2. Let, for example, $v_2 \in V_1$, $v_6 \in V_2$.

Subcase 2.a. One of the vertices v_7, v_8 belongs to V_1, for example $v_7 \in V_1$. If $v_8 \in V_1$ or $v_1 \in V_1$, then V_1 will contain respectively the 3-clique $\{v_2, v_7, v_8\}$ or the 3-clique $\{v_1, v_2, v_7\}$. Therefore we assume that $v_1, v_8 \in V_2$.

Let $v_4 \in V_1$. If $u_6 \in V_1$, then $\{u_6, v_4, v_7\}$ is a 3-clique of Γ_2, contained in V_1. If $u_6 \in V_2$, then $\{u_6, v_1, v_6\}$ is a 3-clique of Γ_2, contained in V_2.

Let $v_4 \in V_2$. If $u_1 \in V_1$, then $\{u_1, v_2, v_7\}$ is a 3-clique of Γ_2, contained in V_1. If $u_1 \in V_2$, then $\{u_1, v_4, v_6\}$ is a 3-clique of Γ_2, contained in V_2.

Subcase 2.b. $v_7, v_8 \in V_2$. Assume first that at least one of the vertices v_4, v_5 belongs to V_2 and let, for example, $v_4 \in V_2$. If $v_3 \in V_2$, then $\{v_3, v_4, v_6\}$ is a 3-clique of Γ_2, contained in V_2. Thus we assume that $v_3 \in V_1$. Now, if $u_5 \in V_1$, then $\{u_5, v_2, v_3\}$ is a 3-clique of Γ_2, contained in V_1. If $u_5 \in V_2$, then $\{u_5, v_4, v_7\}$ is a 3-clique of Γ_2, contained in V_2.

Finally, we consider the case when $v_4, v_5 \in V_1$. If $u_7 \in V_1$, then $\{u_7, v_4, v_5\}$ is a 3-clique of Γ_2, contained in V_1. If $u_7 \in V_2$, then $\{u_7, v_7, v_8\}$ is a 3-clique of Γ_2, contained in V_2. ■

REFERENCES

1. Erdös, P., C. Rogers. The construction of certain graphs. *Canad. J. Math.*, **14**, 1962, 702–707.
2. Goodman, A. On sets of acquitances and strangers at any party. *Amer. Math. Montly*, **66**, 1959, 778–783.

3. Greenwood, R., A. Gleason. Combinatorial relation and chromatic graphs. *Canad. J. Math.*, **7**, 1955, 1–7.
4. Harary, F. The two-triangle case of acquitance graph. *Math. Magazine*, **45**, 1972, 130–135.
5. Irving, R. On a bound of Graham and Spencer for a graph coloring constant. *J. Combin. Theory*, Ser. B, **15**, 1973, 200–203.
6. Kery, G. Ramsey egy grafelmelety teteléröl. *Math. Lapok*, **15**, 1964, 204–224.
7. Khadjiivanov, N. Ramsey numbers. Narodna prosveta Press, Sofia, 1982.
8. Khadjiivanov, N. Extremal graph theory. "Kliment Ohridski" University Press, Sofia, 1990.
9. Nenov, N. An example of 15-vertex (3,3)-Ramsey graph with clique number 4. *C. R. Acad. Bulg. Sci.*, **34**, 1981, 1487–1489.
10. Nenov, N. Some applications of Zykov numbers to Ramsey theory. *Ann. Sof. Univ.*, Fac. Math., **74**, 1980, 29–50.
11. Nenov, N. The chromatic number of any 10-vertex graph without 4-cliques is at most 4. *C. R. Acad. Bulg. Sci.*, **37**, 1984, 301–304.
12. Nenov, N., N. Khadjiivanov. Any Ramsey graph without 5-cliques has more than 11 vertices. *Serdika*, **11**, 1985, 341–356. Erratum, *Serdika*, **12**, 1986, 204.
13. Nenov, N., N. Khadjiivanov. On the 4-anticliques of some graphs without triangles. *Ann. Sof. Univ.*, Fac. Math., **78**, 1984, 186–208.
14. Ramsey, F. On a problem of formal logic. *Proc. London Math. Soc.*, **30**, 1930, 264–286.

Received March 17, 1998

Section of Algebra
Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 Blvd James Bourchier
BG-1164 Sofia, Bulgaria
E-mail address: **nedialkov@fmi.uni-sofia.bg**

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 91, 1997

ON THE TWO-POINT CORRELATION FUNCTIONS IN RANDOM ARRAYS OF NONOVERLAPPING SPHERES

KONSTANTIN Z. MARKOV

For a random dispersion of identical spheres, the known two-point correlation functions like "particle-center," "center-surface," "particle-surface," etc., are studied. Geometrically, they give the probability density that two points, thrown at random, hit in various combinations a sphere's center, a sphere, or a sphere's surface. The basic result of the paper is a set of simple and integral representations of one and the same type for these correlations by means of the radial distribution function for the set of sphere's centers. The derivations are based on the geometrical reasoning, recently employed by Markov and Willis when studying the "particle-particle" correlation. An application, concerning the effective absorption strength of a random array of spherical sinks, is finally given.

Keywords: random media, dispersions of spheres, correlation variational bounds, absorption problem

1991/1995 Math. Subject Classification: 60G60, 60H15, 49K45

1. INTRODUCTION

In many cases of great practical interest the macroscopic behaviour of a two-phase medium is strongly influenced by the amount and the internal distribution of the interfacial surface. A classical problem of such a kind is supplied first of all by the theory of diffusion-controlled reactions, as initiated by Smoluchowski in 1916. Formally, this is equivalent to the problem, concerning a species (defects) diffusing in the presence of an array of ideally absorbing traps (sinks). Another classical problem is the quest for the permeability of porous solids. The reason is that in both problems the observed macroscopic response is ruled by the events that take place at the boundary between the phases: in the first case chemical reactants' encounter (or absorption of defects) happens there and in the second case the viscous fluid flows around the particle surfaces, where no-slip boundary condition is to be satisfied. Hence it is natural that in studying both these phenomena the interfacial statistics should essentially enter the appropriate theories. Perhaps the first example was provided by Doi [3] who derived bounds on both the effective sink strength and the permeability. These bounds were put on a firmer base and generalized by Torquato and co-authors [9, 16, 10, 1]. The bounds include integrals of the interfacial two-point statistical correlations, which later on were thoroughly studied within a more general framework by Torquato [15, 14]. An alternative approach in the absorption context has been proposed by Talbot and Willis [12] who, using a Hashin-Shtrikman's type variational principle, derived a bound on the effective sink strength for a dispersion of nonoverlapping spheres which eventually utilizes only an integral incorporating the total correlation function. At a first glance this bound is entirely different from Doi's one since no interfacial statistics is even mentioned in Talbot and Willis' reasoning. As we shall see below, the Talbot and Willis bound turns out, however, to be identical to that of Doi.

The evaluation of the interfacial statistical characteristics for realistic two-phase random models meets with considerable difficulties. Only for the simplest model of fully penetrable spheres (the Boolean model) the needed quantities can be comparatively easily evaluated, as done by Doi himself. For dispersions of nonoverlapping spheres — a model that very often is appropriate for particulate type media — such an evaluation is much more involving, and the reason can be well seen from the already mentioned paper of Torquato [15]. In the same paper the author notes that the needed interfacial correlations have a convolution structure which allows, in principle, to reduce them to single integrals containing the total correlation functions for the dispersions, provided the Fourier transform is employed in the statistically isotropic case. No further details are given in [14], however, apart from appropriate formulae valid for a dilute dispersion, and numerical results for the semi-empirical Verlet-Weis distribution [18], see also [13]. (Note that the dilute results have been derived by Berryman [2] by means of a different approach.)

In the recent paper [7], a simple geometrical reasoning was proposed, which allowed the authors to represent the two-point correlation function of the region, occupied by the spheres (that is, the "particle-particle" correlation), as a simple integral that contains the radial distribution function of the spheres. The aim of the present work is to demonstrate that the same geometrical reasoning can be straightforwardly applied when considering the two-point interfacial correlations, if combined with a formula, noted by Doi [3]. In this way the said correlations will be reduced to even simpler integrals of the *same* type as that for the "particle-particle" one. To accomplish this, the definitions of the three basic interfacial characteristics are first introduced in Section 2, preceded by that of the simple "particle-center" correlation. The investigation of the latter in Section 3 serves as a model for a similar treatment of the interfacial characteristics, performed in Sections 4–6. (The study of the "particle-center" correlation, detailed here, is outlined in the author's paper [6].) The formulae for all two-point correlations have a fully similar structure, which is summarized in Table 1 (Section 9). In Section 7 the first two moments of the various two-point correlations are directly evaluated by means of an alternative and simpler method which is applicable in the 2-D case as well. As an elementary application of the obtained formulae it is finally shown (Section 8) that the Doi's bound on the effective sink strength of the dispersion coincides with that of Talbot and Willis.

2. DEFINITIONS OF THE BASIC TWO-POINT STATISTICAL CHARACTERISTICS

Consider a dispersion of equal and nonoverlapping spheres of radius a in $\mathbb{R}^3$, whose centers form the random set of points $\{x_\alpha\}$. The assumption of statistical isotropy and homogeneity is adopted henceforth. Introduce after Stratonovich [11] the so-called random density field for the dispersion

$$\psi(x) = \sum_\alpha \delta(x - x_\alpha), \tag{2.1}$$

$\delta(x)$ is the Dirac delta-function. All multipoint moments of the field $\psi(x)$ can be easily expressed by means of the multipoint probability densities of the random set $\{x_\alpha\}$, but in what follows only the first two simplest formulae of this kind will be needed, namely,

$$\langle\psi(x)\rangle = n, \quad F^{\rm cc}(x) = \langle\psi(x)\psi(0)\rangle = n\delta(x) + n^2 g(x), \tag{2.2}$$

where n is the number density of the spheres, and $g(x) = g(r)$, $r = |x|$, is their radial distribution function, see [11]. The brackets $\langle\cdot\rangle$ signify ensemble averaging. Note that the assumption of nonoverlapping implies that $g(x) = 0$ if $|x| \le 2a$. The notation $F^{\rm cc}(x)$ in (2.2) is justified by the interpretation of the quantity $\langle\psi(x)\psi(0)\rangle$ — this is the "center-center" correlation, in the sense that it obviously gives the probability densities of finding centers of particles both at the origin and at the point x.

Let

$$I_1(x) = \begin{cases} 1, & \text{if } x \in \mathcal{K}_1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

be the characteristic function of the region $\mathcal{K}_1$, occupied by the spheres. Then

$$I_1(x) = (h_a * \psi)(x) = \int h_a(x-y)\psi(y)\,\mathrm{d}y, \quad I_1'(x) = \int h_a(x-y)\psi'(y)\,\mathrm{d}y, \tag{2.4}$$

where $\psi'(y) = \psi(y) - n$ is the fluctuating part of the field $\psi(y)$ and $h_a(y)$ is the characteristic function of a single sphere of radius a, located at the origin. All integrals hereafter are over the whole $\mathbb{R}^3$ and, as usual, $f * g$ denotes the convolution of the functions f and g. The simple integral representation (2.4), combined with the formulae (2.2), serves as a basis for evaluating the needed interfacial statistical characteristics in what follows. Its simplest consequence reads

$$\eta_1 = \langle I_1(x) \rangle = nV_a, \quad V_a = \tfrac{4}{3}\pi a^3, \tag{2.5}$$

having taken averages of both sides of (2.4); η_1 is the volume fraction of the spheres.

In turn, the two-point correlation most often used is

$$F^{\,\mathrm{pp}}(x) = \langle I_1(0) I_1(x) \rangle \,. \tag{2.6}$$

The interpretation of $\langle I_1(0)I_1(x)\rangle$ is obvious — this is the probability that two points, separated by the vector x, when thrown into the medium both fall within a sphere. That is why $\langle I_1(0)I_1(x)\rangle$ can be called "particle-particle" correlation, which explains its notation $F^{\,\mathrm{pp}}(x)$ in (2.6).

Before introducing the interfacial characteristics, it is noted that another correlation, closely related to $F^{\,\mathrm{pp}}(x)$, will be useful as well. This is the "particle-center" one

$$F^{\,\mathrm{pc}}(x) = \langle I_1(x)\psi(0) \rangle \,, \tag{2.7}$$

which obviously gives the probability that for a pair of points, separated by the vector x, one hits a sphere's center while the other falls into a sphere.

It is natural to represent the above introduced correlations as

$$F^{\,\mathrm{cc}}(x) = n^2 + \overline{F}^{\,\mathrm{cc}}(x), \quad F^{\,\mathrm{pc}}(x) = n\eta_1 + \overline{F}^{\,\mathrm{pc}}(x), \quad F^{\,\mathrm{pp}}(x) = \eta_1^2 + \overline{F}^{\,\mathrm{pp}}(x), \tag{2.8}$$

where, as it follows from (2.2), (2.4), (2.6) and (2.7) ,

$$\begin{aligned}
&\overline{F}^{\,\mathrm{cc}}(x) = \langle \psi'(0)\psi'(x) \rangle = n\delta(x) + n^2\nu_2(x), \\
&\overline{F}^{\,\mathrm{pc}}(x) = \langle I_1'(x)\psi'(0) \rangle = (h_a * \overline{F}^{\,\mathrm{cc}})(x) = nh_a(x) + n^2 \int h_a(x-y)\nu_2(y)\,\mathrm{d}y, \\
&\overline{F}^{\,\mathrm{pp}}(x) = \langle I_1'(x) I_1'(0) \rangle = \big(h_a * \overline{F}^{\,\mathrm{pc}}\big)(x) = \big(h_a * h_a * \overline{F}^{\,\mathrm{cc}}\big)(x).
\end{aligned} \tag{2.9}$$

Here

$$\nu_2(y) = g(y) - 1 \tag{2.10}$$

is the so-called binary (or total) correlation function for the dispersion. Due to the no long-range assumption, all $\nu_2(x)$, $\overline{F}^{\,\mathrm{cc}}(x)$, $\overline{F}^{\,\mathrm{pc}}(x)$ and $\overline{F}^{\,\mathrm{pp}}(x)$ vanish as $x \to \infty$, since the constants in the right-hand sides of (2.8) are just their long-range values.

Let us recall now the definitions of the interfacial correlations. The first one,

$$F^{\,\mathrm{sc}}(x) = \langle\, |\nabla I_1(x)|\, \psi(0) \rangle \,, \tag{2.11}$$

can be called "surface-center." Since $|\nabla I_1(x)|$ and $\psi(x)$ are delta-functions, the former concentrated over the surface $\partial\mathcal{K}_1$ of the spheres and the latter over the set $\{x_\alpha\}$, the interpretation of $F^{\mathrm{sc}}(x)$ is obvious — this is the probability that if two points, separated by the vector x, are thrown into the medium, one of them falls on the surface of a sphere, while the other hits a center x_α of a sphere. This interpretation explains the terminology used here (note that it differs from that used by Torquato [15], where (2.11) is called "surface-particle" correlation).

The second interfacial correlation is

$$F^{\mathrm{sp}}(x) = \langle\, |\nabla I_1(x)|\, I_1(0)\rangle \tag{2.12}$$

— obviously the "surface-particle" one. The reason is that it gives the probability that one of the two points, separated by the vector x, when thrown into the medium, falls on the surface of a sphere, and the other falls within a sphere. (Note again the difference in terminology used here: Torquato [15] calls (2.11) "surface-particle" correlation, while (2.12) is very closely connected to the "surface-void" correlation of Doi [3].)

Finally, let

$$F^{\mathrm{ss}}(x) = \langle\, |\nabla I_1(x)|\, |\nabla I_1(0)|\,\rangle \tag{2.13}$$

be the "surface-surface" correlation, which gives the probability that the two points, separated by the vector x, thrown into the medium, both fall on the spheres' surfaces. (The terminology agrees here with that of Doi [3] and Torquato [15].)

Let now $h_b(x)$ be the characteristic function of the sphere of variable radius b, located at the origin. Then

$$\frac{\partial}{\partial b}\, h_b(x)\,\bigg|_{b\,=\,a} = \delta(|x| - a). \tag{2.14}$$

As a matter of fact, the formula (2.14) was noted by Doi [3] who employed it for evaluating the interfacial correlations for the Boolean model of fully penetrable spheres. Coupled with Stratonovich density field (2.1), it gives

$$|\nabla I_1(x)| = \int \frac{\partial}{\partial b}\, h_b(x-y)\,\psi(y)\,\mathrm{d}y\,\bigg|_{b\,=\,a}, \tag{2.15}$$

since $|\nabla I_1(x)|$ is a sum of delta functions, concentrated on the surfaces of the spheres. The formula (2.15) will play a central role in our study. Its first and simplest consequence is the formula for the specific surface, S, of the dispersion, i.e. the amount of the interface in a unit volume. Due to the nonoverlapping assumption, obviously $S = 4\pi a^2 n$. Formally, the latter formula immediately follows after averaging (2.15):

$$S = \langle |\nabla I_1(x)| \rangle = n\frac{\partial}{\partial b}\int h_b(x-y)\,\mathrm{d}y\,\bigg|_{b\,=\,a} = n\frac{\mathrm{d}}{\mathrm{d}b}\left(\tfrac{4}{3}\pi b^3\right)\bigg|_{b\,=\,a} = 4\pi a^2 n. \tag{2.16}$$

Similarly to (2.8), represent the interfacial correlations in the form

$$F^{\mathrm{sc}}(x) = nS + \overline{F}^{\mathrm{sc}}(x), \quad F^{\mathrm{sp}}(x) = \eta_1 S + \overline{F}^{\mathrm{sp}}(x), \quad F^{\mathrm{ss}}(x) = S^2 + \overline{F}^{\mathrm{ss}}(x), \tag{2.17}$$

where, as it follows from (2.11), (2.4), (2.12) and (2.13) ,

$$\begin{aligned}
\overline{F}^{\mathrm{sc}}(x) &= \langle |\nabla I_1(x)| \psi'(0)\rangle \, n \frac{\partial}{\partial b}\, h_b(x) \,\Big|_{b=a} + n^2 \frac{\partial}{\partial b} \int h_b(x-y)\nu_2(y)\,\mathrm{d}y \,\Big|_{b=a}, \\
\overline{F}^{\mathrm{sp}}(x) &= \langle |\nabla I_1'(x)| I_1'(0)\rangle = (h_a * \overline{F}^{\mathrm{sc}})(x) = \int h_a(x-y)\overline{F}^{\mathrm{sc}}(y)\,\mathrm{d}y, \\
\overline{F}^{\mathrm{ss}}(x) &= \langle |\nabla I_1(x)|\,(|\nabla I_1(0)| - S)\rangle = (\frac{\partial}{\partial b}\, h_b * \overline{F}^{\mathrm{sc}})(x) \,\Big|_{b=a} \qquad (2.18) \\
&= \int \frac{\partial}{\partial b} h_a(x-y)\overline{F}^{\mathrm{sc}}(y)\,\mathrm{d}y \,\Big|_{b=a} .
\end{aligned}$$

Similarly to (2.8), all $\overline{F}^{\mathrm{sc}}(x)$, $\overline{F}^{\mathrm{sp}}(x)$, $\overline{F}^{\mathrm{ss}}(x)$ vanish at infinity, since the constants in the right-hand sides of (2.17) are the appropriate long-range values.

It is noted after Torquato [15] that the "surface-center" correlation (2.11) is the most important in the sense of (2.18), i.e. the other two — $F^{\mathrm{sp}}(x)$ and $F^{\mathrm{ss}}(x)$ — can be easily represented by means of $F^{\mathrm{sc}}(x)$.

It should be pointed out also that all the correlation functions, mentioned in this section, are particular case of the much more general statistical characteristics for two-phase random media, as introduced by Torquato [15]. Our aim here will be however much more specific, namely, derivation of simple integral representations of these correlations by means of the total correlation function for the set $\{x_\alpha\}$ of sphere's centers of the type of Eq. (3.13) below.

3. THE "PARTICLE-CENTER" CORRELATION

Let us split the radial distribution function, $g(x)$, as

$$g(x) = g^{\mathrm{ws}}(x) + \widetilde{g}(x), \qquad (3.1)$$

where

$$g^{\mathrm{ws}}(x) = 1 - h_{2a}(x) = \begin{cases} 0, & \text{if } |x| \leq 2a, \\ 1, & \text{if } |x| > 2a, \end{cases} \qquad (3.2)$$

corresponds to the simplest "well-stirred" distribution of spheres; $\widetilde{g}(x)$ is then the "correction" to the latter. In turn, the total correlation $\nu_2(x)$, defined in (2.10), is represented as

$$\nu_2(x) = -h_{2a}(x) + \widetilde{\nu}_2(x). \qquad (3.3)$$

Moreover, one has

$$\begin{aligned}
\nu_2(x) &= \widetilde{\nu}_2(x) = \widetilde{g}(x), \quad \text{if} \quad |x| \geq 2a, \\
\widetilde{\nu}_2(x) &= g(x), \quad \text{if} \quad |x| < 2a,
\end{aligned} \qquad (3.4)$$

as a consequence of the nonoverlapping assumption. The formula $(3.4)_1$ will allow us to replace below $\widetilde{g}(x)$ by the binary correlation $\nu_2(x)$ when $|x| = r \geq 2$.

Let us recall now the well-known formula for the common volume of two spheres of radii b and ξ, the first centered at the origin, the other at the point x, $|x| = r$:

$$(h_b * h_\xi)(x) = \int h_b(x-y)h_\xi(y)\,\mathrm{d}y = V_a \begin{cases} \tau^3, & \text{if } 0 \leq \rho \leq \mu - \tau, \\ \Psi(\rho;\mu,\tau), & \text{if } \mu - \tau \leq \rho \leq \mu + \tau, \\ 0, & \text{if } \rho > \mu + \tau, \end{cases} \tag{3.5}$$

where

$$\Psi(\rho;\mu,\tau) = \frac{1}{16\rho}(\mu+\tau-\rho)^2(\rho^2 + 2(\mu+\tau)\rho - 3(\mu-\tau)^2), \tag{3.6}$$

with the dimensionless variables

$$\rho = r/a, \quad \mu = \xi/a, \quad \tau = b/a. \tag{3.7}$$

It is assumed in (3.5) that $\xi \geq b$, i.e. $\mu \geq \tau$. The elementary formulae (3.5) and (3.6) will play a central role in the sequel.

From (2.8), (3.1) and (3.2) it now follows

$$\overline{F}^{\,\mathrm{pc}}(x) = \overline{F}^{\,\mathrm{pc}}_{\mathrm{ws}}(x) + \widetilde{F}^{\,\mathrm{pc}}(x), \tag{3.8}$$

where

$$\begin{aligned} \overline{F}^{\,\mathrm{pc}}_{\mathrm{ws}}(x) &= nh_a(x) - n^2(h_a * h_{2a})(x) \\ &= n\eta_2 h_a(x) - \frac{n\eta_1}{16\rho}(3-\rho)^2(\rho^2+6\rho-3)\big[h_{3a}(x) - h_a(x)\big], \end{aligned} \tag{3.9}$$

$$\widetilde{F}^{\,\mathrm{pc}}(x) = n^2 \int h_a(x-y)\widetilde{g}(y)\,\mathrm{d}y. \tag{3.10}$$

This formula implies that

$$\widetilde{F}^{\,\mathrm{pc}}(x) = 0, \quad \text{if} \quad |x| \leq a, \tag{3.11}$$

since $\widetilde{g}(x) = 0$ at $|x| \leq 2a$, see $(3.4)_2$.

To represent $\widetilde{F}^{\,\mathrm{pc}}(x)$ as a simple one-tuple integral, containing the function $\widetilde{g}(x)$, write down the latter as

$$\widetilde{g}(y) = \int_{2a}^{\infty} \widetilde{g}(A)\,\frac{\partial}{\partial A} h_A(y)\,\mathrm{d}A, \tag{3.12}$$

which follows from (2.14). Then, in virtue of (3.5) (at $\tau = 1$) and $(3.4)_2$,

$$\begin{aligned} \widetilde{F}^{\,\mathrm{pc}}(x) &= n^2 \int_2^{\infty} \mathrm{d}\mu\, \widetilde{g}(\mu) \frac{\partial}{\partial\mu}(h_a * h_\xi)(r) \\ &= \frac{3n\eta_1}{4\rho} \int_{\max\{2,\rho-1\}}^{\rho+1} \left[1 - (\mu-\rho)^2\right] \mu\nu_2(\mu)\,\mathrm{d}\mu. \end{aligned} \tag{3.13}$$

The obtained simple representation of $F^{\text{pc}}(x)$ by means of the total correlation allows one to interconnect the moments

$$\theta_k^{\text{pc}} = \int_0^\infty \rho^k \overline{F}^{\text{pc}}(r)\,\mathrm{d}\rho, \quad k = 0, 1, \ldots, \tag{3.14}$$

of $F^{\text{pc}}(r)$ on the semiaxis $(0, \infty)$ with the appropriate moments of the total correlation. Indeed, due to (3.8) and (3.9),

$$\theta_k^{\text{pc}} = \theta_{k,\text{ws}}^{\text{pc}} + \widetilde{\theta}_k^{\text{pc}}. \tag{3.15}$$

The first term in (3.15) corresponds to the well-stirred distribution when $\overline{F}^{\text{pc}}(r) = \overline{F}_{\text{ws}}^{\text{pc}}(x)$ is given in (3.9); the appropriate integration is elementary. In turn, $\widetilde{\theta}_m^{\text{pc}}$ corresponds to the deviation $\widetilde{g}(r)$ of the radial distribution function from the well-stirred statistics. Using (3.13) and changing the order of integration give

$$\begin{aligned} \widetilde{\theta}_k^{\text{pc}} &= n\eta_1 \int_2^\infty H_k^{\text{pc}}(\mu)\mu\nu_2(\mu)\mathrm{d}\mu, \\ H_k^{\text{pc}}(\mu) &= \frac{3}{4}\int_{\mu-1}^{\mu+1} \rho^{k-1}\left[1 - (\mu-\rho)^2\right]\,\mathrm{d}\rho. \end{aligned} \tag{3.16}$$

The functions $H_k^{\text{pc}}(\mu)$ in (3.16) are polynomials whose explicit evaluation is straightforward. In particular,

$$H_1^{\text{pc}}(\mu) = 1, \quad H_2^{\text{pc}}(\mu) = \mu, \quad \text{etc.} \tag{3.17}$$

Hence, if

$$m_k = \int_2^\infty \rho^k \nu_2(\rho)\,\mathrm{d}\rho, \quad k = 0, 1, \ldots, \tag{3.18}$$

are the moments on $(2, \infty)$ of the binary correlation $\nu_2(\rho)$ or, which is the same, of the "correction" $\widetilde{g}(\rho)$ to the radial distribution function, then the formulae (3.16) and (3.17), together with (3.9), imply

$$\theta_1^{\text{pc}} = n\eta_1\left(\frac{5 - 19\eta_1}{10\eta_1} + m_1\right), \quad \theta_2^{\text{pc}} = n\eta_1\left(\frac{1 - 8\eta_1}{3\eta_1} + m_2\right), \quad \text{etc.} \tag{3.19}$$

4. THE "SURFACE-CENTER" CORRELATION

Inserting (3.3) into $(2.18)_1$ gives

$$\overline{F}^{\text{sc}}(x) = \overline{F}_{\text{ws}}^{\text{sc}}(x) + \widetilde{F}^{\text{sc}}(x), \tag{4.1}$$

where

$$\overline{F}_{\text{ws}}^{\text{sc}}(x) = n\delta(r-a) - n^2\frac{\partial}{\partial b}\int h_b(x-y)h_{2a}(y)\,\mathrm{d}y\,\bigg|_{b=a}, \tag{4.2}$$

$$\widetilde{F}^{\,\mathrm{sc}}(x) = n^2 \frac{\partial}{\partial b} \int h_b(x-y)\widetilde{g}(y)\,\mathrm{d}y \,\Bigg|_{b=a}. \tag{4.3}$$

Hence, the first term, $F^{\,\mathrm{sc}}_{\mathrm{ws}}(x)$, in (4.1) corresponds to the well-stirred distribution, while the second one, $\widetilde{F}^{\,\mathrm{sc}}(x)$, is due to the deviation, $\widetilde{g}(x)$, of the radial distribution function from the latter.

Combining $(2.18)_1$ and (4.2), and using (4.3) (at $\xi = 2a$) give eventually the "surface-center" correlation (2.11) in the well-stirred case:

$$\overline{F}^{\mathrm{sc}}_{\mathrm{ws}}(x) = n\delta(|x|-a) - nS \begin{cases} 0, & \text{if } 0 \le \rho \le 1, \\ \dfrac{(\rho+1)(3-\rho)}{4\rho}, & \text{if } 1 < \rho \le 3, \\ 0, & \text{if } \rho > 3. \end{cases} \tag{4.4}$$

To evaluate the deviation $\widetilde{F}^{\,\mathrm{sc}}(x)$ from (4.3), we shall use once again the representation (3.12):

$$\widetilde{F}^{\,\mathrm{sc}}(x) = n^2 \int_{2a}^{\infty} \mathrm{d}\xi\, \widetilde{g}(\xi) \left\{ \frac{\partial^2}{\partial b \partial \xi} \int h_b(x-y) h_\xi(y)\,\mathrm{d}y \right\}_{b=a}. \tag{4.5}$$

Applying (3.5) yields the needed formula

$$\widetilde{F}^{\,\mathrm{sc}}(x) = \frac{nS}{2\rho} \begin{cases} 0, & \text{if } 0 \le \rho \le 1, \\ \displaystyle\int_{\max\{2,\rho-1\}}^{\rho+1} \mu \nu_2(\mu)\,\mathrm{d}\mu, & \text{if } \rho > 1. \end{cases} \tag{4.6}$$

Similarly to Section 3, consider the evaluation of the moments of $F^{\,\mathrm{sc}}(x)$, i.e. the quantities

$$\theta^{\mathrm{sc}}_k = \int_0^{\infty} \rho^k\, \overline{F}^{\,\mathrm{sc}}(r)\mathrm{d}\rho, \quad k = 0, 1, \ldots \tag{4.7}$$

Due to (4.1), again

$$\theta^{\mathrm{sc}}_k = \theta^{\mathrm{sc}}_{k,\mathrm{ws}} + \widetilde{\theta}^{\mathrm{sc}}_k \tag{4.8}$$

— the first term in (4.8) corresponds to the well-stirred distribution and its evaluation is elementary; the second is due to the "deviation" $\widetilde{g}(r)$. To evaluate the latter, insert (4.6) into (4.7) and change again the order of integration:

$$\begin{gathered} \widetilde{\theta}^{\mathrm{sc}}_k = nS \int_2^{\infty} H^{\mathrm{sc}}_k(\mu)\widetilde{g}(\mu)\mathrm{d}\mu, \\ H^{\mathrm{sc}}_k(\mu) = \frac{1}{2} \int_{\mu-1}^{\mu+1} \rho^{k-1}\,\mathrm{d}\rho = \frac{(\mu+1)^k - (\mu-1)^k}{2k}. \end{gathered} \tag{4.9}$$

Hence $H^{\mathrm{sc}}_1(\mu) = 1$, $H^{\mathrm{sc}}_2(\mu) = \mu$, etc. Together with (4.8), (4.4) and (4.9), this implies

$$\theta^{\mathrm{sc}}_1 = nS \left(\frac{1 - 11\eta_1/2}{3\eta_1} + m_1 \right), \quad \theta^{\mathrm{sc}}_2 = nS \left(\frac{1 - 8\eta_1}{3\eta_1} + m_2 \right), \quad \text{etc.} \tag{4.10}$$

First, let us evaluate $F^{\text{sp}}(0)$:

$$F^{\text{sp}}(0) = \langle\, |\nabla I_1(0)|\, I_1(0)\rangle = \frac{\partial}{\partial b} \iint h_a(y_1) h_b(y_2)\ \langle \psi(y_1)\psi(y_2)\rangle\ \mathrm{d}y_1 \mathrm{d}y_2 \,\Big|_{b=a}$$

$$= \frac{\partial}{\partial b} \int h_a(y) h_b(y)\,\mathrm{d}y \,\Big|_{b=a}, \tag{5.1}$$

having used (2.2) and the fact that $g(y_1 - y_2) = 0$ if $|y_1 - y_2| \leq 2a$, due to the nonoverlapping assumption. But

$$\frac{\partial}{\partial b} \int h_a(y) h_b(y)\,\mathrm{d}y = \frac{\partial}{\partial b} \begin{cases} \frac{4}{3}\pi a^3, & \text{if } b > a, \\ \frac{4}{3}\pi b^3, & \text{if } b < a, \end{cases} \tag{5.2}$$

which equals 0 if $b > a$, and $4\pi b^2$ if $b < a$. Hence, a question appears, which of the two values, 0 or $S = 4\pi a^2 n$, should be attributed to $F^{\text{sp}}(0)$ when putting $b = a$ in (5.1) and (5.2). The correct answer is *one-half* of these two values, i.e.

$$F^{\text{sp}}(0) = \frac{1}{2} S. \tag{5.3}$$

This will be confirmed by the formal calculations below. Roughly speaking, 1/2 in (5.3) means that the boundary $\partial\mathcal{K}$ is "equally shared" between the constituents. We imagine, in other words, that if a point lies in $\partial\mathcal{K}$, "half" of it belongs to $\mathcal{K}_1$ and the other "half" to $\mathcal{K}_2$.

To evaluate $\overline{F}^{\text{sp}}(x)$, employ its definition from (2.18) and the formula (2.2):

$$\overline{F}^{\text{sp}}(x) = \frac{\partial}{\partial b} \iint h_a(y_1) h_b(x - y_2)\ \langle \psi'(y_1)\psi'(y_2)\rangle\ \mathrm{d}y_1 \mathrm{d}y_2 \,\Big|_{b=a} = A_1 n + A_2 n^2, \tag{5.4}$$

where

$$A_1 = \frac{\partial}{\partial b} \int h_a(y) h_b(x - y)\,\mathrm{d}y \,\Big|_{b=a}, \tag{5.5}$$

$$A_2 = \frac{\partial}{\partial b} \iint h_a(y_1) h_b(x - y_2)\, \nu_2(y_1 - y_2)\,\mathrm{d}y_1 \mathrm{d}y_2 \,\Big|_{b=a}. \tag{5.6}$$

The coefficient A_1 can be immediately found differentiating (3.5) at $\xi = b$ and putting $b = a$ in the result:

$$\frac{\partial}{\partial b}(h_a * h_b)(r) \,\Big|_{b=a} = \pi a^2 \begin{cases} 2 - \rho, & \text{if } 0 \leq r \leq 2a, \\ 0, & \text{if } r > 2a, \end{cases} \tag{5.7}$$

and hence

$$A_1 n = \frac{1}{2} S \begin{cases} 1 - \rho/2, & \text{if } 0 \leq r \leq 2a, \\ 0, & \text{if } r > 2a. \end{cases} \tag{5.8}$$

The formula (5.8) means that

$$F^{\mathrm{sp}}(x) = \overline{F}^{\mathrm{sp}}(x) = \frac{1}{2} S\left(1 - \frac{r}{2a}\right) h_{2a}(x) + o(n),$$

which agrees with the result of Berryman [2], see also [14], found by means of different arguments.

To evaluate the coefficient A_2 from (5.6), we shall literally follow the reasoning of [7]. Consider to this end the triple convolution

$$\left(h_a * \frac{\partial}{\partial b} h_b * h_A\right)(r)\Bigg|_{b=a} = \left(\left(\varphi^{\mathrm{sp}}(t) h_{2a}\right) * h_A\right)(r), \tag{5.9}$$

where, according to (5.7),

$$\varphi^{\mathrm{sp}}(t) = \left(h_a * \frac{\partial}{\partial b} h_b\right)(r)\Bigg|_{b=a} = \pi a^2(2-t), \quad t = r/a. \tag{5.10}$$

Similarly to [7], we treat $\varphi^{\mathrm{sp}}(t)$ as pertaining to an inhomogeneous and radially-symmetric ball whose density decreases along the radius according to (5.10). This inhomogeneous ball is then approximated, for a given division $0 = \xi_0 < \xi_1 < \dots \xi_{N-1} < \xi_N = 2a$ of the interval $(0, 2a)$, by a family of concentric spherical layers $\xi_i < r < \xi_{i+1}$, each one homogeneous and of density $\varphi^{\mathrm{sp}}(\xi_i)$. In the limit $\triangle\xi_i = \xi_i - \xi_{i-1} \to 0$ one finds

$$\begin{aligned}\Big(\left(\varphi^{\mathrm{sp}}(t) h_{2a}\right) * h_A\Big)(r) &= \int_0^{2a} \varphi^{\mathrm{sp}}(\xi/a)\, \frac{\partial}{\partial \xi}\left(h_\xi * h_A\right)(r)\, \mathrm{d}\xi \\ = \varphi^{\mathrm{sp}}(\xi/a)\left(h_\xi * h_A\right)(r)\,\Big|_{\xi=0}^{\xi=2a} &- \frac{1}{a}\int_0^{2a} \left(h_\xi * h_A\right)(r) \frac{\partial}{\partial \xi}\varphi^{\mathrm{sp}}(\xi/a)\, \mathrm{d}\xi \\ = \pi a^2 \int_0^2 \left(h_\xi * h_A\right)(r)\, \mathrm{d}\mu &= 4\pi a^2 V_a U_{\mathrm{sp}}(\rho;\tau),\end{aligned} \tag{5.11}$$

since $\varphi^{\mathrm{sp}}(2) = 0$ and $h_\xi * h_A \Big|_{\xi=0} = 0$. In accordance with the notations (3.7), $\mu = \xi/a$ and $\tau = A/a \geq 2$. The evaluation of the function $U_{\mathrm{sp}}(\rho;\tau)$ is obvious, using (3.5) at $b = A$ in (5.11), and the final result reads

$$U_{\mathrm{sp}}(\rho;\tau) = \begin{cases} U_{\mathrm{sp}}^{(I)}(\rho;\tau), & \text{if } 0 \leq \rho \leq \tau - 2, \\ U_{\mathrm{sp}}^{(II)}(\rho;\tau), & \text{if } \tau - 2 \leq \rho \leq \tau, \\ U_{\mathrm{sp}}^{(III)}(\rho;\tau), & \text{if } \tau \leq \rho \leq \tau + 2\ , \\ 0, & \text{if } \rho > \tau + 2, \end{cases} \tag{5.12}$$

where

$$U_{\rm sp}^{(I)}(\rho;\tau) = \frac{1}{4}\int_0^2 \mu^3 \mathrm{d}\mu = 1,$$

$$U_{\rm sp}^{(II)}(\rho;\tau) = \frac{1}{4}\int_0^{\tau-\rho} \mu^3\,d\mu + \frac{1}{4}\int_{\tau-\rho}^2 \Psi(\rho;\tau,\mu)\,\mathrm{d}\mu, \tag{5.13}$$

$$U_{\rm sp}^{(III)}(\rho;\tau) = \frac{1}{4}\int_{\rho-\tau}^2 \Psi(\rho;\tau,\mu)\,\mathrm{d}\mu,$$

with $\Psi(\rho;\tau,\mu)$ defined in (3.6). The integrals in (5.13) can be analytically evaluated, but the only formulae that will be important for the sequel are

$$\left(h_a * \frac{\partial}{\partial b} h_b * h_{2a}\right)(r)\Bigg|_{b=a} = 4\pi a^2 V_a U_{\rm sp}(\rho;2),$$

$$U_{\rm sp}(\rho;2) = \begin{cases} 1 - \frac{1}{4}\rho^2 + \frac{5}{160}\rho^3 + \frac{1}{160}\rho^4, & \text{if } 0 \le \rho \le 2, \\ \frac{(4-\rho)^3(\rho^2+7\rho-4)}{160\rho}, & \text{if } 2 < \rho < 4, \\ 0, & \text{if } \rho \ge 4. \end{cases} \tag{5.14}$$

Also, it turns out that

$$\frac{\partial}{\partial\tau} U_{\rm sp}(\rho;\tau) = \frac{3\tau}{4\rho} G^{\rm sp}(\rho-\tau),$$

$$G^{\rm sp}(t) = \begin{cases} f^{\rm sp}(t), & \text{if } -2 \le t \le 0, \\ f^{\rm sp}(-t), & \text{if } 0 \le t \le 2, \\ 0, & \text{if } |t| \ge 2, \end{cases} \tag{5.15}$$

$$f^{\rm sp}(t) = \frac{1}{8}(2+t)^2\,(1-t).$$

As a first application of the foregoing formulae, consider the well-stirred approximation, see (3.2). The coefficient A_2 from (5.6) then becomes

$$A_2 n^2 = -n^2 \left(h_a * \frac{\partial}{\partial b} h_b * h_{2a}\right)(r)\Bigg|_{b=a}$$

and application of (5.4), (5.8) and (5.14) gives eventually

$$F_{\rm ws}^{\rm sp}(r) = \eta_1 S + \overline{F}_{\rm ws}^{\rm sp}(r),$$

$$\overline{F}_{\rm ws}^{\rm sp}(r) = S \begin{cases} \frac{1}{2} - \frac{\rho}{4} - \eta_1\left[1 - \frac{1}{4}\rho^2 + \frac{5}{160}\rho^3 + \frac{1}{160}\rho^4\right], & \text{if } 0 \le \rho \le 2, \\ \frac{(4-\rho)^3(4-7\rho-\rho^2)}{160\rho}\eta_1, & \text{if } 2 < \rho < 4, \\ 0, & \text{if } \rho \ge 4. \end{cases} \tag{5.16}$$

In the general case the radial correlation function $g(r)$ is decomposed again as the sum (3.1), so that

$$F^{\rm sp}(r) = F^{\rm sp}_{\rm ws}(r) + \widetilde{F}^{\rm sp}(r), \tag{5.17}$$

with the well-stirred contribution, given in (5.16), and

$$\widetilde{F}^{\rm sp}(r) = n^2 \frac{\partial}{\partial b} \iint h_a(y_1) h_b(x - y_2)\, \widetilde{g}(y_1 - y_2)\, \mathrm{d}y_1 \mathrm{d}y_2 \,\bigg|_{b=a} . \tag{5.18}$$

The evaluation of this integral follows the reasoning of Section 3. Namely, inserting (3.12) in the right-hand side of (5.18) yields

$$\begin{aligned} \widetilde{F}^{\rm sp}(r) &= n^2 \int_{2a}^{\infty} \widetilde{g}(A) \frac{\partial}{\partial A} \left(h_a * \frac{\partial h_b}{\partial b} * h_A \right)(r)\, \mathrm{d}A \,\bigg|_{b=a} \\ &= 4\pi a^2 n^2 V_a \int_2^{\infty} \widetilde{g}(\tau) \frac{\partial}{\partial \tau} U_{\rm sp}(\rho; \tau)\, \mathrm{d}\tau \\ &= \frac{\eta_1 S}{\rho} \int_{\max\{\rho-2,2\}}^{\rho+2} G^{\rm sp}(\rho - \tau)\, \tau \nu_2(\tau)\, \mathrm{d}\tau, \end{aligned} \tag{5.19}$$

as it follows from (2.16) and (5.15).

The formulae (5.16), (5.17), (5.15) and (5.19) provide the needed representation of the "surface-particle" correlation $F^{\rm sp}(r)$ for an arbitrary dispersion of nonoverlapping spheres. They imply, in particular, that indeed $F^{\rm sp}(0) = S/2$, as it was argued in the beginning of this Section, see (5.3). The correction to the total correlation function, $\widetilde{g}(r) = \nu_2(r)$, see (3.4), for the set of sphere centers features in the expression for $F^{\rm sp}(r)$ through a simple one-tuple integral in (5.19). It is noted that the obtained formula for $F^{\rm sp}(r)$ is fully similar to that of Markov and Willis [7] for the "particle-particle" correlation $F^{\rm pp}(r)$ defined in (2.6). (In the latter case, let us recall, the counterpart of the function $f^{\rm sp}(t)$ from (5.15) is $f(t) = f^{\rm pp}(t) = (2+t)^3(4 - 6t + t^2)$, see [7, eq. (33b)].)

Similarly to the previous Sections, the formula (5.19) allows us to evaluate the moments of $\overline{F}^{\rm sp}(x)$ on the semiaxis $(0, \infty)$ to be

$$\theta_k^{\rm sp} = \int_0^{\infty} \rho^k\, \overline{F}^{\rm sp}(\rho) \mathrm{d}\rho = \theta^{\rm sp}_{k,\rm ws} + \widetilde{\theta}_k^{\rm sp}, \tag{5.20}$$

$k = 0, 1, \ldots$ The well-stirred contribution $\theta^{\rm sp}_{k,\rm ws}$ can be found by means of an elementary integration, using (5.16). For the "corrections" $\theta_k^{\rm sp}$ we have

$$\begin{aligned} \widetilde{\theta}_k^{\rm sp} &= \eta_1 S \int_2^{\infty} H_k^{\rm sp}(\mu) \mu \nu_2(\mu) \mathrm{d}\mu, \\ H_k^{\rm sp}(\mu) &= \int_{\mu-2}^{\mu} \rho^{k-1} f^{\rm sp}(\rho - \mu)\, \mathrm{d}\rho + \int_{\mu}^{\mu-2} \rho^{k-1} f^{\rm sp}(\mu - \rho)\, \mathrm{d}\rho, \end{aligned} \tag{5.21}$$

as it follows from (5.19) and (5.20). Recalling the form of $f^{\mathrm{sp}}(t)$ from (5.15), one easily finds, in particular, $H_1^{\mathrm{sp}}(\mu) = 1$, $H_2^{\mathrm{sp}}(\mu) = \mu$, etc., and hence, using (5.16),

$$\theta_1^{\mathrm{sp}} = S\eta_1 \left(\frac{5 - 26\eta_1}{15\eta_1} + m_1 \right), \quad \theta_2^{\mathrm{sp}} = S\eta_1 \left(\frac{1 - 8\eta_1}{3\eta_1} + m_2 \right), \quad \text{etc.,} \tag{5.22}$$

where m_k are the moments (3.18).

6. THE "SURFACE-SURFACE" CORRELATION

Due to (2.17), (2.15) and (2.2), we have in this case

$$\overline{F}^{\mathrm{ss}}(x) = \frac{\partial^2}{\partial b \partial c} \iint h_b(y_1) h_c(x - y_2) \left\langle \psi'(y_1) \psi'(y_2) \right\rangle \mathrm{d}y_1 \mathrm{d}y_2 \Bigg|_{b,c=a} = B_1 n + B_2 n^2, \tag{6.1}$$

where

$$B_1 = \frac{\partial^2}{\partial b \partial c} \int h_b(y) h_c(x - y) \, \mathrm{d}y \Bigg|_{b,c=a}, \tag{6.2}$$

$$B_2 = \frac{\partial^2}{\partial b \partial c} \iint h_b(y_1) h_c(x - y_2) \, \nu_2(y_1 - y_2) \, \mathrm{d}y_1 \mathrm{d}y_2 \Bigg|_{b,c=a}. \tag{6.3}$$

The coefficient B_1 can be immediately found, evaluating the second mixed derivative $\partial^2 / \partial\mu\partial\tau$ of the function Ψ, see (3.5), and putting $\mu = \tau = 1$ in the result:

$$B_1 = \frac{2\pi a}{\rho} \begin{cases} 1, & \text{if } \rho \leq 2, \\ 0, & \text{if } \rho > 2, \end{cases} \tag{6.4}$$

which means that in the dilute case

$$F^{\mathrm{ss}}(x) = \overline{F}^{\mathrm{ss}}(x) = \frac{S}{2r} h_{2a}(x) + o(n).$$

The latter agrees with the result of Berryman [2], see also [14], found by means of different arguments.

To calculate B_2, consider again the appropriate triple convolution, similar to (5.9):

$$\begin{aligned} \Big(\frac{\partial}{\partial b} h_b * \frac{\partial}{\partial c} h_c * h_A \Big)(r) \Bigg|_{b,c=a} &= \Big(\big(\varphi^{\mathrm{ss}}(\xi/a) h_{2a}\big) * h_A \Big)(r) \\ &= \varphi^{\mathrm{ss}}(\xi/a) \big(h_\xi * h_A\big)(r) \Bigg|_{\xi=0}^{\xi=2a} - \int_0^2 \big(h_\xi * h_A\big)(r) \frac{\mathrm{d}}{\mathrm{d}\mu} \varphi^{\mathrm{ss}}(\mu) \, \mathrm{d}\mu \\ &= \pi a \big(h_{2a} * h_A\big)(r) + 4\pi a V_a U_{\mathrm{ss}}(\rho; \tau), \end{aligned} \tag{6.5}$$

having used that

$$\varphi^{\mathrm{ss}}(t) = \left(\frac{\partial}{\partial b} h_b * \frac{\partial}{\partial c} h_c \right)(r) \Bigg|_{b,c=a} = \frac{2\pi a}{t} h_{2a}(t),$$

$t = r/a$, see (6.2) and (6.4). The function $U_{\text{ss}}(\rho;\tau)$ in (6.5) has the same form as that of its "surface-particle" counterpart $U_{\text{sp}}(\rho;\tau)$ in (5.12), with the functions

$$\begin{aligned} U_{\text{ss}}^{(I)}(\rho;\tau) &= \frac{1}{2}\int_0^2 \mu d\mu = 1, \\ U_{\text{ss}}^{(II)}(\rho;\tau) &= \frac{1}{2}\int_0^{\tau-\rho} \mu\, d\mu + \frac{1}{2}\int_{\tau-\rho}^2 \frac{1}{\mu^2}\,\Psi(\rho;\tau,\mu)\,d\mu, \\ U_{\text{ss}}^{(III)}(\rho;\tau) &= \frac{1}{2}\int_{\rho-\tau}^2 \frac{1}{\mu^2}\,\Psi(\rho;\tau,\mu)\,d\mu, \end{aligned} \tag{6.6}$$

where $\Psi(\rho;\tau,\mu)$ is defined in (3.6). The integrals in (6.6) can be analytically evaluated, similarly to those in (5.13), but again the only formulae important for the sequel are, first,

$$\left(\frac{\partial}{\partial b}h_b * \frac{\partial}{\partial c}h_c * h_{2a}\right)(r)\bigg|_{b,c=a} = \pi a\big(h_{2a} * h_{2a}\big)(r) + 4\pi a V_a U_{\text{ss}}(\rho;2),$$

$$U_{\text{ss}}(\rho;2) = \begin{cases} 1 - \dfrac{1}{8}\rho^2 - \dfrac{1}{64}\rho^3, & \text{if } 0 \le \rho \le 2, \\ \dfrac{(4-\rho)^3(\rho+4)}{64\rho}, & \text{if } 2 < \rho < 4, \\ 0, & \text{if } \rho \ge 4. \end{cases} \tag{6.7}$$

Second, it turns out that

$$\begin{aligned} \frac{\partial}{\partial\tau}U_{\text{ss}}(\rho;\tau) &= \frac{3\tau}{16\rho}\,G_0^{\text{ss}}(\rho-\tau), \\ G_0^{\text{ss}}(t) &= \begin{cases} f_0^{\text{ss}}(t), & \text{if } -2 \le t \le 0, \\ f_0^{\text{ss}}(-t), & \text{if } 0 \le t \le 2, \\ 0, & \text{if } |t| \ge 2, \end{cases} \\ f_0^{\text{ss}}(t) &= (2+t)^2. \end{aligned} \tag{6.8}$$

In the well-srirred case, as it follows from (3.5), (6.1), (6.3), (6.4) and (6.7),

$$\overline{F}_{\text{ws}}^{\text{ss}}(x) = \frac{S^2}{12\eta_1}\left\{\frac{2}{\rho}h_{2a}(x) - \eta_1\left[\frac{1}{16}(\rho-4)^2(\rho+8)h_{4a}(x) + 4U_{\text{ss}}(\rho;2)\right]\right\}. \tag{6.9}$$

In the general case $g(r)$ is once again decomposed into the form (3.1), so that

$$\overline{F}^{\text{ss}}(r) = \overline{F}_{\text{ws}}^{\text{ss}}(r) + \widetilde{F}^{\text{ss}}(r), \tag{6.10}$$

with the well-stirred part, $\overline{F}^{\mathrm{ss}}_{\mathrm{ws}}(r)$, given in (6.7), and

$$\begin{aligned}\widetilde{F}^{\mathrm{ss}}(r) &= n^2 \int_{2a}^{\infty} \widetilde{g}(A) \frac{\partial}{\partial A} \left(\frac{\partial h_b}{\partial b} * \frac{\partial h_c}{\partial c} * h_A \right)(r)\,\mathrm{d}A \,\bigg|_{b,c=a} \\ &= \int_{2}^{\infty} \widetilde{g}(\tau) \frac{\partial}{\partial \tau} \Big\{ \pi a (h_{2a} * h_A)(r) + 4\pi a V_a U_{\mathrm{ss}}(\rho;\tau)\,\mathrm{d}\tau \Big\} \qquad (6.11) \\ &= \frac{1}{16\rho} S^2 \int_{\max\{\rho-2,2\}}^{\rho+2} \left[4 - (\rho-\tau)^2 + G_0^{\mathrm{ss}}(\rho-\tau) \right] \tau \widetilde{g}(\tau)\,\mathrm{d}\tau,\end{aligned}$$

as it follows from (2.16), (3.5), (3.6) and (6.8), $\tau = A/a$. Taking into account (6.8), we can recast (6.11) into the following final form:

$$\widetilde{F}^{\mathrm{ss}}(r) = \frac{S^2}{\rho} \int_{\max\{\rho-2,2\}}^{\rho+2} G^{\mathrm{ss}}(\rho-\tau)\, \tau \nu_2(\tau)\,\mathrm{d}\tau, \qquad (6.12)$$

where the function G^{ss} has the same form as G_0^{ss} in (6.8), but with the function $f_0^{\mathrm{ss}}(t)$ replaced by

$$f^{\mathrm{ss}}(t) = \frac{1}{4}(2+t). \qquad (6.13)$$

For the moments of $\overline{F}^{\mathrm{sp}}(x)$ on the semiaxis $(0,\infty)$ we have, similarly to the previous sections,

$$\begin{aligned} \theta_k^{\mathrm{ss}} &= \int_0^{\infty} \rho^k\, \overline{F}^{\mathrm{ss}}(r) \mathrm{d}\rho = \theta_{k,\mathrm{ws}}^{\mathrm{ss}} + \widetilde{\theta}_k^{\mathrm{ss}}, \quad k = 0,1,\ldots, \\ \widetilde{\theta}_k^{\mathrm{ss}} &= S^2 \int_2^{\infty} H_k^{\mathrm{ss}}(\mu) \mu \widetilde{g}(\mu) \mathrm{d}\mu, \qquad (6.14) \\ H_k^{\mathrm{ss}}(\mu) &= \frac{1}{4} \left\{ \int_{\mu-2}^{\mu} \rho^{k-1}(2+\rho-\mu)\,\mathrm{d}\rho + \int_{\mu}^{\mu-2} \rho^{k-1}(2+\mu-\rho)\,\mathrm{d}\rho \right\}, \\ H_1^{\mathrm{sp}}(\mu) &= 1, \quad H_2^{\mathrm{sp}}(\mu) = \mu, \quad \text{etc.} \end{aligned}$$

The well-stirred contribution, $\theta_{k,\mathrm{ws}}^{\mathrm{ss}}$, can be elementary found by means of (6.9). In particular,

$$\theta_1^{\mathrm{ss}} = S^2 \left(\frac{1-5\eta_1}{3\eta_1} + m_1 \right), \quad \theta_2^{\mathrm{ss}} = S^2 \left(\frac{1-8\eta_1}{3\eta_1} + m_2 \right), \qquad (6.15)$$

where m_k are the moments (3.18).

7. DIRECT EVALUATION OF THE FIRST TWO MOMENTS OF THE CORRELATION FUNCTIONS

In the application to be dealt with below (Section 8), the first moments like θ_1^{pp}, θ_1^{ps}, etc., will be of central importance. They were evaluated in the preceeding sections as consequences of the appropriate integral representations of the two-point correlations through the radial distribution functions. There exists, however, a simpler and more direct method, based on the interconnections (2.9) and (2.18). The method works equally well in the 2-D case, when the derivation of the counterparts of the above integral representations for the two-point correlations should be considerably more complicated. (The reason is that the common surface of two circles in the plane is not already a rational function of the distance between the circle's centers and their radii, in contrast with the 3-D simple function (3.6) that gives the common volume of two balls.)

Integrate $(2.9)_2$ over the whole $\mathbb{R}^3$ and introduce (3.1) in the result:

$$\int \overline{F}^{\,\mathrm{pc}}(x)\,\mathrm{d}x = 4\pi a^3\theta_2^{\mathrm{pc}} = nV_a + n^2V_a(-V_{2a} + 4\pi a^3 m_2),$$

having used the definition of m_2, see (3.18). Since $V_{2a} = 8V_a$ and $nV_a = \eta_1$, the already known formula for θ_2^{pc} immediately follows, cf. (3.19).

Integrate next $(2.9)_3$ over $\mathbb{R}^3$:

$$\int \overline{F}^{\,\mathrm{pp}}(x)\,\mathrm{d}x = V_a\int \overline{F}^{\,\mathrm{pc}}(x)\,\mathrm{d}x, \quad \text{i.e.} \quad \theta_2^{\mathrm{pp}} = V_a\theta_2^{\mathrm{pc}},$$

or

$$\theta_2^{\mathrm{pp}} = \eta_1^2\left(\frac{1-8\eta_1}{3\eta_1} + m_2\right) \tag{7.1}$$

— a formula derived in [7] by means of the appropriate integral representation of $F^{\mathrm{pp}}(x)$ through the radial distribution function.

The reasoning is fully similar in 2-D; only the volume $V_a = \frac{4}{3}\pi a^3$ is replaced by the surface $S_a = \pi a^2$, $\eta_1 = nS_a$ and $S_{2a} = 4S_a$, which yields

$$\begin{aligned} &\int \overline{F}^{\,\mathrm{pc}}(x)\,\mathrm{d}x = 2\pi a^2\int \rho\overline{F}^{\,\mathrm{pc}}(x)\,\mathrm{d}\rho = 2\pi a^2\theta_1^{\mathrm{pc}}, \qquad \theta_1^{\mathrm{pp}} = S_a\theta_1^{\mathrm{pc}}, \\ &\theta_1^{\mathrm{pc}} = n\left(\frac{1-4\eta_1}{2} + \eta_1 m_1\right), \quad \theta_1^{\mathrm{pp}} = \eta_1^2\left(\frac{1-4\eta_1}{2\eta_1} + m_1\right) \quad \text{in 2-D.} \end{aligned} \tag{7.2}$$

Note that the correlation function $F^{\mathrm{pp}}(x)$ should be positive definite for any realistic random constitution, see, e.g. [17]. This implies, in particular, that in the 3-D case $\theta_2^{\mathrm{pp}} > 0$, because θ_2^{pp} is proportional to the value of the Fourier transform of $F^{\mathrm{pp}}(x)$ at the origin; similarly, $\theta_1^{\mathrm{pp}} > 0$ in 2-D. From (7.1) and (7.2) it follows then that the well-stirred approximation (3.2) (for which $m_1 = m_2 = 0$) is admissible only if $\eta_1 < 1/8$ in 3-D and $\eta_1 < 1/4$ in 2-D (more generally, if $\eta_1 < 1/2^d$ in a d-dimensional space). Both these critical 3-D and 2-D values have been conjectured by Willis [19] who noticed that the quasi-crystalline approximation in the wave propagation problem in random dispersions fails if η_1 is bigger. A rigorous justification of this conjecture in 3-D was proposed, e.g., in [5] and [7].

For the interfacial correlation, the formulae (2.18) are to be employed in a similar manner. Namely, integrating $(2.18)_1$ over $\mathbb{R}^3$, together with (3.1), gives

$$\begin{aligned} 4\pi a^3\theta_2^{\rm sc} &= n\frac{\partial}{\partial b}\left(\tfrac{4}{3}\pi b^3\right)\Big|_{b=a} + n^2\int \frac{\partial}{\partial b}\left(\tfrac{4}{3}\pi b^3\right)\Big|_{b=a} \nu_2(y)\,\mathrm{d}y \\ &= 4\pi a^2 n + 4\pi a^2 n^2(-8V_a + 4\pi a^3 m_2), \end{aligned} \tag{7.3}$$

and it remains to notice that $n/a = nS/(3\eta_1)$ in order to reproduce the formula for $\theta_2^{\rm sc}$, cf. (4.10).

Integrate next $(2.18)_2$ over $\mathbb{R}^3$:

$$\int \overline{F}^{\rm sp}(x)\,\mathrm{d}x = V_a\int \overline{F}^{\rm sc}(x)\,\mathrm{d}x, \quad \text{i.e.} \quad \theta_2^{\rm sp} = V_a\theta_2^{\rm sc}, \tag{7.4}$$

cf. (5.22). Finally, from $(2.18)_2$ it follows

$$\int \overline{F}^{\rm ss}(x)\,\mathrm{d}x = 4\pi a^2\int \overline{F}^{\rm sc}(x)\,\mathrm{d}x, \quad \text{i.e.} \quad \theta_2^{\rm ss} = \frac{S}{n}\theta_2^{\rm sc} = \frac{S}{\eta_1}\theta_2^{\rm sp},$$

cf. (7.4) and (6.15).

The 2-D counterparts of the above moments are immediately derived. The counterpart of (7.3) now reads

$$\begin{aligned} 2\pi a^2\theta_1^{\rm sc} &= n\frac{\partial}{\partial b}\left(\pi b^2\right)\Big|_{b=a} + n^2\int \frac{\partial}{\partial b}\left(\pi b^2\right)\Big|_{b=a} \nu_2(y)\,\mathrm{d}y \\ &= 2\pi a n + 2\pi a n^2(-4S_a + 2\pi a^2 m_1), \end{aligned}$$

so that

$$\theta_1^{\rm sc} = nL\left(\frac{1-4\eta_1}{2\eta_1} + m_1\right) \quad \text{in 2-D}, \tag{7.5}$$

where $L = 2\pi a n$ is the "specific length" — the 2-D counterpart of the specific surface $S = 4\pi a^2 n$ in the dispersion; we have also noted that $1/a = L/(2\eta_1)$ in this case. In turn,

$$\theta_1^{\rm sp} = S_a\theta_1^{\rm sc} = L\eta_1\left(\frac{1-4\eta_1}{2\eta_1} + m_1\right), \quad \theta_1^{\rm ss} = \frac{L}{n}\theta_1^{\rm sc} = L^2\left(\frac{1-4\eta_1}{2\eta_1} + m_1\right) \quad \text{in 2-D}. \tag{7.6}$$

To find in 3-D the moments $\theta_1^{\rm pc}$, $\theta_1^{\rm pp}$, etc., multiply first the formula $(2.9)_2$ by $G(x) = 1/(4\pi|x|)$ and integrate the result over $\mathbb{R}^3$:

$$\begin{aligned} a^2\theta_1^{\rm pc} &= \int G(x)\overline{F}^{\rm pc}(x)\,\mathrm{d}x = n\int G(x)h_a(x)\,\mathrm{d}x + n^2\int \varphi_a(y)\nu_2(y)\,\mathrm{d}y \\ &= n\frac{a^2}{2} - n^2\int \varphi_a(y)h_{2a}(y)\,\mathrm{d}y + n^2\int_{|y|\geq 2a} \varphi_a(y)\nu_2(y)\,\mathrm{d}y, \end{aligned}$$

where

$$\varphi_a(x) = (G * h_a)(x) = \begin{cases} (3a^2 - r^2)/6, & \text{if } r < a, \\ a^3/(3r), & \text{if } r \geq a, \end{cases} \tag{7.7}$$

is the well-known harmonic potential of a sphere of radius a. Elementary integration, using (7.7), reproduces the formula for θ_1^{pc}, cf. (3.19).

In turn, multiply $(2.18)_3$ by $G(x)$ and integrate over $\mathbb{R}^3$:

$$a^2\theta_1^{\mathrm{pp}} = \int G(x)\overline{F}^{\mathrm{pp}}(x)\,\mathrm{d}x = \int \varphi_a(y)\overline{F}^{\mathrm{pc}}(y)\,\mathrm{d}y. \tag{7.8}$$

But, as it follows from (7.7),

$$\varphi_a(x) = V_aG(x) + \left[\frac{1}{6}(3a^2 - r^2) - \frac{V_a}{4\pi r}\right] h_a(x), \tag{7.9}$$

which is introduced into (7.8):

$$a^2\theta_1^{\mathrm{pp}} = a^2V_a\theta_1^{\mathrm{pc}} + \int \left[\frac{1}{6}(3a^2 - r^2) - \frac{V_a}{4\pi r}\right] h_a(x)\,\mathrm{d}x.$$

It remains to notice that $\overline{F}^{\mathrm{pc}}(x) = n\eta_2$ if $|x| \leq a$, as it follows from (3.8), (3.9) and (3.11), so that the integral in the last formula equals $-a^2\eta_1/10$ and therefore

$$\theta_1^{\mathrm{pp}} = V_a\theta_1^{\mathrm{pc}} - \frac{\eta_1}{10} = \eta_1^2\left(\frac{2-9\eta_1}{5\eta_1} + m_1\right) \tag{7.10}$$

— a result, also derived in [7] by means of the appropriate integral representation of $F^{\mathrm{pp}}(x)$.

For the interfacial correlation we have, first of all,

$$\begin{aligned} a^2\theta_1^{\mathrm{sc}} &= \int G(x)\overline{F}^{\mathrm{sc}}(x)\,\mathrm{d}x \\ &= n\int G(x)\frac{\partial}{\partial b}h_b(x)\Big|_{b=a}\,\mathrm{d}x + n^2\int \frac{\partial}{\partial b}\varphi_b(y)\Big|_{b=a}\nu_2(y)\,\mathrm{d}y, \end{aligned} \tag{7.11}$$

see $(2.18)_1$. Using (7.7) and (3.1) reproduces the formula (4.10) for θ_1^{sc} after simple integration. In turn, from $(2.18)_2$ it follows

$$a^2\theta_1^{\mathrm{sp}} = \int G(x)\overline{F}^{\mathrm{sp}}(x)\,\mathrm{d}x = \int \varphi(y)\overline{F}^{\mathrm{sc}}(y)\,\mathrm{d}y.$$

Inserting here (7.9) elementary yields the already known formula for θ_1^{sp}, cf. (5.22).

Finally, from $(2.18)_3$ one has

$$\begin{aligned} a^2\theta_1^{\mathrm{ss}} &= \int G(x)\overline{F}^{\mathrm{ss}}(x)\,\mathrm{d}x = \int \frac{\partial}{\partial b}\varphi_b(y)\Big|_{b=a}\overline{F}^{\mathrm{sc}}(y)\,\mathrm{d}y \\ &= 4\pi a^2\int G(x)\overline{F}^{\mathrm{sc}}(x)\,\mathrm{d}x + \int a\left(1-\frac{a}{r}\right)\overline{F}^{\mathrm{sc}}(x)h_a(x)\,\mathrm{d}x, \end{aligned}$$

having used that

$$\left.\frac{\partial\varphi_b(x)}{\partial b}\right|_{b=a} = 4\pi a^2 G(x) + a\left(1-\frac{a}{r}\right)h_a(x), \tag{7.12}$$

which follows from (7.9). But $\overline{F}^{\text{sc}}(x) = n\delta(r-a) - nS$, as it is seen from (4.1), (4.2) and (4.6), and the known formula (6.15) for θ_1^{ss} shows up once again.

8. THE DOI-TALBOT-WILLIS BOUND

As a first and simplest application of the integral representations of the various kinds of two-point correlations, derived in Sections 2 to 6, consider a dispersion of ideal and nonoverlapping spherical sinks (the phase '1'), immersed into an unbounded matrix. The governing equations of this well-known problem read

$$\Delta c(x) + K = 0, \quad x \in \mathcal{K}_2, \quad c(x)\Big|_{\partial\mathcal{K}_2} = 0. \tag{8.1}$$

This equation describes the steady-state behaviour of a species (defects), generated at the rate K within the matrix phase '2', occupying the region $\mathcal{K}_2$, and absorbed by the sinks (the "trapping" phase '2') in the region $\mathcal{K}_1 = \mathbb{R}^3\backslash\mathcal{K}_2$. Then the creation of defects is *exactly* compensated by their removal from the sinks, so that in the steady-stae limit under study

$$k^{*2}\langle c(x)\rangle = K(1-\eta_1). \tag{8.2}$$

The rate constant k^{*2} is just the effective absorption coefficient (the sink strength) of the medium. Its evaluation and bounding for special kinds of random constitution and, above all, for random dispersion of spheres, have been the subject of numerous works, starting with classical studies of Smoluchowski (1916), see, e.g. [4, 3, 12, 9, 16] *et al.* (Note that we have added the factor $1-\eta_1$ in (8.2), due to the fact that in the case under study, defects are created *only* within the phase '2' (the sink-free region), see Richards and Torquato [8] for a discussion.)

We shall confine the analysis to variational bounding of the sink strength k^{*2}, taking into account the foregoing two-point statistical characteristics. Recall to this end the variational principle of Rubinstein and Torquato [9].

Let $\mathcal{A}$ be the class of smooth and statistically homogeneous trial fields such that

$$\mathcal{A} = \Big\{u(x) \mid \Delta u(x) + K = 0, x \in \mathcal{K}_2\Big\}. \tag{8.3}$$

Then

$$k^{*2} \geq \frac{K^2(1-\eta_1)}{\langle I_2(x)|\nabla u(x)|^2\rangle}. \tag{8.4}$$

The equality sign in (8.4) is achieved if $u(x) = c(x)$ is the actual field that solves the problem (8.1).

Since $\left\langle I_2(x)|\nabla u(x)|^2\right\rangle \leq \left\langle|\nabla u(x)|^2\right\rangle$, another bound immediately follows from (8.4), namely,

$$k^{*2} \geq \frac{K^2(1-\eta_1)}{\langle|\nabla u(x)|^2\rangle}, \tag{8.5}$$

see [9]. Though weaker than (8.4), the evaluation of the bound (8.5) is simpler, because it obviously employs smaller amount of statistical information about the medium's constitution.

Following Doi [3] and Rubinstein and Torquato [9], consider the trial fields

$$u(x) = K\int G(x-y)\left[I_2(y) - \xi|\nabla I_1(y)|\right] \mathrm{d}y, \tag{8.6}$$

where $G(x) = 1/(4\pi|x|)$. Since $\Delta G_0(x) + \delta(x) = 0$, it is easily seen that $\Delta u(x) = K$ if $x \in \mathcal{K}_2$, and therefore the fields $u(x)$ in (8.6) are admissible. The constant ξ is uniquely defined from the condition that the integrand in (8.6) should possess zero mean value:

$$\langle I_2(y)\rangle - \xi\,\langle|\nabla I_2(y)\rangle = \eta_2 - \xi S = 0, \quad \text{i.e.} \quad \xi = \xi_0 = \eta_2/S. \tag{8.7}$$

For this choice of ξ, the trial field (8.6) becomes

$$u(x) = -K\int G(x-y)\left[I_1'(y) + \xi_0\big(|\nabla I_1(y)| - S\big)\right) \mathrm{d}y,$$

and hence

$$\left\langle|\nabla u(x)|^2\right\rangle = K^2\left(\theta_1^{\mathrm{pp}} + 2\xi_0\theta_1^{\mathrm{sp}} + \xi_0^2\theta_1^{\mathrm{ss}}\right),$$

after an obvious integration by parts. Using (8.7), (8.5) and the formulae for the appropriate moments (7.10), (5.22) and (6.15) leads eventually to the bound

$$k^{*2}a^2 \geq \frac{3\eta_1(1-\eta_1)}{1-5\eta_1-\eta_1^2/5+3\eta_1 m_1}, \tag{8.8}$$

which coincides with the bound derived by Talbot and Willis [12] by means of an ingineous variational procedure of Hashin-Shtrikman's type, see [6] for more details and discussion. The fact that the original Doi's result, for a dispersion of nonoverlapping spheres, can be recast in the elegant Talbot and Willis' form (8.8) was noticed by Talbot (unpublished manuscript) and, independently, by Beasley and Torquato [1], who apparently were not aware of the paper [12]. Due to all these reasons it seems proper to call (8.8) Doi-Talbot-Willis bound. Another variational procedure that leads to (8.8) has been recently proposed by the author [6].

9. CONCLUDING REMARKS

In the present paper we have represented all two-point correlation functions (2.9) and (2.18) for a random dispersion of nonoverlapping spheres as single integrals containing the binary correlation function $\nu_2(r)$ for the random set of sphere's centers. The reasoning of the recent paper [7], where only the "particle-particle" correlation has been treated in detail, has served as a basis of the analysis. The representations for all two-point correlations have one and the same structure, which can be summarized in the following formulae:

$$
\begin{aligned}
F^{\text{cor}}(\rho) &= F_{\infty}^{\text{cor}} + \overline{F}^{\text{cor}}(\rho), \quad \lim_{\rho\to\infty} \overline{F}^{\text{cor}}(\rho) = 0, \\
F^{\text{cor}}(\rho) &= F_{\text{ws}}^{\text{cor}} + \widetilde{F}^{\text{cor}}(\rho), \\
\widetilde{F}^{\text{cor}}(\rho) &= F_{\infty}^{\text{cor}} \int_{\max\{\rho-\beta,2\}}^{\rho+\beta} G^{\text{cor}}(\rho-\tau)\,\tau\nu_2(\tau)\,\mathrm{d}\tau,
\end{aligned}
\tag{9.1}
$$

where

$$
G^{\text{cor}}(t) = \begin{cases} f^{\text{cor}}(t), & \text{if } -\beta \le t \le 0, \\ f^{\text{cor}}(-t), & \text{if } 0 \le t \le \beta, \\ 0, & \text{if } |t| \ge \beta. \end{cases}
\tag{9.2}
$$

In (9.1) and (9.2), F_{∞}^{cor} is the long-range value of the appropriate correlation, $\overline{F}^{\text{cor}}(\rho)$ — its part that decays at infinity; $F_{\text{ws}}^{\text{cor}}$ is the contribution to the latter, generated by the well-stirred part (3.2) of the radial distribution function $g(r)$ for the set of sphere's centers, and $\widetilde{F}^{\text{cor}}(\rho)$ is due to the "deviation" $\widetilde{g}(r)$ of $g(r)$ from the well-stirred one, cf. (3.1) (recall that $\widetilde{g}(r) = \nu_2(r)$ if $r \ge 2a$, see (3.4)). The parameter β takes the values 1 or 2, depending on the kind of correlation under study. We note also that

$$
G^{\text{cor}}(t) = f^{\text{cor}}(t), \quad \text{if} \quad |t| \le \beta,
$$

provided $f^{\text{cor}}(t)$ is even, which is the case with "particle-center" and "surface-center" correlations (for which $\beta = 1$), see (3.13) and (4.6).

For the sake of completeness, the function $f^{\text{cor}}(t)$ for the "particle-particle" correlation $\overline{F}^{\text{pp}}(x)$ is also given, see [7]. In this case, the well-stirred contribution reads

$$
\overline{F}_{\text{ws}}^{\text{pp}}(r) = \begin{cases} 1 - \dfrac{3\rho}{4(1-\eta_1)} + \dfrac{(1+3\eta_1)\rho^3}{16(1-\eta_1)} - \dfrac{9\eta_1\rho^4}{160(1-\eta_1)} \\ \quad + \dfrac{\eta_1\rho^6}{2240(1-\eta_1)}, & \text{if } 0 \le \rho \le 2, \\ \dfrac{\eta_1}{1-\eta_1}\,\dfrac{(\rho-4)^4\,(36 - 34\,\rho - 16\,\rho^2 - \rho^3)}{2240\rho}, & \text{if } 2 \le \rho \le 4, \\ 0, & \text{if } \rho \ge 4, \end{cases}
\tag{9.3}
$$

see once again [7] for details and references.

Another set of useful formulae, derived in the paper, concerns the moments

$$
\theta_k^{\text{cor}} = \int_0^{\infty} \rho^k\,\overline{F}^{\text{cor}}(\rho)\,\mathrm{d}\rho, \quad k = 1, 2, \ldots,
\tag{9.4}
$$

TABLE 1. Notations, parameters and functions in the integral representations (9.1) of the various two-point correlations

Correlation	Notation	$F_\infty^{\rm cor}$	$\overline{F}_{\rm ws}^{\rm cor}(r)$	$f^{\rm cor}(t)$	β
center-center	$F^{\rm cc}$	n^2	$n\delta(x) - n^2 h_{2a}(x)$	–	–
particle-center	$F^{\rm pc}$	$n\eta_1$	Eq. (3.9)	$\frac{3}{4}(1-t^2)$	1
surface-center	$F^{\rm sc}$	nS	Eq. (4.4)	$\frac{1}{2}$	1
particle-particle	$F^{\rm pp}$	η_1^2	Eq. (9.3)	$\frac{3}{160}(2+t)^3(4-6t+t^2)$	2
surface-particle	$F^{\rm sp}$	η_1^2	Eq. (5.16)	$\frac{1}{8}(2+t)^2(1-t)$	2
surface-surface	$F^{\rm ss}$	η_1^2	Eq. (6.9)	$\frac{1}{4}(2+t)$	2

of the two-point correlations (2.9) and (2.18). For an arbitrary k, they can be evaluated by means of the representations (9.1), summarized in Table 1, and thus interconnected to the appropriate moments (3.18) of the binary correlation. In the cases $k = 1$ and $k = 2$, which seem to be most interesting for applications, evaluation of (9.4) does not need however the aforementioned representations, but can be done directly, using, as a matter of fact, just their definitions. This was illustrated in Section 7. The results, concerning $\theta_2^{\rm cor}$ (in 3-D) and $\theta_1^{\rm cor}$ (in 2-D), can be concisely summarized in the simple formulae

$$\begin{aligned} \theta_2^{\rm cor} &= F_\infty^{\rm cor}\left(\frac{1-8\eta_1}{3\eta_1} + m_2\right) \quad \text{in 3-D,} \\ \theta_1^{\rm cor} &= F_\infty^{\rm cor}\left(\frac{1-2\eta_1}{2\eta_1} + m_1\right) \quad \text{in 2-D,} \end{aligned} \tag{9.5}$$

where $F_\infty^{\rm cor}$ are the long-range values of the appropriate correlation, see Table 1 and Eqs. (7.1), (7.2), (3.19), (4.10), (6.15), (7.5) and (7.6).

In 3-D the moments $\theta_1^{\rm cor}$ have a form, similar to (9.5):

$$\theta_1^{\rm cor} = F_\infty^{\rm cor}\left(T_1^{\rm cor}(\eta_1) + m_1\right), \tag{9.6}$$

but now the functions $T_1^{\rm cor}(\eta_1)$ are specific for different correlations. They are listed in Table 2, in which the foregoing formulae (3.19), (4.10), (6.15) and (7.10) are simply put together.

TABLE 2. The functions $T_1^{\mathrm{cor}}(\eta_1)$ in Eq. (9.6) for the various two-point correlations

Correlation	F^{pc}	F^{sc}	F^{pp}	F^{sp}	F^{ss}
$T_1^{\mathrm{cor}}(\eta_1)$	$\frac{5-19\eta_1}{10\eta_1}$	$\frac{1-11\eta_1/2}{3\eta_1}$	$\frac{2-9\eta_1}{5\eta_1}$	$\frac{5-26\eta_1}{15\eta_1}$	$\frac{1-5\eta_1}{3\eta_1}$

ACKNOWLEDGEMENTS. Special thanks are due to D.R.S. Talbot who provided the author with his unpublished notes concerning coincidence between Doi's and Talbot & Willis' bound for dispersions of spheres. The support of the Bulgarian Ministry of Education, Science and Technology under Grant No MM 416-94, is gratefully acknowledged as well.

REFERENCES

1. Beasley, J. D., S. Torquato. New bounds on the permeability of a random array of spheres. *Phys. Fluids*, **A 1**, 1989, 199–207.
2. Berryman, J. Computing variational bounds for flow through random aggregates of spheres. *J. Comp. Physics*, **52**, 1983, 142–162.
3. Doi, M. A new variational approach to the diffusion and the flow problem in porous media. *J. Phys. Soc. Japan*, **40**, 1976, 567–572.
4. Felderhof, B. U., J. M. Deutch. Concentration dependence of the rate of diffusion-controlled reactions. *J. Chem. Phys.*, **64**, 1976, 4551–4558.
5. Markov, K. Z. On a statistical parameter in the theory of random dispersions of spheres. In: *Continuum Models of Discrete Systems, Proc. 8th Int. Symposium* (Varna, 1995), K. Z. Markov, ed., World Sci., 1996, 241–249.
6. Markov, K. Z. On a two-point correlation function in random dispersions and an application. In: *Continuum Models and Discrete Systems, Proc. 9th Int. Symposium* (Istanbul, 1998), E. Inan and K. Z. Markov, eds., World Sci., 1998, 206–215.
7. Markov, K. Z., J. R. Willis. On the two-point correlation function for dispersions of nonoverlapping spheres. *Mathematical Models and Methods in Applied Sciences.* **8**, 1998, 359–377.
8. Richards, P. M., S. Torquato. Upper and lower bounds for the rate of diffusion-controlled reactions. *J. Chem. Phys.*, **87**, 1987, 4612–4614.
9. Rubinstein, J., S. Torquato. Diffusion-controlled reactions: Mathematical formulation, variational principles, and rigorous bounds. *J. Chem. Phys.*, **88**, 1988, 6372–6380.
10. Rubinstein, J., S. Torquato. Flow in random porous media: Mathematical formulation, variational principles, and rigorous bounds. *J. Fluid Mech.*, **206**, 1989, 25–46.
11. Stratonovich, R. L. Topics in Theory of Random Noises. Vol. 1, Gordon and Breach, New York, 1963.
12. Talbot, D. R. S., J. R. Willis. The effective sink strength of a random array of voids in irradiated material. *Proc. R. Soc. London*, **A370**, 1980, 351–374.
13. Torquato, S. Concentration dependence of diffusion-controlled reactions among static reactive sinks. *J. Chem. Phys.*, **85**, 1986, 7178–7179.

14. Torquato, S. Interfacial surface statistics arising in diffusion and flow problems in porous media. *J. Chem. Phys.*, **85**, 1986, 4622–4628.

15. Torquato, S. Microstructure characterization and bulk properties of disordered two-phase media. *J. Stat. Phys.*, **45**, 1986, 843–873.

16. Torquato, S., J. Rubinstein. Diffusion-controlled reactions: II. Further bounds on the rate constant. *J. Chem. Phys.*, **90**, 1989, 1644–1647.

17. Vanmarcke, E. Random Fields: Analysis and Synthesis. MIT Press, Cambridge, Massachusetts and London, England, 1983.

18. Verlet, L., J.-J. Weis. Equilibrium theory of simple liquids. *Phys. Rev. A*, **5**, 1972, 939–952.

19. Willis, J. R. A polarization approach to the scattering of elastic waves. II. Multiple scattering from inclusions. *J. Mech. Phys. Solids*, **28**, 1980, 307–327.

Received January 28, 1998
Revised May 28, 1998

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 Blvd J. Bourchier, P.O. Box 48
BG-1164 Sofia, Bulgaria
E-mail: kmarkov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 91, 1997

A NEW APPROACH FOR DERIVING C^2-BOUNDS ON THE EFFECTIVE CONDUCTIVITY OF RANDOM DISPERSIONS

KRASSIMIR D. ZVYATKOV

A new variational procedure for evaluating the effective conductivity of a dilute random dispersion of spheres is proposed. The classical variational principles are employed, in which a class of trial fields in the form of suitably truncated factorial series is introduced. In general, this class leads to a rigorous formula for the effective conductivity, which is correct to the order "square of sphere fraction," and makes use of the disturbance to the temperature field in an unbounded matrix, generated by two spherical inhomogeneities. The basic idea in the present study consists in replacing this "two-sphere" field by a superposition of disturbances, generated by the same two spheres, but considered as single already, together with the disturbance due to another single sphere, centered between them and radially inhomogeneous. In this way new variational bounds on the effective conductivity are derived and discussed in more detail for a special choice of the middle sphere's properties. The obtained bounds improve, in particular, on the known three-point bounds on the effective conductivity of the dispersion.

Keywords: random media, dispersions of spheres, variational bounds, effective conductivity

1991/95 Math. Subject Classification: 60G60, 60H15, 49K45

1. INTRODUCTION

Consider a statistically homogeneous dispersion of equi-sized nonoverlapping spheres of conductivity κ_f and radii a, immersed at random into a matrix of conductivity κ_m. In the heat conductivity context and absence of body sources, the temperature field, $\theta(\mathbf{x})$, in the dispersion is governed by the equations

$$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0, \quad \mathbf{q}(\mathbf{x}) = \kappa(\mathbf{x})\nabla\theta(\mathbf{x}), \quad \langle\nabla\theta(\mathbf{x})\rangle = \mathbf{G}, \tag{1.1}$$

where $\kappa(\mathbf{x})$ is the random conductivity field of the medium, $\mathbf{q}(\mathbf{x})$ — the heat flux vector, $\mathbf{G}$ is the prescribed macroscopic value of the temperature gradient, and the brackets $\langle\cdot\rangle$ denote statistical averaging [1]. Since the field $\kappa(\mathbf{x})$ takes the values κ_f or κ_m depending on whether $\mathbf{x}$ lies in a sphere or in the matrix respectively, it allows the representation

$$\kappa(\mathbf{x}) = \langle\kappa\rangle + [\kappa]\int h(\mathbf{x}-\mathbf{y})\psi'(\mathbf{y})\, d^3\mathbf{y}, \tag{1.2}$$

where $[\kappa] = \kappa_f - \kappa_m$, $h(\mathbf{x})$ is the characteristic function of a single sphere of radius a located at the origin, and $\psi'(\mathbf{x})$ is the fluctuating part of the random density field

$$\psi(\mathbf{x}) = \sum_j \delta(\mathbf{x}-\mathbf{x}_j),$$

generated by the random field $\{\mathbf{x}_j\}$ of sphere's centers [2]. The integrals hereafter are over the whole $\mathbb{R}^3$ if the integration domain is not explicitly indicated.

The solution of Eq. (1.1) is understood in a statistical sense, so that one is to evaluate all multipoint moments (correlation functions) of $\theta(\mathbf{x})$ and the joint moments of $\kappa(\mathbf{x})$ and $\theta(\mathbf{x})$, see, e.g., [1]. Let c be the volume fraction of the spheres, then $n = c/V_a$ is their number density. As discussed in [3–5], the solution $\theta(\mathbf{x})$ of the random problem (1.1), asymptotically valid to the order c^2, can be found in the form of truncated functional series:

$$\begin{aligned}\theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int T_1(\mathbf{x}-\mathbf{y})D_\psi^{(1)}(\mathbf{y})\, d^3\mathbf{y} \\ + \int\int T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)D_\psi^{(2)}(\mathbf{y}_1,\mathbf{y}_2)\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2,\end{aligned} \tag{1.3}$$

where T_1 and T_2 are certain non-random kernels and the fields

$$\begin{aligned}D_\psi^{(0)} = 1, \quad D_\psi^{(1)}(\mathbf{y}) = \psi'(\mathbf{y}), \quad D_\psi^{(2)}(\mathbf{y}_1,\mathbf{y}_2) = \psi(\mathbf{y}_1)[\psi(\mathbf{y}_2) - \delta(\mathbf{y}_1-\mathbf{y}_2)] \\ -ng_0(\mathbf{y}_1-\mathbf{y}_2)[D_\psi^{(1)}(\mathbf{y}_1) + D_\psi^{(1)}(\mathbf{y}_2)] - n^2 g_0(\mathbf{y}_1-\mathbf{y}_2)\end{aligned} \tag{1.4}$$

are the first three terms in the c^2-orthogonal system, formed as a result of the appropriate virial orthogonalization, see again [3–5] for details and discussion. In Eq. (1.4) $g_0(r)$ is the leading part of the well-known radial distribution function $g(r) = f_2(r)/n^2$ for the dispersion in the dilute case $n \to 0$, i.e. $g(r) = g_0(r) + O(n)$; $f_2(r)$ denotes the two-point probability density for the set of sphere centers and $r = |\mathbf{y}_1 - \mathbf{y}_2|$.

The identification of the kernels T_1 and T_2 is performed in [4] and [5] by means of a procedure, proposed by Christov and Markov [6]. It consists in inserting the truncated series (1.3) into the random equation (1.1), multiplying the result by the fields $D_\psi^{(p)}$, $p = 0, 1, 2$, and averaging the results. In this way a certain system of integro-differential equations for the needed kernels of the truncated series can be straightforwardly derived. The solution is analytically obtained in [4] and hence the full statistical solution of the problem (1.1), asymptotically correct to the order c^2, is known. In particular, this solution allows one to derive the effective conductivity κ^* of the dispersion, to the same order c^2, through evaluating the one-point moment

$$\langle \kappa(\mathbf{x}) \nabla\theta(\mathbf{x}) \rangle = \kappa^* \langle \nabla\theta(\mathbf{x}) \rangle = \kappa^* \mathbf{G}.$$

As a result, the renormalized c^2-formula of Jeffrey [7] for the effective conductivity of the dispersion was rederived, but with rigorous justification of the integration mode in the appropriate conditionally convergent integrals.

As shown in [8], the same result is obtained when the truncated series (1.3) are employed as trial fields in the classical variational principle, corresponding to the problem (1.1):

$$W_A[\theta(\cdot)] = \left\langle \kappa(\mathbf{x}) |\nabla\theta(\mathbf{x})|^2 \right\rangle \longrightarrow \min, \quad \langle \nabla\theta(\mathbf{x}) \rangle = \mathbf{G}, \tag{1.5}$$

$\min W_A = \kappa^* G^2$, see, e.g., [1]. Moreover, the leading parts in the virial expansions

$$T_1(\mathbf{x}) = T_1(\mathbf{x}; n) = T_{1,0}(\mathbf{x}) + T_{1,1}(\mathbf{x})n + \cdots, \tag{1.6}$$

$$T_2(\mathbf{x}, \mathbf{y}) = T_2(\mathbf{x}, \mathbf{y}; n) = T_{2,0}(\mathbf{x}, \mathbf{y}) + T_{2,1}(\mathbf{x}, \mathbf{y})n + \cdots \tag{1.7}$$

of the optimal kernels T_1 and T_2 suffice to determine the effective conductivity κ^* to the order c^2. In this way the equations for the virial coefficients $T_{1,0}$ and $T_{2,0}$, already found in [4], have been rederived. It turned out that $T_{1,0}(\mathbf{x})$ coincides with the disturbance $T^{(1)}(\mathbf{x})$ to the temperature field $\mathbf{G} \cdot \mathbf{x}$ in an unbounded matrix, introduced by a single spherical inhomogeneity, located at the origin:

$$T_{1,0}(\mathbf{x}) = T^{(1)}(\mathbf{x}) = 3\beta G \cdot \nabla\varphi(\mathbf{x}), \tag{1.8}$$

where $\varphi(\mathbf{x}) = \varphi(\mathbf{x}, a) = h * \dfrac{1}{4\pi|\mathbf{x}|}$ is the Newtonian potential for the single sphere of the radius a and $\beta = [\kappa]/(\kappa_f + 2\kappa_m)$. For the coefficient $T_{2,0}$ one has

$$2T_{2,0}(\mathbf{x}, \mathbf{x} - \mathbf{z}) = T^{(2)}(\mathbf{x}; \mathbf{z}) - T^{(1)}(\mathbf{x}) - T^{(1)}(\mathbf{x} - \mathbf{z}), \tag{1.9}$$

where $T^{(2)}(\mathbf{x}; \mathbf{z})$ is the disturbance to the temperature field $\mathbf{G} \cdot \mathbf{x}$ in an unbounded matrix of conductivity κ_m, generated by a pair of spherical inhomogeneities of conductivity κ_f, centered at the origin and at the point $\mathbf{z}$, $|\mathbf{z}| > 2a$.

It is important to point out that the variational derivation, involving the truncated series (1.3), leads to a c^2-formula for the effective conductivity that contains absolutely convergent integrals solely. Namely, let

$$\frac{\kappa^*}{\kappa_m} = 1 + 3\beta c + a_{2\kappa} c^2 + \cdots, \quad a_{2\kappa} = 3\beta^2 + a'_{2\kappa}, \tag{1.10}$$

be the virial expansion of κ^*. For the c^2-deviation $a'_{2\kappa}$. from the well-known Maxwell formula one has

$$a'_{2\kappa} G^2 = \frac{[\kappa]}{\kappa_m} \frac{1}{V_a^2} \int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y}) \nabla_x T^{(1)}(\mathbf{x}-\mathbf{y}) \cdot \nabla_x T^{(2)}(\mathbf{x};\mathbf{y})\, d^3\mathbf{y}, \tag{1.11}$$

where $V_a = \frac{4}{3}\pi a^3$. (See also [9,10], where the Hashin-Shtrikman variational principle was employed to derive the same formula (1.11).)

In order to calculate the c^2-coefficient $a_{2\kappa}$, one needs the field $T^{(2)}(\mathbf{x};\mathbf{z})$. The latter can be explicitly found, e.g. by means of the method of twin expansions. The calculations, based on this solution and the formula (1.11), however, will be not simpler than the ones in the well-known works [7] and [11], based on the "renormalized" formula of Jeffrey [7]. That is why our aim here is to look for an appropriate approximation for the field $T^{(2)}(\mathbf{x};\mathbf{z})$ which, when combined with (1.9), will produce a class of trial field in the form (1.3). However, this class will be narrower than (1.3) and as a result certain variational bounds on $a_{2\kappa}$ will follow only.

Consider first the simplest case when the kernel T_1 in (1.3) is adjustable and the kernel T_2 vanishes: $T_2 = 0$, i.e.

$$\theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int T_1(\mathbf{x}-\mathbf{y}) D_\psi^{(1)}(\mathbf{y})\, d^3\mathbf{y}. \tag{1.12}$$

This class has been introduced and discussed in detail by Markov in [12], where it is shown that minimizing the functional $W_A[\theta(\cdot)]$ over the class (1.12) gives the best three-point upper bound $\kappa^{(3)}$ on the effective conductivity κ^*, i.e. the most restrictive one which uses three-point statistical information for the medium. According to (1.9), this bound corresponds to the approximation

$$T^{(2)}(\mathbf{x};\mathbf{z}) \approx T^{(1)}(\mathbf{x}) + T^{(1)}(\mathbf{x}-\mathbf{z}) \tag{1.13}$$

of the disturbance $T^{(2)}(\mathbf{x};\mathbf{z})$. We will come back to the three-point bounds again in Section 2.1.

Obviously, the approximation (1.13) is appropriate when the two spheres are far away, i.e. $|\mathbf{z}| \gg 2a$. Here we propose an improvement of this approximation that consists in adding the disturbance $\widetilde{T}^{(1)}(\mathbf{x}-\mathbf{z}/2)$ to the adjustable temperature field $\mathbf{\Phi}(\mathbf{z})\cdot\mathbf{x}$, generated by a single radial inhomogeneous sphere, centered between two spheres, i.e. at the point $\mathbf{z}/2$. Thus, we assume the approximation

$$T^{(2)}(\mathbf{x};\mathbf{z}) \approx \widetilde{T}^{(1)}\left(\mathbf{x}-\frac{\mathbf{z}}{2}\right) + T^{(1)}(\mathbf{x}) + T^{(1)}(\mathbf{x}-\mathbf{z}). \tag{1.14}$$

This idea is suggested by some successful models in the theory of dispersions, see, e.g., [13] and [14], where, in fact, the interactions of the spheres are taken into account by introducing a single radial inhomogeneous sphere, immersed into effective medium. According to (1.9), the approximation (1.14) leads to the following choice of the kernel T_2 in (1.3):

$$T_2(\mathbf{x}-\mathbf{y}_1,\mathbf{x}-\mathbf{y}_2)=\frac{1}{2}\widetilde{T}^{(1)}\left(\mathbf{x}-\frac{\mathbf{y}_1+\mathbf{y}_2}{2}\right). \tag{1.15}$$

In Section 2.2 a new variational procedure will be considered. It is based on the possibility to vary both the field $\mathbf{\Phi}(\mathbf{z})$ and the conductivity distribution of the middle sphere. Its counterpart that yields lower bounds will be discussed in Section 2.3. Finally, in Section 3 a simple case will be considered, when the middle sphere is homogeneous and encompasses the other two spheres. This case allows us to obtain quite easily explicit results, which will be then compared with some of the known variational bounds.

2. THE VARIATIONAL PROCEDURE

The disturbance $\widetilde{T}^{(1)}(\mathbf{x}-\mathbf{z}/2)$ to the temperature field $\mathbf{\Phi}(\mathbf{z})\cdot\mathbf{x}$, generated by a single radial inhomogeneous sphere, centered at the point $\mathbf{z}/2$, has the form

$$\widetilde{T}^{(1)}\left(\mathbf{x}-\frac{\mathbf{z}}{2}\right)=\mathbf{\Phi}(\mathbf{z})\cdot\nabla f\left(\mathbf{x}-\frac{\mathbf{z}}{2},\mathbf{z}\right), \tag{2.1}$$

where $f(\mathbf{w},\mathbf{z})=f(|\mathbf{w}|,\mathbf{z})$ is a function, specified by the radial distribution of the conductivity coefficient of the sphere. The dependence of $f(\mathbf{w},\mathbf{z})$ on its second argument $\mathbf{z}$ indicates explicitly the possibility that the latter distribution is arbitrary for the moment. Hereafter the differentiation of the function $f(\mathbf{w},\mathbf{z})$ is with respect to its first argument, $\nabla=\nabla_w$.

According to (1.15) and (2.1), we should employ the classical variation principle (1.5) over the class of trial field (1.3), provided the kernel T_2 has the form

$$T_2(\mathbf{x},\mathbf{x}-\mathbf{z})=\frac{1}{2}\mathbf{\Phi}(\mathbf{z})\cdot\nabla f\left(\mathbf{x}-\frac{\mathbf{z}}{2},\mathbf{z}\right),$$

i.e.

$$T_2(\mathbf{x}-\mathbf{y}_1,\mathbf{x}-\mathbf{y}_2)=\frac{1}{2}\mathbf{\Phi}(\mathbf{y}_2-\mathbf{y}_1)\cdot\nabla f\left(\mathbf{x}-\frac{\mathbf{y}_1+\mathbf{y}_2}{2},\mathbf{y}_2-\mathbf{y}_1\right). \tag{2.2}$$

Here the kernel $T_1(\mathbf{y})$, the functions $\mathbf{\Phi}$ and f are adjustable. To this end it is appropriate to remind briefly the variational procedure, connected with the derivation of the so-called optimal three-point bounds.

Making use of Eq. (1.2) and the formulae for the moments of the random density field $\psi(\mathbf{x})$, we find an expression for the restriction $W_A^{(1)}[T_1(\cdot)]$ of the functional W_A over the class (1.12), see [12] for details. The optimal kernel $T_1(\mathbf{x})$, i.e. the solution of the Euler-Lagrange equation for the functional $W_A^{(1)}$, is looked for in the virial form (1.6). This representation of $T_1(\mathbf{x})$ generates the appropriate virial expansion of the restriction $W_A^{(1)}[T_1(\cdot)]$, namely,

$$W_A^{(1)}[T_1(\cdot)] = \langle\kappa\rangle\, G^2 + W_A^{(1,1)}[T_{1,0}(\cdot)]n + W_A^{(1,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot)]n^2 + \cdots, \qquad (2.3)$$

see [15, Eqs. (4.2)-(4.5)]. An analysis of the coefficient $W_A^{(1,1)}$ shows that

$$\delta W_A^{(1,1)}[T_{1,0}(\cdot)] = 0 \Longleftrightarrow T_{1,0}(\mathbf{x}) = T^{(1)}(\mathbf{x}), \qquad (2.4)$$

where $T^{(1)}(\mathbf{x})$ is the disturbance (1.8), generated by a single spherical inhomogeneity. It turns out, however, that at $T_{1,0}(\mathbf{x}) = T^{(1)}(\mathbf{x})$ the virial coefficient $W_A^{(1,2)}$ does not depend on $T_{1,1}(\mathbf{x})$, i.e.

$$W_A^{(1,2)}[T^{(1)}(\cdot), T_{1,1}(\cdot)] = \overline{W}_A^{(1,2)}[T^{(1)}(\cdot)] = 3\beta^2\kappa_m\left(1 + \frac{[\kappa]}{\kappa_m}m_2\right)V_a^2 G^2, \qquad (2.5)$$

where

$$m_2 = m_2[g_0(\cdot)] = 2\int_2^\infty \frac{\lambda^2}{(\lambda^2-1)^3}\, g_0(\lambda a)d\lambda, \quad \lambda = |\mathbf{y}|/a, \qquad (2.6)$$

is a statistical parameter for the dispersion, introduced in [12]. Hence, according to Eq. (2.3), we have for the optimal upper tree-point bound $\kappa^{(3)}$

$$\kappa^* G^2 \le \kappa^{(3)} G^2 = \langle\kappa\rangle\, G^2 + \frac{1}{V_a}W_A^{(1,1)}[T^{(1)}(\cdot)]c + \frac{1}{V_a^2}\overline{W}_A^{(1,2)}[T^{(1)}(\cdot)]c^2 + o(c^2). \ (2.7)$$

On the base of this analysis it is shown in [15] that the Beran's bounds [16] are c^2-optimal in the above explained sense. Eqs. (2.5) and (2.7) yield straightforwardly the following estimate for the c^2-coefficient $a_{2\kappa}$ in the virial expansion (1.10) of κ^* (see [12, 15]) :

$$a_{2\kappa} \le a_{2\kappa}^u, \quad a_{2\kappa}^u = 3\beta^2\left(1 + \frac{[\kappa]}{\kappa_m}m_2\right). \qquad (2.8)$$

Let us note that the formula (2.8) for the upper bound $a_{2\kappa}^u$ can be obtained also if we insert (1.13) into (1.11), taking into account (1.10) and the identities

$$\int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y})\nabla T^{(1)}(\mathbf{x}-\mathbf{y})\cdot\nabla T^{(1)}(\mathbf{x})\, d^3\mathbf{x} = 0, \qquad (2.9a)$$

$$\int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y})|\nabla T^{(1)}(\mathbf{x}-\mathbf{y})|^2\, d^3\mathbf{y} = 3\beta^2 V_a^2 m_2. \qquad (2.9b)$$

It is interesting to point out that inserting (1.13) into the "renormalized" formula of Jeffrey [7] leads, however, to the Maxwell c^2-value $a_{2\kappa} = 3\beta^2$ that corresponds to the Hashin-Shtrikman bound.

2.2. NEW UPPER BOUND FOR THE DISPERSION

Using the formulae for the moments of the fields $D_\psi^{(1)}$ and $D_\psi^{(2)}$, see [4, Eqs. (3.4)], the restriction $W_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)]$ of the functional W_A over the general class (1.3) becomes

$$W_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)] = W_A^{(1)}[T_1(\cdot)] + \widetilde{W}_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)], \tag{2.10}$$

where

$$\begin{aligned}\widetilde{W}_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)] &= 2n^2\kappa_m \iint g_0(\mathbf{y}_1 - \mathbf{y}_2)\,|\nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)|^2\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2\\
&+ 2n^2[\kappa]\iint g_0(\mathbf{y}_1 - \mathbf{y}_2)\big[h(\mathbf{x}-\mathbf{y}_1) + h(\mathbf{x}-\mathbf{y}_2)\big]\,|\nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)|^2\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2\\
&+ 2n^2[\kappa]\iint g_0(\mathbf{y}_1 - \mathbf{y}_2)\Big[h(\mathbf{x}-\mathbf{y}_1)\nabla T_1(\mathbf{x}-\mathbf{y}_2) + h(\mathbf{x}-\mathbf{y}_2)\nabla T_1(\mathbf{x}-\mathbf{y}_1)\Big]\\
&\cdot \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2 + o(n^2),\end{aligned} \tag{2.11}$$

see [8, Section 3]; here we have used the fact that the kernel $T_2(\mathbf{y}_1, \mathbf{y}_2)$ is a symmetric function of its arguments.

Let us consider now the narrower class (1.3) when the kernel T_2 has the form (2.2). Then, according to Eqs. (2.3) and (2.10), for the restriction $W_A^{(2)}[T_1(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)]$ of the functional W_A over this class we get

$$\begin{aligned}W_A^{(2)}[T_1(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)] &= \langle\kappa\rangle G^2 + W_A^{(1,1)}[T_{1,0}(\cdot)]\, n\\
&+ W_A^{(2,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)]\, n^2 + o(n^2),\end{aligned} \tag{2.12}$$

where

$$\begin{aligned}&W_A^{(2,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)]\\
&= W_A^{(1,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot)] + \widehat{W}_A^{(2)}[T_{1,0}(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)],\end{aligned} \tag{2.13a}$$

$$\widehat{W}_A^{(2)}[T_{1,0}(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)] = \frac{1}{n^2}\widetilde{W}_A^{(2)}\left[T_{1,0}(\cdot), \frac{1}{2}\boldsymbol{\Phi}(\cdot)\cdot\nabla f(\cdot,\cdot)\right]. \tag{2.13b}$$

Here $W_A^{(1,1)}$ and $W_A^{(1,2)}$ are the virial coefficients from Eq. (2.3) for which, let us recall, Eqs. (2.4) and (2.5) hold. Hence, the minimization of the functional $W_A^{(2)}$ is reduced to that of the functional

$$\widehat{W}_A^{(2)\dagger}[\boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)] = \widehat{W}_A^{(2)}\left[T^{(1)}(\cdot), \boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)\right]. \tag{2.14}$$

Taking into account Eqs. (1.8) and (2.13b), after an appropriate change of integrand variables in (2.11), we find the following form of functional (2.14):

$$
\begin{aligned}
&\widehat{W}_A^{(2)\dagger}\left[\boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)\right] \\
&= \frac{1}{2}\iint g_0(\mathbf{z})\left[\kappa_m + 2[\kappa]h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\right]\left|\boldsymbol{\Phi}(\mathbf{z})\cdot\nabla\nabla f(\mathbf{w},\mathbf{z})\right|^2 d^3\mathbf{w}\,d^3\mathbf{z} \qquad (2.15)\\
&+ 6\beta[\kappa]\mathbf{G}\cdot\iint g_0(\mathbf{z})h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\nabla\nabla\varphi\left(\mathbf{w}+\frac{\mathbf{z}}{2}\right)\cdot\nabla\nabla f\left(\mathbf{w},\mathbf{z}\right)\cdot\boldsymbol{\Phi}(\mathbf{z})\,d^3\mathbf{w}\,d^3\mathbf{z}\,.
\end{aligned}
$$

The minimizing functions $\boldsymbol{\Phi}$ and f satisfy the Euler-Lagrange equations

$$
\delta_{\Phi}\widehat{W}_A^{(2)\dagger} = 0, \qquad \delta_f\widehat{W}_A^{(2)\dagger} = 0\,. \tag{2.16}
$$

The first of these equations yields straightforwardly

$$
\begin{aligned}
&\boldsymbol{\Phi}(\mathbf{z})\cdot\int\left[\kappa_m + 2[\kappa]h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\right]\nabla\nabla f(\mathbf{w},\mathbf{z})\cdot\nabla\nabla f(\mathbf{w},\mathbf{z})\,d^3\mathbf{w} \\
&= -6\beta[\kappa]\mathbf{G}\cdot\int h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\nabla\nabla\varphi\left(\mathbf{w}+\frac{\mathbf{z}}{2}\right)\cdot\nabla\nabla f\left(\mathbf{w},\mathbf{z}\right)\,d^3\mathbf{w}
\end{aligned} \tag{2.17}
$$

at $|\mathbf{z}| > 2a$, whose solution $\boldsymbol{\Phi}(\mathbf{z})$ can be easily found for a given function f. Taking into account that $f(\mathbf{w},\mathbf{z}) = f(|\mathbf{w}|,\mathbf{z})$, the second equation in (2.16) is recast as

$$
\begin{aligned}
&\Phi_i(\mathbf{z})\Phi_j(\mathbf{z})\int\limits_{\Omega_{1w}}\left\{\kappa_m\left(\Delta f(|\mathbf{w}|,z)\right)_{,ij} + 2[\kappa]\left(h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)f_{,ik}(|\mathbf{w}|,\mathbf{z})\right)_{,kj}\right\}dS_w \\
&= -6\beta[\kappa]G_i\Phi_j(\mathbf{z})\int\limits_{\Omega_{1w}}\left(h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\varphi_{,ik}\left(\mathbf{w}+\frac{\mathbf{z}}{2}\right)\right)_{,kj}dS_w
\end{aligned} \tag{2.18}
$$

at $|\mathbf{z}| > 2a$, where Ω_{1w} is the sphere $|\mathbf{w}| = 1$.

Eqs. (2.17) and (2.18) form a very complicated system of integro-differential equations for the optimal functions $\boldsymbol{\Phi}$ and f. That is why we shall consider a simpler procedure in which the function f is fixed.

Making use of Eq. (2.17), the minimum value of the functional $\widehat{W}_A^{(2)\dagger}$ can be recast now in the form in which the solution $\boldsymbol{\Phi}(\mathbf{z})$ of this equation enters linearly:

$$
\begin{aligned}
&\min_{\Phi}\widehat{W}_A^{(2)\dagger}\left[\boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)\right] \\
&= 3\beta[\kappa]\mathbf{G}\cdot\iint g_0(\mathbf{z})h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\nabla\nabla\varphi\left(\mathbf{w}+\frac{\mathbf{z}}{2}\right)\cdot\nabla\nabla f\left(\mathbf{w},\mathbf{z}\right)\cdot\boldsymbol{\Phi}(\mathbf{z})\,d^3\mathbf{w}\,d^3\mathbf{z}\,.
\end{aligned} \tag{2.19}
$$

With the notations

$$
\mathcal{R}(\mathbf{z}) = \frac{1}{V_a}\int\left[1 + 2\frac{[\kappa]}{\kappa_m}h\left(\mathbf{w}-\frac{\mathbf{z}}{2}\right)\right]\nabla\nabla f(\mathbf{w},\mathbf{z})\cdot\nabla\nabla f(\mathbf{w},\mathbf{z})\,d^3\mathbf{w}, \tag{2.20a}
$$

$$\mathcal{J}(\mathbf{z}) = \frac{1}{V_a} \int h\left(\mathbf{w} - \frac{\mathbf{z}}{2}\right) \nabla\nabla\varphi\left(\mathbf{w} + \frac{\mathbf{z}}{2}\right) \cdot \nabla\nabla f(\mathbf{w}, \mathbf{z})\, d^3\mathbf{w}, \tag{2.20b}$$

Eqs. (2.17) and (2.19) can be written in the form

$$\boldsymbol{\Phi}(\mathbf{z}) \cdot \mathcal{R}(\mathbf{z}) = -6\beta \frac{[\kappa]}{\kappa_m} \mathbf{G} \cdot \mathcal{J}(\mathbf{z}),$$

$$\min_{\Phi} \widehat{W}_A^{(2)\dagger} [\boldsymbol{\Phi}(\cdot), f(\cdot, \cdot)] = 3\beta[\kappa] V_a \mathbf{G} \cdot \int g_0(\mathbf{z}) \mathcal{J}(\mathbf{z}) \cdot \boldsymbol{\Phi}(\mathbf{z})\, d^3\mathbf{z}\,.$$

Thus the solution of Eq. (2.17) is

$$\boldsymbol{\Phi}(\mathbf{z}) = -6\beta \frac{[\kappa]}{\kappa_m} \mathbf{G} \cdot \mathcal{J}(\mathbf{z}) \cdot \mathcal{R}^{-1}(\mathbf{z}) \tag{2.21}$$

and the minimum value of the functional $\widehat{W}_A^{(2)\dagger}$ is

$$\min_{\Phi} \widehat{W}_A^{(2)\dagger} [\boldsymbol{\Phi}(\cdot), f(\cdot, \cdot)] = -\, 18\beta^2 \frac{[\kappa]^2}{\kappa_m} V_a \mathbf{G} \cdot \int g_0(\mathbf{z}) \mathcal{J}(\mathbf{z}) \cdot \mathcal{R}^{-1}(\mathbf{z}) \cdot \mathcal{J}(\mathbf{z})\, d^3\mathbf{z} \cdot \mathbf{G}\,. \tag{2.22}$$

Hence, according to Eqs. (2.4), (2.5), (2.7), (2.12)–(2.14), we obtain the following upper bound on the effective conductivity κ^*:

$$\kappa^* G^2 \leq \kappa^\dagger G^2, \quad \kappa^\dagger G^2 = \langle\kappa\rangle\, G^2 + \frac{1}{V_a} W_A^{(1,1)}[T^{(1)}(\cdot)]\, c \tag{2.23}$$

$$+ \frac{1}{V_a^2} \left\{ \overline{W}_A^{(1,2)}[T^{(1)}(\cdot)] + \min \widehat{W}_A^{(2)\dagger} \right\} c^2 + o(c^2) = \kappa^{(3)} + \frac{1}{V_a}^{2} \min \widehat{W}_A^{(2)\dagger}\, c^2 + o(c^2).$$

In turn, Eqs. (2.5), (2.22) and (2.23) yield straightforwardly an upper bound for the c^2-coefficient $a_{2\kappa}$ in the virial expansion (1.10) of κ^*, namely,

$$a_{2\kappa} \leq a_{2\kappa}^{u\dagger}, \quad a_{2\kappa}^{u\dagger} = 3\beta^2 \left(1 + \frac{[\kappa]}{\kappa_m} m_2 - \left(\frac{[\kappa]}{\kappa_m}\right)^2 \widetilde{m}_2^u \right), \tag{2.24}$$

where

$$\widetilde{m}_2^u = \widetilde{m}_2^u[g_0(\cdot), f(\cdot, \cdot), \alpha] = \frac{2}{V_a} \int g_0(\mathbf{z})\, \mathrm{tr}\left[\mathcal{J}(\mathbf{z}) \cdot \mathcal{R}^{-1}(\mathbf{z}) \cdot \mathcal{J}(\mathbf{z})\right]\, d^3\mathbf{z} \tag{2.25}$$

is a new statistical parameter for the dispersion, $\alpha = \kappa_f/\kappa_m$. This parameter depends not only on the leading part $g_0(r)$ of the radial distribution function g, but on the given function $f(\mathbf{w}, \mathbf{z})$ and on the ratio α for the dispersion as well, see Eqs. (2.20).

In order to obtain a similar lower bound on κ^*, we shall employ the classical dual variational principle for the problem (1.1), formulated with respect to the heat flux $\mathbf{q}(\mathbf{x}) = \nabla \times \mathbf{U}(\mathbf{x})$,

$$W_B[\mathbf{U}(\cdot)] = \left\langle k(\mathbf{x})|\nabla \times \mathbf{U}(\mathbf{x})|^2 \right\rangle \longrightarrow \min, \quad \langle \mathbf{q}(\mathbf{x}) \rangle = \mathbf{Q}, \tag{2.26}$$

$\min W_B = k^* Q^2$, $k^* = 1/\kappa^*$. The compliance field $k(\mathbf{x}) = 1/\kappa(\mathbf{x})$ has the form (1.2), i.e.

$$k(\mathbf{x}) = \langle k \rangle + [k] \int h(\mathbf{x}-\mathbf{y})\psi'(\mathbf{y})\, d^3\mathbf{y}, \quad [k] = k_f - k_m. \tag{2.27}$$

Similarly to the above-performed analysis, consider the functional W_B over the class of trial field

$$\begin{aligned} \mathbf{U}(\mathbf{x}) = \frac{1}{2}\mathbf{Q} \times \mathbf{x} + \int \mathbf{S}_1(\mathbf{x}-\mathbf{y}) D_\psi^{(1)}(\mathbf{y})\, d^3\mathbf{y} \\ + \iint \mathbf{S}_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2) D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2)\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2 \end{aligned} \tag{2.28}$$

— the counterpart of the class (1.3). Similarly, if the kernels $\mathbf{S}_1$ and $\mathbf{S}_2$ are arbitrary adjustable functions, the class (2.28) leads to the exact c^2-value of the effective compliance k^*, as it was the case with the effective conductivity κ^*. For the restriction $W_B^{(2)}[\mathbf{S}_1(\cdot), \mathbf{S}_2(\cdot,\cdot)]$ of the functional W_B over this class one has

$$W_B^{(2)}[\mathbf{S}_1(\cdot), \mathbf{S}_2(\cdot,\cdot)] = W_B^{(1)}[\mathbf{S}_1(\cdot)] + \widetilde{W}_B^{(2)}[\mathbf{S}_1(\cdot), \mathbf{S}_2(\cdot,\cdot)], \tag{2.29}$$

where $W_B^{(1)}[\mathbf{S}_1(\cdot)]$ is the restriction of W_B over the class

$$\mathbf{U}(\mathbf{x}) = \frac{1}{2}\mathbf{Q} \times \mathbf{x} + \int \mathbf{S}_1(\mathbf{x}-\mathbf{y}) D_\psi^{(1)}(\mathbf{y})\, d^3\mathbf{y} \tag{2.30}$$

and

$$\begin{aligned} \widetilde{W}_B^{(2)}[\mathbf{S}_1(\cdot), \mathbf{S}_2(\cdot,\cdot)] &= 2n^2 k_m \iint g_0(\mathbf{y}_1-\mathbf{y}_2)\, |\nabla_x \times \mathbf{S}_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)|^2\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2 \\ &+ 2n^2[k] \iint g_0(\mathbf{y}_1-\mathbf{y}_2)\left[h(\mathbf{x}-\mathbf{y}_1) + h(\mathbf{x}-\mathbf{y}_2)\right] |\nabla_x \times \mathbf{S}_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)|^2\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2 \\ &+ 2n^2[k] \iint g_0(\mathbf{y}_1-\mathbf{y}_2)\Big[h(\mathbf{x}-\mathbf{y}_1)\nabla \times \mathbf{S}_1(\mathbf{x}-\mathbf{y}_2) + h(\mathbf{x}-\mathbf{y}_2)\nabla \times \mathbf{S}_1(\mathbf{x}-\mathbf{y}_1)\Big] \\ &\cdot \nabla_x \times \mathbf{S}_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2 + o(n^2). \end{aligned} \tag{2.31}$$

The class (2.30) is the counterpart of (1.12) and leads to the c^2-optimal three-point lower bound $1/k^{(3)}$ on the effective conductivity κ^*, see [12]. The solution of the Euler-Lagrange equation $\delta W_B^{(1)}[\mathbf{S}_1(\cdot)] = 0$ now has the form

$$\mathbf{S}_1(\mathbf{x}) = \mathbf{S}^{(1)}(\mathbf{x}) + O(c),$$

where $\mathbf{q}^{(1)}(\mathbf{x}) = \nabla \times \mathbf{S}^{(1)}(\mathbf{x})$ is the disturbance to the constant heat flux $\mathbf{Q}$ in an unbounded matrix, introduced by a single spherical inhomogeneity, located at the origin:

$$\mathbf{q}^{(1)}(\mathbf{x}) = 3\beta\mathbf{Q}\cdot\left[\nabla\nabla\varphi(\mathbf{x}) + h(\mathbf{x})\mathbf{I}\right], \quad \text{i.e.} \quad \mathbf{S}^{(1)}(\mathbf{x}) = -3\beta\mathbf{Q}\times\nabla\varphi(\mathbf{x}). \tag{2.32}$$

Then

$$\begin{aligned} k^* \le k^{(3)}, \quad k^{(3)}Q^2 &= \min W_B^{(1)}[\mathbf{S}_1(\cdot)] = W_B^{(1)}[\mathbf{S}^{(1)}(\cdot)] + o(c^2) \\ &= k_m\left\{1 - 3\beta c + 3\beta^2\left(2 + \frac{[k]}{k_m}m_2\right)c^2\right\}Q^2 + o(c^2), \end{aligned} \tag{2.33}$$

where m_2 is the statistical parameter (2.6). In virtue of these relations, the optimal three-point lower bounds for the c^2-coefficient $a_{2\kappa}$ in the virial expansion (1.10) of κ^* are straightforwardly obtained (see [12, 15]):

$$a_{2\kappa}^l \le a_{2\kappa}, \quad a_{2\kappa}^l = 3\beta^2\left(1 + \frac{[\kappa]}{\kappa_f}m_2\right). \tag{2.34}$$

Eqs. (2.29) and (2.31) are the counterparts of Eqs. (2.10) and (2.11) respectively. A fully similar analysis shows in turn that the leading part $\mathbf{S}_{2,0}$ of the optimal kernel $\mathbf{S}_2$, $\mathbf{S}_2(\mathbf{x},\mathbf{y}) = \mathbf{S}_{2,0}(\mathbf{x},\mathbf{y}) + O(c)$, has now the form

$$2\mathbf{S}_{2,0}(\mathbf{x},\mathbf{x}-\mathbf{z}) = \mathbf{S}^{(2)}(\mathbf{x};\mathbf{z}) - \mathbf{S}^{(1)}(\mathbf{x}) - \mathbf{S}^{(1)}(\mathbf{x}-\mathbf{z}), \tag{2.35}$$

where $\mathbf{q}^{(2)}(\mathbf{x};\mathbf{z}) = \nabla_x \times \mathbf{S}^{(2)}(\mathbf{x};\mathbf{z})$ is the disturbance to the constant heat flux $\mathbf{Q}$ in an unbounded matrix of conductivity κ_m, generated by a pair of spherical inhomogeneities of conductivity κ_f, centered at the origin and at the point $\mathbf{z}$.

In order to improve on the optimal lower bound (2.34), similarly to Eqs. (1.8), (2.1)–(2.3) for the upper one and (2.32), we can make the following choice of the kernel $\mathbf{S}_2$ in (2.28):

$$\mathbf{S}_2(\mathbf{x},\mathbf{x}-\mathbf{z}) = \frac{1}{2}\boldsymbol{\Phi}(\mathbf{z})\times\nabla f\left(\mathbf{x}-\frac{\mathbf{z}}{2},\mathbf{z}\right), \tag{2.36a}$$

where the functions $\boldsymbol{\Phi}$ and f can be again treated as adjustable. Let us note that now the field

$$\begin{aligned} \widetilde{\mathbf{q}}_2(\mathbf{x},\mathbf{x}-\mathbf{z}) &= 2\nabla_x \times \mathbf{S}_2(\mathbf{x},\mathbf{x}-\mathbf{z}) \\ &= -\boldsymbol{\Phi}(\mathbf{z})\cdot\left[\nabla\nabla f\left(\mathbf{x}-\frac{\mathbf{z}}{2},\mathbf{z}\right) - \Delta f\left(\mathbf{x}-\frac{\mathbf{z}}{2},\mathbf{z}\right)\mathbf{I}\right], \end{aligned} \tag{2.36b}$$

in general, is *not* the disturbance to a certain heat flux in an unbounded matrix, introduced by a single radial inhomogeneous sphere, centered at the point $\mathbf{z}/2$. A simple check shows, however, that for a homogeneous middle sphere the field $\widetilde{\mathbf{q}}_2(\mathbf{x},\mathbf{x}-\mathbf{z})$ is indeed such a disturbance, see Eqs. (2.32) and (2.36a). An example of this kind will be considered in Section 3.

The further analysis is fully similar to the one, already performed in Section 2.2. That is way we shall present the basic results only. The explicit form of the functional

$$\widehat{W}_B^{(2)\dagger}\left[\boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)\right] = \frac{1}{n^2}\widetilde{W}_B^{(2)}\left[\mathbf{S}^{(1)}(\cdot), \frac{1}{2}\boldsymbol{\Phi}(\cdot)\cdot\nabla f(\cdot,\cdot)\right] \tag{2.37}$$

is obtained straightforwardly by means of Eqs. (2.29), (2.31), (2.32) and (2.36); it is of the same form (2.15), provided we replace κ by k, $\mathbf{G}$ by $\mathbf{Q}$, $\nabla\nabla\varphi$ by $\nabla\nabla\varphi + h\mathbf{I}$ and $\nabla\nabla f$ by $-(\nabla\nabla f - \Delta f\mathbf{I})$.

With the notations

$$\begin{aligned}\mathfrak{R}(\mathbf{z}) = \frac{1}{V_a}\int\left[1 - 2\frac{[\kappa]}{\kappa_f}h\left(\mathbf{w} - \frac{\mathbf{z}}{2}\right)\right]\left[\nabla\nabla f(\mathbf{w},\mathbf{z}) - \Delta f(\mathbf{w},\mathbf{z})\mathbf{I}\right]\\ \cdot\left[\nabla\nabla f(\mathbf{w},\mathbf{z}) - \Delta f(\mathbf{w},\mathbf{z})\mathbf{I}\right] d^3\mathbf{w},\end{aligned} \tag{2.38a}$$

$$\begin{aligned}\Im(\mathbf{z}) = \frac{1}{V_a}\int h\left(\mathbf{w} - \frac{\mathbf{z}}{2}\right)\left[\nabla\nabla\varphi\left(\mathbf{w} + \frac{\mathbf{z}}{2}, \mathbf{z}\right) + h\left(\mathbf{w} + \frac{\mathbf{z}}{2}\right)\mathbf{I}\right]\\ \cdot\left[\nabla\nabla f(\mathbf{w},\mathbf{z}) - \Delta f(\mathbf{w},\mathbf{z})\mathbf{I}\right] d^3\mathbf{w},\end{aligned} \tag{2.38b}$$

the Euler-Lagrange equation $\delta_{\Phi}\widehat{W}_B^{(2)\dagger} = 0$ reads

$$\boldsymbol{\Phi}(\mathbf{z})\cdot\mathfrak{R}(\mathbf{z}) = 6\beta\frac{[k]}{k_m}\mathbf{Q}\cdot\Im(\mathbf{z}),$$

whose solution is

$$\boldsymbol{\Phi}(\mathbf{z}) = 6\beta\frac{[k]}{\kappa_m}\mathbf{Q}\cdot\Im(\mathbf{z})\cdot\mathfrak{R}^{-1}(\mathbf{z}). \tag{2.39}$$

Then the minimum value of the functional $\widehat{W}_B^{(2)\dagger}$ is

$$\min_{\Phi}\widehat{W}_B^{(2)\dagger}\left[\boldsymbol{\Phi}(\cdot), f(\cdot,\cdot)\right] = -18\beta^2\frac{[k]^2}{k_m}V_a\mathbf{Q}\cdot\int g_0(\mathbf{z})\Im(\mathbf{z})\cdot\mathfrak{R}^{-1}(\mathbf{z})\cdot\Im(\mathbf{z})\,d^3\mathbf{z}\cdot\mathbf{Q}\,. \tag{2.40}$$

According to Eqs. (2.29), (2.33), (2.37) and (2.40), an upper bound $k^\dagger$ on the effective compliance k^* immediately follows

$$k^*Q^2 \le k^\dagger Q^2 = k^{(3)} + \frac{1}{V_a^2}\min\widehat{W}_B^{(2)\dagger}c^2 + o(c^2)$$

$$= k_m\left\{1 - 3\beta c + 3\beta^2\left(2 + \frac{[k]}{k_m}m_2 - \frac{[k]}{k_m}^2\widetilde{m}_2^l\right)c^2\right\}Q^2 + o(c^2).$$

Here

$$\widetilde{m}_2^l = \widetilde{m}_2^l[g_0(\cdot), f(\cdot,\cdot), \alpha] = \frac{2}{V_a}\int g_0(\mathbf{z})\,\mathrm{tr}\left[\Im(\mathbf{z})\cdot\mathfrak{R}^{-1}(\mathbf{z})\cdot\Im(\mathbf{z})\right]\,d^3\mathbf{z} \tag{2.41}$$

is the counterpart of the statistical parameter $\widetilde{m}_2^u$, see (2.25). In virtue of these relations we obtain straightforwardly the following lower bound for the c^2-coefficient $a_{2\kappa}$ in the virial expansion (1.10) of κ^*:

$$a_{2\kappa}^{l\dagger} \leq a_{2\kappa}, \quad a_{2\kappa}^{l\dagger} = 3\beta^2 \left(1 + \frac{[\kappa]}{\kappa_f} m_2 + \left(\frac{[\kappa]}{\kappa_f}\right)^2 \widetilde{m}_2^l\right). \tag{2.42}$$

Let us note that the bounds (2.24) and (2.42) are *five-point* bounds in the sense that they require knowledge of the first ℓ-point moments for the random density field $\psi(\mathbf{x})$ up to $\ell = 5$, see Eqs. (1.2)–(1.5), (2.26)–(2.28). To get explicitly the parameters $\widetilde{m}_2^l$ and $\widetilde{m}_2^u$ for a given function f, an analytical evaluation of the integrals (2.20) and (2.38) is needed however.

3. A SIMPLE EXAMPLE

Let us choose now the function $f(\mathbf{w}, \mathbf{z})$ in Eqs. (2.2) and (2.36) in the form

$$f(\mathbf{w}, \mathbf{z}) = \varphi\left(\mathbf{w}, \frac{|\mathbf{z}|}{2} + A\right)$$

at $|\mathbf{z}| \geq 2a$, i.e.

$$f\left(\mathbf{x} - \frac{\mathbf{z}}{2}, \mathbf{z}\right) = \varphi\left(\mathbf{x} - \frac{\mathbf{z}}{2}, \frac{|\mathbf{z}|}{2} + A\right), \tag{3.1}$$

where A is a scalar parameter, $A \geq a$, so that $|\mathbf{z}|/22 + A \geq 2a$.

According to the foregoing analysis, this choice means that the disturbance $T^{(2)}(\mathbf{x}; \mathbf{z})$, generated by two spheres centered at the origin and at the point $\mathbf{z}$, is approximated by the superposition of the disturbances $T^{(1)}(\mathbf{x})$ and $T^{(1)}(\mathbf{x} - \mathbf{z})$, generated by the same two spheres, but considered as singly, and the disturbance $T^{(1)}(\mathbf{x} - \mathbf{z}/2)$, generated by a single homogeneous sphere, centered exactly between them and encompassing the same spheres, see Eqs. (1.8), (1.14), (2.1) and (3.1). At that, let us recall, the middle sphere is immersed into adjustable temperature field $\boldsymbol{\Phi}(\mathbf{z})$ that has been varied in order to derive the best c^2-bounds on the effective conductivity. Now we shall obtain this bounds as functions of the parameter $s = A/a$, $s \geq 1$.

After simple change of the integrand variable the fields $\mathcal{R}(\mathbf{z})$ and $\mathcal{J}(\mathbf{z})$ in (2.20) are recast as

$$\begin{aligned} \mathcal{R}(\mathbf{z}) &= \frac{1}{V_a}\int \left[1 + 2\frac{[\kappa]}{\kappa_m} h(\mathbf{u})\right] \nabla\nabla f\left(\frac{\mathbf{z}}{2} - \mathbf{u}, \mathbf{z}\right) \cdot \nabla\nabla f\left(\frac{\mathbf{z}}{2} - \mathbf{u}, \mathbf{z}\right) d^3\mathbf{u}, \\ \mathcal{J}(\mathbf{z}) &= \frac{1}{V_a}\int h(\mathbf{u}) \nabla\nabla\varphi(\mathbf{z} - \mathbf{u}) \cdot \nabla\nabla f\left(\frac{\mathbf{z}}{2} - \mathbf{u}, \mathbf{z}\right) d^3\mathbf{u}. \end{aligned} \tag{3.2}$$

Taking into account that $\nabla\nabla\varphi(\mathbf{u},\mathbf{z}) = -\frac{1}{3}\mathbf{I}$ at $|\mathbf{u}| < |\mathbf{z}|$ and the Eqs. (3.1) and (3.2), we get

$$\begin{gathered} \mathcal{R}(\mathbf{z}) = \frac{1}{9}\left\{3\left(\frac{|\mathbf{z}|}{2a} + s\right)^3 + 2\,\frac{[\kappa]}{\kappa_m}\right\}\mathbf{I}\,, \\ \mathcal{J}(\mathbf{z}) = -\,\frac{1}{3}\boldsymbol{\omega}(\mathbf{z}), \quad \boldsymbol{\omega}(\mathbf{z}) = \frac{1}{V_a}\int h(u)\nabla\nabla\varphi(\mathbf{z}-\mathbf{u})\,d^3\mathbf{u}\,. \end{gathered} \tag{3.3}$$

The field $\boldsymbol{\omega}(\mathbf{z})$ is the same one that appears in the variational procedure of Willis [17], see [9, 10] also, whose explicit form is

$$\boldsymbol{\omega}(\mathbf{z}) = \frac{1}{3}\left(\frac{a}{|\mathbf{z}|}\right)^3 \left(3\mathbf{e}_r\mathbf{e}_r - \mathbf{I}\right), \quad \mathbf{e}_r = \mathbf{z}/|\mathbf{z}|. \tag{3.4}$$

In the same way one obtains for the fields $\mathfrak{R}(\mathbf{z})$ and $\mathfrak{S}(\mathbf{z})$ in (2.38) the following formulae:

$$\mathfrak{R}(\mathbf{z}) = \frac{2}{9}\left\{3\left(\frac{|\mathbf{z}|}{2a} + s\right)^3 - 4\,\frac{[\kappa]}{\kappa_f}\right\}\mathbf{I}, \quad \mathfrak{S}(\mathbf{z}) = \frac{2}{3}\boldsymbol{\omega}(\mathbf{z})\,. \tag{3.5}$$

After simple algebra, based on Eqs. (2.25), (2.41), (3.3)–(3.5), we get eventually the needed parameters $\widetilde{m}_2^u$ and $\widetilde{m}_2^l$:

$$\begin{gathered} \widetilde{m}_2^u = 32\int_0^{1/2} g_0\left(\frac{a}{\rho}\right)\frac{\rho^5}{3(1+2s\rho)^3 + 16\rho^3[\kappa]/\kappa_m}\,d\rho\,, \\ \widetilde{m}_2^l = 64\int_0^{1/2} g_0\left(\frac{a}{\rho}\right)\frac{\rho^5}{3(1+2s\rho)^3 - 32\rho^3[\kappa]/k_f}\,d\rho\,. \end{gathered} \tag{3.6}$$

Thus, for the simple choice (3.1) of the function $f(\mathbf{w},\mathbf{z})$, we have obtained the c^2-bounds (2.24), (2.42) explicitly. A simple check shows that the integrands in (3.6) are always positive, and so are the parameters $\widetilde{m}_2^l$ and $\widetilde{m}_2^u$. Then, from (2.8), (2.24), (2.34) and (2.42), we can conclude that the obtained bounds always improve on the optimal three-point bounds. Moreover, it is immediately seen that the parameters $\widetilde{m}_2^l$ and $\widetilde{m}_2^u$ are decreasing functions of the parameter s, vanishing as $s \to \infty$. Therefore the obtained bounds are the best if $s = 1$, i.e. when the middle sphere, encompassing the other two ones, touches them. This fact suggests that the consideration of the case when the middle sphere overlaps the other two spheres could lead to better results. The calculations in this case, however, are more complicated. In the limiting case $s \to \infty$ our bounds coincide with the optimal three-point bounds.

The behaviour of our bounds is illustrated in the well-stirred case when $g(\mathbf{z}) = 1$ at $|\mathbf{z}| > 2a$, see Table 1. It is seen that the new lower bound (at $s = 1$) improves

TABLE 1. Comparison of various bounds on the c^2-coefficient a_2 for a well-stirred dispersion of spheres; the exact values are due to Felderhof *et al.* [11] and the value of the parameter m_2 is $m_2 \approx 0.14045$ [12]

β	α	Lower bounds			exact	Upper bounds		
		3-point (2.34)	present (2.42)	Willis [9, (7.21)]		present (2.24)	3-point (2.8)	Willis [9, (7.21)]
-0.5	0	$-\infty$	$-\infty$	–	0.588	0.641	0.645	0.659
-0.49	0.013	-6.715	-2.934	–		0.617	0.620	0.634
-0.4	0.143	0.076	0.165	–	0.399	0.421	0.422	0.433
-0.3	0.308	0.185	0.194	–	0.236	0.243	0.244	0.250
-0.2	0.500	0.103	0.104	–	0.110	0.111	0.112	0.114
-0.1	0.727	0.028	0.028	–	0.029	0.029	0.029	0.029
0	1	0	0	0	0	0	0	0
0.2	1.75	0.127	0.127	0.126	0.130	0.132	0.133	–
0.4	3.00	0.525	0.527	0.529	0.563	0.607	0.615	–
0.6	5.50	1.204	1.211	1.249	1.370	1.686	1.763	–
0.8	13	2.169	2.185	2.328	2.638	4.437	5.156	–
0.9	28	2.759	2.782	3.016	3.485	8.576	11.645	–
0.99	298	3.352	3.382	3.726		58.705	125.592	–
1.0	∞	3.420	3.450	3.811	4.506	∞	∞	–

considerably on the respective three-point bound when $\alpha \to 0$; a similar improvement takes place for the upper ones at $\alpha \to \infty$. In Table 1 the bound of Willis [9] is also given. Recall that it improves on the lower three-point bound, but the upper one is worse.

Finally, we shall note that the proposed approach to derive variational bounds can be employed on the base of the variational principle of Hashin-Shtrikman. In this case it can be easily shown, for example, that the bounds of Willis correspond to the approximation $\nabla T^{(2)}(\mathbf{x};\mathbf{z}) = \nabla T^{(1)}(\mathbf{x}) + \boldsymbol{\Phi}(\mathbf{z})$, see [9, 10]. This means that the bounds of Willis can be treated as the exact HS-counterpart of our bounds, derived in Section 3. More details will be given elsewhere.

ACKNOWLEDGEMENTS. The support of this work by the Bulgarian Ministry of Education, Science and Technology under Grant No MM416-94 is gratefully acknowledged. The author thanks K. Z. Markov for the helpful and stimulating discussions.

REFERENCES

1. Beran, M. Statistical continuum theories. John Wiley, New York, 1968.
2. Stratonovich, R. L. Topics in theory of random noises, Vol. 1, Gordon and Breach, New York, 1967.
3. Markov, K. Z. On the factorial functional series and their application to random media. *SIAM J. Appl. Math.*, **51**, 1991, 172–186.

4. Markov, K. Z. On the heat propagation problem for random dispersions of spheres. *Math. Balkanica (New Series)*, **3**, 1989, 399–417.

5. Christov, C. I., K. Z. Markov. Stochastic functional expansion for random media with perfectly disordered constitution. *SIAM J. Appl. Math.*, **45**, 1985, 289–311.

6. Markov, K. Z., C. I. Christov. On the problem of heat conduction for random dispersions of spheres allowed to overlap. *Math. Models and Methods in Applied Sciences*, **2**, 1992, 249–269.

7. Jeffrey, D. J. Conduction through a random suspension of spheres. *Proc. Roy. Soc. London*, **A335**, 1973, 355–367.

8. Zvyatkov, K. D. Variational principles and the c^2-formula for the effective conductivity of a random dispersion. In: *Continuum Models and Discrete Systems*, ed. K. Z. Markov, World Sci., 1996, 324–331.

9. Markov, K. Z., K. D. Zvyatkov. Functional series and Hashin-Shtrikman's type bounds on the effective conductivity of random media. *Europ. J. Appl. Math.*, **6**, 1995, 611–629.

10. Markov, K. Z., K. D. Zvyatkov. Functional series and Hashin-Shtrikman's type bounds on the effective properties of random media. In: *Advances in Mathematical Modeling of Composite Materials*, ed. K. Z. Markov, World Sci., 1994, 59-106.

11. Felderhof, B. U., G. W. Ford, E. G. D. Cohen. Two-particle cluster integral in the expansion of the dielectric constant. *J. Stat. Phys.*, **28**, 1982, 1649–1672.

12. Markov, K. Z. Application of Volterra-Wiener series for bounding the overall conductivity of heterogeneous media. I. General procedure. II. Suspensions of equi-sized spheres. *SIAM J. Appl. Math.*, **47**, 1987, 831–850, 851–870.

13. Hashin, Z. Assessment of the self-consistent approximation. *J. Composite Materials*, **2**, 1968, 284–300.

14. Acrivos, A., E. Chang. A model for estimating transport quantities in two-phase materials. *Phys. Fluids*, **29**, 1986, 3–4.

15. Markov, K. Z., K. D. Zvyatkov. Optimal third-order bounds on the effective properties of some composite media, and related problems. *Advances in Mechanics (Warsaw)*, **14**, No 4, 1991, 3–46.

16. Beran, M. Use of a variational approach to determine bounds for the effective permittivity of a random medium. *Nuovo Cimento*, **38**, 1965, 771–782.

17. Willis, J. R. Variational principles and bounds for the overall properties of composites. In: *Continuum Models and Discrete Systems*, ed. J. Provan, University of Waterloo Press, Ontario, 1978, 185–215.

Received February 19, 1998

Faculty of Mathematics and Informatics,
"K. Preslavski" University of Shumen,
BG-9700 Shumen, Bulgaria
E-mail: zvjatkov@uni-shoumen.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 91, 1997

KNOWLEDGE REPRESENTATION AND PROBLEM SOLVING IN THE INTELLIGENT COMPUTER ALGEBRA SYSTEM STRAMS

MARIA M. NISHEVA-PAVLOVA

The paper discusses the intelligent computer algebra system STRAMS being under development at the Faculty of Mathematics and Informatics, Sofia University. The functional facilities and the architecture of STRAMS are briefly described. The presentation focuses on issues related to the suggested knowledge representation formalism, the structure and the contents of the knowledge base of STRAMS and the implemented mathematical problem solving and learning mechanisms.

Keywords: mathematical knowledge representation, problem solving, intelligent computer algebra system

1995 Math. Subject Classification: main 68T30, secondary 68T35

1. INTRODUCTION

In the last three decades Computer Algebra Systems (CAS) have been widely used in the automation of scientific computation and design. These systems can help in the solution of various problems connected with the execution of complicated and labour-consuming transformations of mathematical expressions. However, irrespective of their good capabilities, "classical" CAS like Reduce, Maple, Mathematica etc. are sometimes difficult for use. The most serious problem here [2, 3] is that "classical" CAS behave as black boxes and therefore the interpretation of the suggested solutions can call for significant efforts.

The reason for this problem is that "classical" CAS have no mathematical knowledge represented in an explicit, declarative way. Their knowledge is embedded implicitly in the algorithms and is inaccessible to the user.

Therefore a series of successful attempts have been made in order to build various kinds of the so-called intelligent CAS. In general, intelligent CAS are systems that are capable to manipulate different types of mathematical knowledge and use a large set of Artificial Intelligence methods and techniques. Because of adequacy and efficiency considerations, intelligent CAS usually are hybrid [1, 4] by means of combining several formalisms and paradigms.

A set of projects aimed at the investigation of different aspects of the integration of the classical approaches for developing CAS with Artificial Intelligence methods and tools have been under development at the Faculty of Mathematics and Informatics, Sofia University. An approach to building intelligent CAS has been developed [5]. The first version of a knowledge-based tool for developing CAS called KAM [6] has been implemented. The experimental intelligent computer algebra system STRAMS discussed in this paper has been under development using KAM. In general, STRAMS is a knowledge-based CAS that can solve various types of mathematical problems, learn and explain the results of its work. The approach used for the development of STRAMS and the major features of this system are described and argumented in details in [5, 6]. Therefore we emphasize here on the analysis of the architecture of STRAMS, the contents of its knowledge base and the implemented mathematical problem solving mechanisms.

2. FUNCTIONAL FACILITIES AND ARCHITECTURE OF STRAMS

STRAMS is a general-purpose intelligent CAS. The definition domain D of the expressions that can be manipulated in STRAMS includes all expressions containing numbers, symbols and the functions: $+$, $-$, $*$, $/$, power function, exponential, logarithmic and trigonometric functions. STRAMS is intended for solving the following main problem types:

— simplification (reduction to a canonical form) of expressions from D;

— symbolic equation solving (solving equations of the form $expr_1 = expr_2$, where $expr_1$ and $expr_2$ are expressions from D);

— symbolic differentiation of expressions from D;

— symbolic integration (formal integration of functions belonging to a particular subset of D).

The architecture of STRAMS is determined by its functional facilities and some additional design requirements like transparency and learning and explanation generation capability. The architecture of the environment KAM used as a tool for the implementation of STRAMS exerts a considerable influence as well.

STRAMS includes the following functional components: a mathematical problem solving engine, an explanation module, an interface module, a control block.

The architecture of STRAMS is shown in Fig. 1.

The mathematical problem solving engine consists of two modules: a knowledge engine and a learning module realizing respectively the problem solving and the

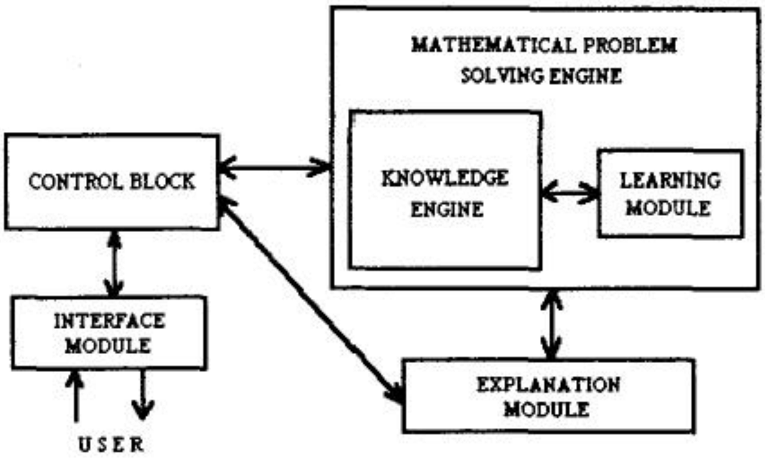

Fig. 1. Architecture of STRAMS

learning capabilities of STRAMS. The structure and the functioning mechanisms of the mathematical problem solving engine are discussed in Section 4 of this paper.

The explanation module realizes the explanation generation capabilities of STRAMS. These capabilities are described in [5, 6] and will not be discussed in this paper.

The interface module is the component of STRAMS the users are in touch with. It analyzes the user requests, converts them into the corresponding internal form and sends appropriate messages to the control block. The current version of the interface module provides only some relatively primitive communication means that will be improved in the further versions.

The control block realizes the general control of the system's work and the interaction between the other components.

3. KNOWLEDGE REPRESENTATION IN STRAMS

The formalism supported by the environment KAM is used for the knowledge representation in STRAMS. It is oriented to the description of knowledge about the properties of the manipulated functions and the methods for mathematical problem solving defined by these properties. This formalism is described and analyzed in details in [5, 6]. Here we present it in brief and give appropriate examples.

The knowledge of STRAMS about the properties of the manipulated functions is described using a special type of rules called rewrite rules. The structure of each rewrite rule includes the description of a correct transformation of a class of mathematical expressions and the formulation of some general preconditions for its performance (if there are any). Examples of rewrite rules:

$(x+y)(x-y) = x^2 - y^2$

$e^x e^y = e^{x+y}$

$\operatorname{tg}(a+b) = \dfrac{\operatorname{tg} a + \operatorname{tg} b}{1 - \operatorname{tg} a \operatorname{tg} b}$ with precondition a, b, $a+b$ different from $\dfrac{(2k+1)\pi}{2}$.

The description of the methods for transformation of the expressions and equations STRAMS can manipulate is realized by the so-called generalized rules (methods). Each generalized rule describes a sequence of transformations of the given expression aimed at its conversion into a particular form. In this sense usually generalized rules contain sequences of properly grouped rewrite rules. More precisely, each generalized rule consists of two parts — a precondition and a body. The precondition is a predicate whose satisfaction is a necessary condition for the application of the generalized rule and for achieving its purpose. The evaluation of the precondition of a given generalized rule is the first step of its application. If the precondition is true, then the body of the generalized rule is performed. The body of a generalized rule may contain:

— a sequence of rewrite rules. Each of them can include some additional control information. In this case the generalized rule is called *declarative*;

— the code of a procedure realizing the application of the rule. Such generalized rules are called *procedural*;

— a set of pairs (*pattern*, *procedure*) such that when the examined expression matches one of the patterns, the corresponding procedure is executed. These generalized rules are called *hybrid*.

Declarative generalized rules are most numerous in STRAMS. As it was mentioned above, the body of such a rule consists of a sequence of rewrite rules that can be divided in three groups: pre-rules, basic rules, post-rules.

The pre-rules are intended to prepare the given expression for the performance of the basic rules. The post-rules are used to remove some "defects" remaining after the performance of the basic rules.

There are three basic types of declarative generalized rules according to the mode of application of their bodies: normal, cyclic and recursive. The body of a normal generalized rule is performed in the following way. First the pre-rules are consecutively applied to the given expression. Each of them is executed on the result returned by the previous one. Then the basic rules are applied in the same way on the result of the execution of the pre-rules. At last the post-rules are applied in the described way.

The body of a cyclic generalized rule contains only one basic rule. It is performed in the following way. First, the pre-rules are executed as in the case of a normal rule. Then the basic rule is executed. If it has not changed its argument, the execution of the body of the generalized rule stops and the current result is returned. In the other case, the corresponding post-rules are performed and then a cyclic execution of the described sequence of steps is carried out until the basic rule returns its argument unchanged.

The body of a recursive generalized rule is first executed on the subexpressions of the given expression and then it is applied to the obtained new argument.

It is possible to construct some combinations between the basic types of declarative generalized rules. In this sense very attractive are the so-called cyclic recursive generalized rules that can be used as a proper mean for the description of some methods for expression simplification (reduction to a canonical form). As an example of such a method we can examine the transformation called expansion.

This transformation can be defined by the equality

$$\begin{aligned}(x_1 + x_2 + \cdots + x_m)(y_1 + y_2 + \cdots + y_n) = {} & x_1y_1 + x_1y_2 + \cdots + x_1y_n \\ & + x_2y_1 + x_2y_2 + \cdots + x_2y_n \\ & \quad \cdots\cdots\cdots\cdots \\ & + x_my_1 + x_my_2 + \cdots + x_my_n.\end{aligned}$$

It is described in STRAMS by a cyclic recursive generalized rule with a body containing the following basic rule:

$$\prod_{i=1}^{n} A_i(x_1 + x_2 + \cdots + x_m) \prod_{j=1}^{k} B_j = \prod_{i=1}^{n} A_i x_1 \prod_{j=1}^{k} B_j + \prod_{i=1}^{n} A_i(x_2 + \cdots + x_m) \prod_{j=1}^{k} B_j.$$

Another classification criterion of the generalized rules is the role they play in the problem solving process of a given, relatively complex task (such tasks in STRAMS are equation solving and symbolic integration). In this sense they can be classified as key and non-key ones. The key generalized rules play a significant role in the control of the search in the state graph of the corresponding problem. In the role of examples of key and non-key generalized rules we give here the descriptions of two generalized rules included in the knowledge base of the equation solving subsystem of STRAMS.

Example 1. *Isolation.* Let an equation $eq : expr_1 = expr_2$ be given and let f be the outermost function in $expr_1$. The execution of the body of the generalized rule consists in the application of the inverse of f to $expr_1$ and $expr_2$. The precondition of the generalized rule is: the unknown occurs in only one of the arguments of f and $expr_2$ does not contain the unknown. The goal is to remain in the left-hand side of eq only the argument containing the unknown.

This generalized rule is a key one and is implemented procedurally due to effectiveness considerations.

Example 2. *Collection.* The goal of this generalized rule is to reduce the number of occurrences of the unknown. Collection is a non-key generalized rule with no explicit precondition. STRAMS applies it only if none of the key generalized rules can be applied. So the precondition of Collection (and of all non-key generalized rules) is: there is no key generalized rule with satisfied preconditions.

This generalized rule is declarative, normal. One of its rewrite rules is:

$AB + AC = A(B + C)$ with the precondition A must contain the unknown.

The knowledge of STRAMS about the problem solving methods for the included types of tasks is described either directly by proper generalized rules or using specific constructions called schemata. A schema is a sequence of non-key generalized rules. It describes a definite step in the problem solving process of a relatively complex task (equation solving or symbolic integration). For a more precise definition of the concept of a schema one can use the following additional considerations:

- each schema is a sequence of at least two non-key generalized rules;

- each schema begins either with the first generalized rule used in the problem solving process or with a generalized rule applied after the application of a key generalized rule;
- after the application of a schema either the corresponding problem is found to be solved or a key generalized rule can be applied.

Thus schemata are a natural generalization of generalized rules. The precondition of a schema is the applicability of its first generalized rule. The goal is to solve the problem or to be able to apply a key generalized rule after the application of the schema.

4. STRUCTURE AND FUNCTIONING MECHANISMS OF THE MATHEMATICAL PROBLEM SOLVING ENGINE

As it was mentioned in Section 2, the mathematical problem solving engine of STRAMS consists of two modules: a knowledge engine and a learning module. The knowledge engine includes the so-called inference control block and the following processing subsystems:

— a simplification subsystem;
— an equation solving subsystem;
— a symbolic differentiation subsystem;
— a symbolic integration subsystem.

The structure of the knowledge engine is shown in Fig. 2.

The processing subsystems realize the main functional facilities of STRAMS listed in Section 2. Each of these subsystems is a relatively autonomous knowledge-based system with its own knowledge base and problem solving program. The typical structure of the processing subsystems of STRAMS is presented in Fig. 3.

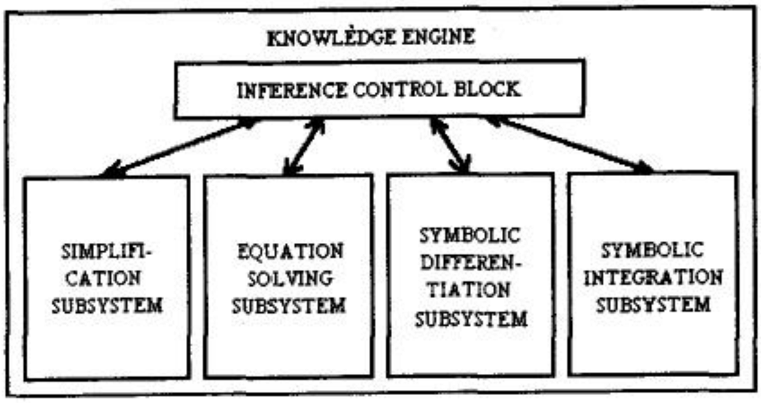
KNOWLEDGE ENGINE
INFERENCE CONTROL BLOCK
SIMPLIFI-CATION SUBSYSTEM
EQUATION SOLVING SUBSYSTEM
SYMBOLIC DIFFEREN-TIATION SUBSYSTEM
SYMBOLIC INTEGRATION SUBSYSTEM

Fig. 2. Structure of the knowledge engine

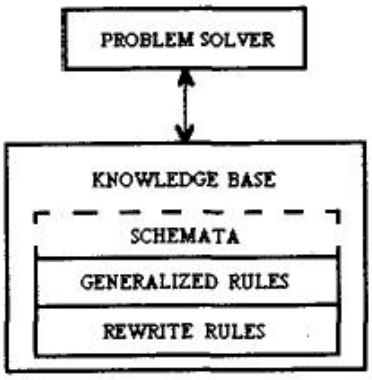

Fig. 3. Structure of the processing subsystems of STRAMS

The knowledge base of each processing subsystem includes the set of generalized rules and rewrite rules that have been used in solving the corresponding type of problems. Additionally, the knowledge bases of the equation solving subsystem and the symbolic integration subsystem contain the corresponding schemata accumulated by the learning module of STRAMS during the system's work.

The problem solver of each processing subsystem realizes the search in the state space of the current problem of the corresponding type. This problem can either be formulated by the user or be generated by some of the processing subsystems. In the role of operators in the state space search the problem solvers use the schemata and generalized rules available at the moment. The application of the chosen generalized rules is performed by the generalized rule interpreter supported by the environment KAM. Some additional search control knowledge has also been used by the problem solvers. It is formulated as a result of some experiments carried out with the particular processing subsystems.

The inference control block realizes the interaction between the knowledge engine and the learning module of STRAMS. The second main function of the inference control block is to manage the interaction between the particular processing subsystems (for example, all processing subsystems generate canonization problems that are solved by the simplification subsystem).

In terms of the functioning mechanisms of the mathematical problem solving engine, the method of work of the problem solvers is most interesting. The particular problem solvers are adjusted copies or simplified versions of one and the same prototype (the control block of the inference engine of KAM [6]). Therefore they perform modifications of one and the same algorithm. The differences are in the kind of the used operators (only generalized rules or schemata and generalized rules) and in the form of the used search control knowledge.

Let us consider as an example the method of work of the problem solver of the equation solving subsystem. There are at least two reasons causing our special interest to this subsystem:

- its problem domain is appropriate for the application of the schemata formalism. Therefore it can do a kind of learning based on the capability for discovering and memorizing the schemata used in the problem solving process;
- it is well known [7] that the state space of some of the types of equations admissible in STRAMS is enormous (includes of the order of 10^{10} states). Therefore the use of some strategic knowledge in order to avoid the exhaustive search is necessary from the point of view of the practical applicability of STRAMS.

The discussed problem solver uses for search control purposes a special evaluation function $Complexity(eq,var)$. $Complexity(eq,var)$ is a linear combination of the number of occurrences $VarOccur(eq,var)$ of the symbol var in the equation eq and the sum of the nesting depths $CommonVarDepth(eq,var)$ of var in eq:

$$Complexity(eq, var) = c_1 VarOccur(eq, var) + c_2 CommonVarDepth(eq, var).$$

This function is used in the examination of all equations. Initially, the value of $Complexity(eq,x)$, where eq is the given equation and x is the unknown, is computed and the variable $initial\_complexity\_factor$ gets this value:

$$initial\_complexity\_factor = Complexity(eq,x).$$

Then the problem solver does not explore all equations eq' in the state graph of eq that do not satisfy the so-called simplicity criterion:

$$f(steps\_done).\ Complexity(eq',x)\text{-}initial\_complexity\_factor$$
$$< c_3.initial\_complexity\_factor,$$

where *steps_done* is the number of transformations reducing *eq* to *eq'*. In this way the simplicity criterion plays the role of a heuristics for pruning a part of the state graph of the current equation in order to avoid the exhaustive search.

Another heuristics used for search control purpose states that the application of a key generalized rule as an operator can significantly shorten the path to the solution. Therefore, when a key generalized rule is applied at a given step, the discussed problem solver continues its work with the exploration of the equation obtained as a result of the application of this generalized rule. If no key generalized rule can be applied at the current step, the problem solver looks for a proper schema leading to the applicability of a key generalized rule.

The general form of the function *Complexity*(*eq,var*) and the simplicity criterion, the definition of the function $f(n)$ and the concrete values of the parameters c_1, c_2, c_3 are suggested in [9].

Let us assume that an equation *eq* has to be solved with respect to the symbol *var*. During its working cycle the discussed problem solver supports a list of equations belonging to the state graph of *eq* that have to be explored. We shall refer to this list as *eq_list* and to its first element as *current_eq*. Then the algorithm of work of our problem solver can be formulated in general as follows.

S1. Initialize *eq_list* to the list containing only *eq*. Initialize *initial_complexity_factor* to *Complexity*(*eq,var*).

S2. If *eq_list* is the empty list, then report failure and quit.

S3. If the equation *current_eq* is solved, then return *current_eq* and quit.

S4. If *current_eq* does not satisfy the simplicity criterion, then remove *current_eq* from *eq_list* and go to S2.

S5. If a key generalized rule is applicable to *current_eq*, then replace *current_eq* in *eq_list* by the equation obtained as a result of the application of the found generalized rule to *current_eq*. Go to S3.

S6. If an existing schema is applicable to *current_eq*, then modify *eq_list* by analogy with S5 and go to S3.

S7. Remove *current_eq* from *eq_list* and add to the end of *eq_list* the equations that can be produced by the application of all non-key generalized rules to *current_eq*. Go to S2.

Whenever an equation is successfully solved, an attempt for the extraction of new schemata is made. For that purpose the inference control block activates the learning module of STRAMS. The learning module analyzes the used sequence of generalized rules, constructs the new schemata candidates (in accordance with the definition of the schema concept) and merges them with the set of existing schemata. In this way STRAMS does a kind of unsupervised learning by accumulation in the corresponding knowledge base of new, successfully applied schemata that can be used in its further work.

5. IMPLEMENTATION OF STRAMS

The implementation of STRAMS has been realized using the environment KAM. The program modules of the mathematical problem solving engine of STRAMS are either exact copies or simplified versions of some of the program modules of KAM. The knowledge bases of the processing subsystems of STRAMS are built by direct recording of the corresponding rewrite rules and generalized rules in internal form. The knowledge base of the simplification subsystem includes rewrite rules and generalized rules described in [8, 9], and the knowledge base of the equation solving subsystem includes rewrite rules and generalized rules described in [9].

The explanation module of STRAMS is a copy of the module of the same name of KAM. The interface module and the control block of STRAMS are developed especially for the purpose. All program modules of STRAMS are written in Common Lisp.

6. SUMMARY AND CONCLUSION

STRAMS is a knowledge-based CAS with the following main features:

- it can solve various types of problems using a set of methods and techniques, traditionally taught in the secondary school and in the introductory university courses;
- it is able to do a kind of learning and explanation generation;
- it can easily be integrated with other software packages;
- its functional facilities can easily be extended.

These features of STRAMS determine its potential applicability in building expert systems, intelligent tutoring systems etc.

Our current activities are directed to the improvement of the user interface of STRAMS and to the extension of its functional facilities.

ACKNOWLEDGEMENTS. This work is partly funded by the Sofia University SRF under Contract No. 274/1997.

REFERENCES

1. Calmet, J., K. Homann, I. Tjandra. Hybrid Representation for Specification and Communication of Mathematical Knowledge. In: K. Homann, S. Jacob, M. Kerber, H. Stoyan (Eds.), *Proceedings of the Workshop on Representation of Mathematical Knowledge*, 12th European Conference on Artificial Intelligence, Budapest, 1996.
2. Homann, K., J. Calmet. Combining Theorem Proving and Symbolic Mathematical Computing. *LNCS*, **958**, Springer-Verlag, 1995, 18–29.
3. Homann, K., J. Calmet. Structures for Symbolic Mathematical Reasoning and Computation. *LNCS*, **1128**, Springer-Verlag, 1996, 216–227.

4. Kapitonova, Y., A. Letichevsky, M. L'vov, V. Volkov. Tools for Solving Problems in the Scope of Algebraic Programming. *LNCS*, **958**, Springer-Verlag, 1995, 30–47.
5. Nisheva-Pavlova, M. A Knowledge-Based Approach to Building Computer Algebra Systems. In: *Proceedings of JCKBSE*'96, Sozopol, 1996, 222–225.
6. Nisheva-Pavlova, M. KAM — A Knowledge-Based Tool for Developing Computer Algebra Systems. *Ann. Sof. Univ., Fac. Math. and Inf.*, **90**, 1996 (to appear).
7. Silver, B. Precondition Analysis: Learning Control Information. In: R. Michalski, J. Carbonell, T. Mitchell (Eds.), *Machine Learning*, Vol. 2, Morgan-Kaufmann, 1986, 647–670.
8. Todorov, B. An Environment for Building Special-Purpose Knowledge-Based Systems for Computer Algebra. MSc Thesis, Sofia University, 1994 (in Bulgarian).
9. Vatchkov, V. Learning in Equation Solving. MSc Thesis, Sofia University, 1993 (in Bulgarian).

Received March 4, 1998
Revised May 12, 1998

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5 Blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: `marian@fmi.uni-sofia.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 91, 1997

ON THE FIELD FLUCTUATIONS IN A RANDOM DISPERSION

KERANKA S. ILIEVA, KONSTANTIN Z. MARKOV

In this note the variances of the basic random fields — temperature and heat flux — in a dilute dispersion of spheres with a small volume fraction $c \ll 1$, subjected to a constant macroscopic temperature gradient are studied. The basic result is an estimate on the c^2-term of these variances, which includes the well-known c^2-term of the effective conductivity, extensively studied in the literature.

Keywords: random dispersion of spheres, effective conductivity, variance of random fields

1991/1995 Math. Subject Classification: 60G60, 60H15

1. INTRODUCTION

The aim of this note is to report some preliminary results concerning field fluctuation in a dispersion of nonoverlapping spheres. The heat conduction context is chosen, above all, for the sake of simplicity. A similar study of any transport phenomenon through the medium in the linear case can be performed along the same line.

Let us recall first how the problem is stated. Assume we have an unbounded matrix material of conductivity κ_m throughout which filler particles of conductivity κ_f are distributed in a statistically isotropic and homogeneous manner. The random conductivity field $\kappa(\mathbf{x})$ of the medium takes then the values κ_m or κ_f, depending on whether $\mathbf{x}$ lies in the matrix or in a particle, respectively. If $\mathbf{G}$ denotes the prescribed macroscopic temperature gradient imposed upon the medium, the governing equations of the problem, at the absence of body sources, read

$$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0, \quad \mathbf{q}(\mathbf{x}) = \kappa(\mathbf{x})\nabla\theta(\mathbf{x}), \tag{1.1a}$$

under the condition

$$\langle \nabla\theta(\mathbf{x})\rangle = \mathbf{G} \tag{1.1b}$$

which plays the role of a "boundary" one. In equations (1.1a) $\mathbf{q}(\mathbf{x})$ is the flux vector and $\theta(\mathbf{x})$ is the random temperature field. Hereafter $\langle\cdot\rangle$ denotes ensemble averaging.

The random problem (1.1) possesses a solution, in a statistical sense, which is unique under the natural condition $0 < k_1 \le \kappa(\mathbf{x}) \le k_2 < \infty$, see [11]. This means, let us recall [1], that all multipoint moments of the temperature field $\theta(\mathbf{x})$, and the joint moments of $\theta(\mathbf{x})$ and $\kappa(\mathbf{x})$, can be specified by means of the known moments of the conductivity field. In particular, among the joint moments, the simplest one-point moment defines the well-known effective conductivity κ^* of the medium through the relation

$$\mathbf{Q} = \langle \mathbf{q}(\mathbf{x})\rangle = \langle \kappa(\mathbf{x})\nabla\theta(\mathbf{x})\rangle = \kappa^*\mathbf{G} \tag{1.2}$$

(having assumed statistical homogeneity and isotropy). Note that the definition (1.2) of the effective conductivity κ^* reflects the "homogenization" of the problem under study, in the sense that from a macroscopic point of view, when only the macroscopic values of the flux and temperature gradient are of interest, the medium behaves as if it were homogeneous with a certain macroscopic conductivity κ^*. This interpretation explains why κ^* and its counterparts, say, the effective elastic moduli, have been extensively studied in the literature on homogenization. There one can find a number of rigorous or approximate schemes of their evaluation, especially, in the context of mechanics of heterogeneous and composite media, see, e.g. [9, 21, 14] *et al.* However, κ^* is only a tiny part of the full statistical solution of the random problem (1.1). Moreover, its evaluation *cannot* be torn away from the full statistical solution of (1.1), i.e. of specifying *all* needed multipoint moments, as pointed out for the first time by Brown [8]. (The latter fact explains the failure of all schemes that try to determine solely the effective property κ^* without trying to solve the whole stochastic problem (1.1).) Besides, there are plenty of reasons why one should pay much more attention to other statistical characteristics of random fields like $\theta(\mathbf{x})$ in (1.1), that appear in problems in random heterogeneous media. For instance, in the context of waves in random media or turbulence phenomena, one of the most important quantities is the variance of local fields, connected with the square of its fluctuation, see [1].

The (undimensional) variances, which we shall discuss hereafter, are defined as

$$\sigma^2_{\nabla\theta} = \frac{\langle |\nabla\theta'(\mathbf{x})|^2\rangle}{G^2}, \quad \sigma^2_q = \frac{\langle |q'(\mathbf{x})|^2\rangle}{Q^2}; \tag{1.3}$$

the primes denote in what follows the fluctuating parts of the respective random fields, so that, in particular, $\nabla\theta'(\mathbf{x}) = \nabla\theta(\mathbf{x}) - \mathbf{G}$, and hence $\langle\nabla\theta'(\mathbf{x})\rangle = 0$.

It is noted that for any two-point medium the variances $\sigma^2_{\nabla\theta}$ and σ^2_q are simply interconnected. Indeed, since the conductivity field $\kappa(\mathbf{x})$ takes only two values, κ_f and κ_m, we have

$$\kappa^2(\mathbf{x}) = (\kappa_f + \kappa_m)\kappa(\mathbf{x}) - \kappa_f\kappa_m$$

and hence

$$\left\langle \mathbf{q}^2(\mathbf{x}) \right\rangle = \left\langle \kappa^2(\mathbf{x})|\nabla\theta(\mathbf{x})|^2 \right\rangle = (\kappa_f + \kappa_m)\kappa^* G^2 - \kappa_f\kappa_m \left\langle |\nabla\theta(\mathbf{x})|^2 \right\rangle,$$

having used (1.2). A simple algebra yields eventually

$$\sigma^2_q = -\frac{\kappa_f\kappa_m}{\kappa^{*2}}\sigma^2_{\nabla\theta} - \frac{(\kappa^* - \kappa_f)(\kappa^* - \kappa_m)}{\kappa^{*2}}. \tag{1.4}$$

Let us point out immediately that the study of variances in particular, and of the multipoint moments in general, is much more complicated than that of the effective properties due to the fact that, as a matter of fact, no variational principles for the former have been proposed and applied in the literature. (Though, see the book [5, p. 143], where an extremely concise exposition and some ideas along this line are indicated.)

To the best of the authors' knowledge an investigation of the variances, in addition to the effective properties in the scalar conductivity context, was initiated by Beran *et al.* [2, 4, 3]. In particular, Beran [2] obtained bounds on the variances through the effective properties, investigated in great detail in the literature. The Beran's estimates are quite crude and this is inevitable since they are applicable to *any* statistically homogeneous and isotropic medium.

More restrictive bounds can be obtained only if additional information about the medium constitution is available and the needed random fields are specified at least to a certain extent. This is the case with random dispersion of spheres which we shall study in more detail later on.

The above mentioned results of Beran indicated that there may exist more intimate connection between variances and effective properties. Indeed, as shown independently by several authors [6, 7, 15], the variance is simply connected to the derivatives of the effective conductivity $\kappa^* = \kappa^*(\kappa_f, \kappa_m)$, treated as a function of the material properties κ_f, κ_m of the two constituents in a binary medium. This is an interesting and important result, but its practical application is limited by the fact that very rarely rigorous analytical formulae for $\kappa^*(\kappa_f, \kappa_m)$ are known for realistic random constitution. Rigorous bounds on $\kappa^*(\kappa_f, \kappa_m)$ are well-known, of course, but they obviously cannot supply any estimates for the appropriate derivatives.

In the present note we shall employ another method for studying variation in random dispersions. Namely, we shall use the fact that for the latter the full statistical solution of the problem (1.1) can be conveniently constructed by means of the functional series approach, see [10, 16, 17]. Moreover, the first two kernels of the series can be explicitly found, which results, in particular, in a formula for the needed variances, which is *exact* to the order c^2, where c is the volume fraction of the spheres. Then the observation that some of the terms in the appropriate formulae are sign-definite produces a bound on the variances which, as it turns out, can be simply expressed by means of the c^2-coefficient of the effective conductivity. The latter, as it is well-known, represents a quantity extensively studied in the literature.

2. C^2–SOLUTION OF THE BASIC PROBLEM (1.1) FOR DISPERSIONS OF SPHERES

To get certain rigorous results for the variance, one should somewhat narrow the class of two-phase random media. To this end, consider in more detail here a random dispersion of spheres as a typical representative of the wide and important class of particulate microinhomogeneous media, extensively studied in the literature.

Let us recall first the so-called virial (or density) expansion of κ^* in powers of the volume fraction c of the spheres:

$$\frac{\kappa^*}{\kappa_m} = 1 + a_1 c + a_2 c^2 + \cdots \tag{2.1}$$

Note that hereafter we shall try to cover simultaneously both 3-D case (dispersion of spheres) and its 2-D counterpart — a matrix containing an array of circular and aligned fibers subjected to a macroscopic gradient perpendicular to fiber axes. Depending on dimension, a will denote either the sphere radius (3-D) or the radius of the cylinder cross-section (2-D). For the volume fraction c of the spheres we have $c = nV_a$, $V_a = \frac{4}{3}\pi a^3$ in the 3-D case, or $c = nS_a$, $S_a = \pi a^2$ in 2-D, n is the number density of the spheres or of the fibers.

As it is well-known, the coefficient a_1 in (2.1) is the only thing rigorously calculated by Maxwell [20] in his classical theory of macroscopic conductivity of a random dispersion. The Maxwell result reads

$$a_1 = d\beta_d, \quad \beta_d = \frac{[\kappa]}{\kappa_f + (d-1)\kappa_m}, \quad [\kappa] = \kappa_f - \kappa_m; \tag{2.2}$$

hereafter $d = 3$ in the 3-D case and $d = 2$ in the 2-D-case.

For higher sphere fraction, the Maxwell theory [20] yields the well-known approximate relation

$$\frac{\kappa^*}{\kappa_m} = 1 + \frac{d\beta_d c}{1 - \beta_d c} \tag{2.3}$$

— the so-called Maxwell (or Clausius-Mossotti) formula [14]. The latter produces in turn the following approximation for the c^2-coefficient, namely:

$$a_2 = d\beta_d^2, \quad d = 2, 3. \tag{2.4}$$

The rigorous evaluation of a_2 has attracted the attention of many authors, because this is the simplest case in which the multiparticle interaction shows up in a nontrivial way. We refer here to the papers [22, 13, 12, 16] *et al.*, where a_2 has been expressed in a closed form, making use of the zero-density limit $g_0(r)$ of the so-called radial distribution function for the spheres, and of the one- and two-inclusion fields for the conductivity problem under study. (Recall that the radial distribution function $g(r) = f_2(r)/n^2$, where $r = |\mathbf{y}_1 - \mathbf{y}_2|$, so that $g(r) = g_0(r) + o(n)$ in the dilute limit $n \to 0$; $f_2(r) = f_2(\mathbf{y}_1 - \mathbf{y}_2)$ denotes the two-point probability density for the set of sphere centers.) In the 2-D case (fiber-reinforced material), the coefficient a_2 has been evaluated analytically by the authors [19], making use of the earlier reasoning of Peterson and Hermans [22].

As already mentioned, the full statistical solution of the problem (1.1) for a random dispersion can be conveniently constructed by means of the functional series approach, see [10, 16, 17] for details. In particular, as shown by one of the authors [16], the temperature gradient fluctuation in the dispersion of spheres, correct to the order c^2, has the form of the truncated functional series:

$$\begin{aligned}\nabla\theta'(\mathbf{x}) &= \int \nabla_x T_1(\mathbf{x}-\mathbf{y}) D_\psi^{(1)}(\mathbf{y})\, d\mathbf{y} \\ &+ \iint \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2) D_\psi^{(2)}(\mathbf{y}_1,\mathbf{y}_2)\, d\mathbf{y}_1 d\mathbf{y}_2,\end{aligned} \tag{2.5}$$

where

$$D_\psi^{(1)} = \psi'(\mathbf{x}) = \psi(\mathbf{x}) - n, \tag{2.6}$$

$$\begin{aligned}D_\psi^{(2)}(\mathbf{y}_1,\mathbf{y}_2) &= \psi(\mathbf{y}_1)[\psi(\mathbf{y}_2) - \delta(\mathbf{y}_1 - \mathbf{y}_2)] \\ &- n g_0(\mathbf{y}_1 - \mathbf{y}_2)[\psi'(\mathbf{y}_1) + \psi'\mathbf{y}_2)] - n^2 g_0(\mathbf{y}_1 - \mathbf{y}_2),\end{aligned} \tag{2.7}$$

and the kernel $T_2(\mathbf{y}_1, \mathbf{y}_2)$ is a symmetric function of its arguments $\mathbf{y}_1$ and $\mathbf{y}_2$. Recall that n is the number density of the spheres, so that their volume fraction is $c = n\,\frac{4}{3}\pi a^3$ in 3-D and $c = n\,\pi a^2$ in 2-D. The integrals hereafter are over the entire space $\mathbb{R}^d$ if the integration domain is not explicitly indicated. In (2.5) to (2.7)

$$\psi(\mathbf{x}) = \sum_\alpha \delta(\mathbf{x} - \mathbf{x}_\alpha)$$

is the random density field of Stratonovich [23], generated by the random set $\{\mathbf{x}_\alpha\}$ of sphere centers. The fields $D_\psi^{(1)}$, $D_\psi^{(2)}$ and the constant field $D_\psi^{(0)} = 1$ constitute a c^2-orthogonal family, i.e.

$$\left\langle D_\psi^{(1)} \right\rangle = \left\langle D_\psi^{(2)} \right\rangle = \left\langle D_\psi^{(1)} D_\psi^{(2)} \right\rangle = o(c^2), \tag{2.8}$$

which means that in the c^2-analysis performed below the averaged values in (2.8) can be neglected. We have also

$$\begin{aligned}&\left\langle D_\psi^{(1)}(\mathbf{y}_1) D_\psi^{(1)}(\mathbf{y}_2) \right\rangle = n\delta_{12} - n^2 R_0(\mathbf{y}_1 - \mathbf{y}_2), \quad R_0(\mathbf{y}) = 1 - g_0(\mathbf{y}), \\ &\left\langle D_\psi^{(2)}(\mathbf{y}_1,\mathbf{y}_2) D_\psi^{(2)}(\mathbf{y}_3,\mathbf{y}_4) \right\rangle = n^2 g_0(\mathbf{y}_1 - \mathbf{y}_2)\big(\delta_{13}\delta_{24} + \delta_{14}\delta_{23}\big),\end{aligned} \tag{2.9}$$

where $\delta_{ij} = \delta(\mathbf{y}_i - \mathbf{y}_j)$, and g_0 is the above mentioned low-density limit of the radial distribution function for the set of sphere centers, which is the only statistical characteristics of the dispersion needed in the c^2-statistical solution for the temperature gradient. The relations (2.9), as well as *all* formulae in the sequel, are correct to the order c^2 only.

The kernel T_1 that enters (2.5) has the form

$$T_1(\mathbf{x}) = T_{10}(\mathbf{x}) + nT_{11}(\mathbf{x}). \tag{2.10}$$

In (2.10) $T_{10}(\mathbf{x})$ is the "one-sphere" solution, i.e. the disturbance field superimposed by a single spherical inclusion of radius a (located at the origin) on a temperature field in the matrix with constant gradient $\mathbf{G}$ at infinity. Recall that $T_{10}(\mathbf{x})$ solves the equation

$$\kappa_m \Delta T_{10}(\mathbf{x}) + [\kappa]\nabla \cdot \{h(\mathbf{x})[\mathbf{G} + \nabla T_{10}]\} = 0 \tag{2.11a}$$

and hence

$$T_{10}(\mathbf{x}) = d\beta_d \mathbf{G} \cdot \nabla\varphi(\mathbf{x}), \tag{2.11b}$$

where $\varphi(\mathbf{x})$ is the Newtonian potential for a sphere (in 3-D) or for a circle (in 2-D) of radius a; $h(\mathbf{x})$ denotes the characteristic function of a single sphere (or disk in 2-D) centered at the origin, and β_d was defined in (2.2). As it is well known, the potential $\varphi(\mathbf{x})$ solves the equation

$$\Delta\varphi(\mathbf{x}) + h(\mathbf{x}) = 0, \tag{2.12}$$

which implies, in particular, that

$$h(\mathbf{x})\nabla T_{10}(\mathbf{x}) = -d\beta_d h(\mathbf{x})\mathbf{G}, \quad \Delta T_{10}(\mathbf{x}) = -d\beta_d \nabla \cdot \big(h(\mathbf{x})\mathbf{G}\big). \tag{2.13}$$

To specify $T_{11}(\mathbf{x})$, we should first note that to the order c^2 the kernel T_2 in (2.5) equals T_{20}. The latter solves the equation

$$\begin{aligned} 2\kappa_m \Delta T_{20}(\mathbf{x}, \mathbf{x}-\mathbf{z}) + [\kappa]\,\nabla \cdot \{2\big[h(\mathbf{x}) + h(\mathbf{x}-\mathbf{z})\big]\nabla T_{20}(\mathbf{x}, \mathbf{x}-\mathbf{z}) \\ + h(\mathbf{x})\nabla T_{10}(\mathbf{x}-\mathbf{z}) + h(\mathbf{x}-\mathbf{z})\nabla T_{10}(\mathbf{x})\} = 0. \end{aligned} \tag{2.14}$$

The differentiation hereafter is with respect to $\mathbf{x}$, and $\mathbf{z}$ plays the role of a parameter. Hence

$$2T_{20}(\mathbf{x}-\mathbf{z}; \mathbf{x}) = T^{(2)}(\mathbf{x}; \mathbf{z}) - T_{10}(\mathbf{x}) - T_{10}(\mathbf{x}-\mathbf{z}) \tag{2.15}$$

with $T^{(2)}(\mathbf{x}; \mathbf{z})$ denoting the "two-sphere" solution, i.e. the disturbance to the temperature field in an unbounded matrix, introduced by a pair of identical spherical inhomogeneities with centers at the origin and at the point $\mathbf{z}$, $|\mathbf{z}| > 2a$, when the temperature gradient at infinity equals $\mathbf{G}$. Thus

$$\kappa_m \Delta T^{(2)}(\mathbf{x}; \mathbf{z}) + [\kappa]\,\nabla \cdot \Big\{[h(\mathbf{x}) + h(\mathbf{x}-\mathbf{z})]\,[\mathbf{G} + \nabla T^{(2)}(\mathbf{x}; \mathbf{z})]\Big\} = 0, \tag{2.16}$$

which is the counterpart of the "single-sphere" equation (2.11a).

The coefficient $T_{11}(\mathbf{x})$ can be represented as

$$T_{11}(\mathbf{x}) = \beta_d V_a T_{10}(\mathbf{x}) + 2L_{20}(\mathbf{x}), \quad L_{20}(\mathbf{x}) = \int_{|\mathbf{z}|>2a} g_0(\mathbf{z}) T_{20}(\mathbf{x}-\mathbf{z};\mathbf{x})\, d\mathbf{z}. \quad (2.17)$$

To calculate the effective conductivity κ^* through the kernels T_1 and T_2, note that the conductivity field $\kappa(\mathbf{x})$ of the dispersion has a form, similar to (2.5), namely,

$$\kappa(\mathbf{x}) = \langle\kappa\rangle + \kappa'(\mathbf{x}), \quad \kappa'(\mathbf{x}) = [\kappa] \int h(\mathbf{x}-\mathbf{y}) D_\psi^{(1)}(\mathbf{y})\, d\mathbf{y}. \quad (2.18)$$

That is why, inserting (2.5) and (2.17) into (1.2) and using the orthogonality of the fields $D_\psi^{(1)}$ and $D_\psi^{(2)}$, see (2.8), give

$$\begin{aligned} \kappa^* \mathbf{G} &= \langle \kappa(\mathbf{x}) \nabla\theta(\mathbf{x}) \rangle = \langle\kappa\rangle\, \mathbf{G} + \langle \kappa'(\mathbf{x}) \nabla\theta'(\mathbf{x}) \rangle \\ &= \langle\kappa\rangle\, \mathbf{G} + n[\kappa] \int h(\mathbf{x}) \nabla S(\mathbf{x})\, d\mathbf{x} \end{aligned} \quad (2.19)$$

with the function

$$S(\mathbf{x}) = T_1(\mathbf{x}) - n \int T_1(\mathbf{x}-\mathbf{y}) R_0(\mathbf{y})\, d\mathbf{y} = S_0(\mathbf{x}) + nS_1(\mathbf{x}), \quad (2.20)$$

so that, due to (2.10),

$$S_0(\mathbf{x}) = T_{10}(\mathbf{x}), \quad S_1(\mathbf{x}) = T_{11}(\mathbf{x}) - \int T_{10}(\mathbf{x}-\mathbf{y}) R_0(\mathbf{y})\, d\mathbf{y}. \quad (2.21)$$

Inserting (2.10) and (2.21) into (2.9) and comparing the result with (2.16) give for the virial coefficients a_1 and a_2:

$$a_1 = \left(1-\beta_d\right) \frac{[\kappa]}{\kappa_m} = d\beta_d, \quad (2.22)$$

which indeed coincides with the exact value, given in (2.17), and

$$a_2 = d\beta_d^2 + a_2', \quad a_2'\mathbf{G} = 2\,\frac{[\kappa]}{\kappa_m} \frac{1}{V_a^2} \int h(\mathbf{x}) \nabla L_{20}(\mathbf{x})\, d\mathbf{x}. \quad (2.23)$$

Note that the integrals in (2.17) and (2.23) are conditionally convergent, the mode of integration being extracted in the course of the statistical solution of the problem (1.1), see [16, 18] for details and discussion. Namely, one should integrate first with respect to the angular coordinates at fixed $r = R$ and only then with respect to the radial coordinate R. This mode of integration will be tacitly used hereafter to avoid convergent difficulties for some of the integrals in Section 4.

Let us point out finally that though the formula for a_2' in (2.23) is written for a 3-D dispersion, it holds as well in the 2-D case, with the only change that the volume V_a of the inclusions is replaced by their area S_a and the integrals are two-tuple. The same will hold true in all formulae in the sequel. Moreover, a closed form analytic formula for a_2' in the 2-D case was derived, let us recall, by the authors in [19].

3. C^2–FORMULA FOR THE VARIANCES $\sigma^2_{\nabla\theta}$ AND σ^2_Q

Insert the representation (2.5) into the definition (1.3) of the variance. Due to the orthogonality of the fields $D^{(1)}_\psi$ and $D^{(2)}_\psi$, see (2.8), we get

$$\sigma^2_{\nabla\theta} = \frac{1}{G^2}\left(\mathcal{A}_1 + \mathcal{A}_2\right), \tag{3.1}$$

where

$$\mathcal{A}_1 = \iint \nabla_x T_1(\mathbf{x}-\mathbf{y}_1)\cdot\nabla_x T_1(\mathbf{x}-\mathbf{y}_2)\left\langle D^{(1)}_\psi(\mathbf{y}_1)D^{(1)}_\psi(\mathbf{y}_2)\right\rangle\, d\mathbf{y}_1 d\mathbf{y}_2, \tag{3.2}$$

$$\begin{aligned}\mathcal{A}_2 &= \iiiint \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2)\cdot\nabla_x T_2(\mathbf{x}-\mathbf{y}_3, \mathbf{x}-\mathbf{y}_4) \\ &\quad\times\left\langle D^{(2)}_\psi(\mathbf{y}_1,\mathbf{y}_2)D^{(2)}_\psi(\mathbf{y}_3,\mathbf{y}_4)\right\rangle\, d\mathbf{y}_1\, d\mathbf{y}_2\, d\mathbf{y}_3\, d\mathbf{y}_4.\end{aligned} \tag{3.3}$$

Note that

$$\mathcal{A}_2 = \left\langle \left| \iint \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2) D^{(2)}_\psi(\mathbf{y}_1,\mathbf{y}_2)\, d\mathbf{y}_1\, d\mathbf{y}_2 \right|^2 \right\rangle,$$

which implies immediately that $\mathcal{A}_2 \geq 0$, and hence

$$\sigma^2_{\nabla\theta} \geq \frac{1}{G^2}\mathcal{A}_1. \tag{3.4}$$

An evaluation of the term $\mathcal{A}_1$ yields thus the lower estimate (3.4) of the variance.

Note that the evaluation of $\mathcal{A}_1$ is much easier than that of $\mathcal{A}_2$. The reason, as we shall see below (Section 4), is that to evaluate $\mathcal{A}_1$ only the single sphere solution T_{10} is needed together with the values, assumed known, of the c^2-term a_2 in (2.16). At the same time $\mathcal{A}_2$ involves already the double-sphere field $T^{(2)}$ in a nontrivial way, which essentially complicates the investigation. Note also that the term $\mathcal{A}_2$ has the order $O(c^2)$ (see (2.9)), so that the lower estimate (3.4) gives correct to the order $O(c)$ results in the dilute case $c \ll 1$. Hence from (3.4) the exact value of the c-coefficient A_1, see (3.12), and a lower bound on the c^2-coefficient A_2 in the virial expansion, see (4.5) below, of the variance $\sigma^2_{\nabla\theta}$ will follow in particular.

4. EVALUATION OF THE TERM $\mathcal{A}_1$

Using (2.9) into (3.2), we get

$$\mathcal{A}_1 = n\int \nabla T_1(\mathbf{x})\cdot\nabla S(\mathbf{x})\, d\mathbf{x} \tag{4.1}$$

with the function $S(\mathbf{x})$ defined in (2.20), and hence due to (2.10) and (2.21)

$$\mathcal{A}_1 = n\big(\mathcal{A}_{11} + n\mathcal{A}_{12}\big), \tag{4.2}$$

$$\mathcal{A}_{11} = \int \nabla T_{10}(\mathbf{x}) \cdot \nabla T_{10}(\mathbf{x})\, d\mathbf{x}, \tag{4.3}$$

$$\mathcal{A}_{12} = 2\int \nabla T_{10}(\mathbf{x}) \cdot \nabla T_{11}(\mathbf{x})\, d\mathbf{x} - \int \nabla T_{10}(\mathbf{x}) \cdot \int \nabla_x T_{10}(\mathbf{x}-\mathbf{y}) R_0(\mathbf{y})\, d\mathbf{y} d\mathbf{x}. \tag{4.4}$$

Let

$$\sigma^2_{\nabla\theta} = A_1 c + A_2 c^2 + \cdots, \quad \sigma_q^2 = B_1 c + B_2 c^2 + \cdots \tag{4.5}$$

be the virial expansions of the variances, similar to the classical expansion (2.1) for the effective conductivity. The leading terms A_1 and B_1 can be easily found, since this requires an evaluation of the integral $\mathcal{A}_{11}$ from (4.3). To this end integrate by parts in (4.3), use (2.13), and once again integrate by parts

$$\mathcal{A}_{11} = -\int T_{10}(\mathbf{x}) \cdot \Delta T_{10}(\mathbf{x})\, d\mathbf{x} = d\beta_d \mathbf{G} \cdot \int h(\mathbf{x}) \nabla T_{10}(\mathbf{x})\, d\mathbf{x},$$

so that using (2.13) once more gives

$$\mathcal{A}_{11} = d\beta_d^2 V_a^2 G^2.$$

Together with (4.2) this gives the leading terms of the virial expansions (4.6) of the variances, namely,

$$A_1 = \beta_d a_1 = d\beta_d^2, \quad B_1 = d(d-1)\beta_d^2 = (d-1)A_1. \tag{4.6}$$

Turning to the evaluation of $\mathcal{A}_{12}$, we start with the first integral in (4.4). Integrating by parts and using (2.13) together with the formula (2.15) for T_{11}, we have

$$\int \nabla T_{10}(\mathbf{x}) \cdot \nabla T_{11}(\mathbf{x})\, d\mathbf{x} = \int \Delta T_{10}(\mathbf{x}) T_{11}(\mathbf{x})\, d\mathbf{x} = -d\beta_d \int \nabla \cdot (h(\mathbf{x})\mathbf{G}) T_{11}(\mathbf{x})\, d\mathbf{x}$$

$$= -d\beta_d \mathbf{G} \cdot \int h(\mathbf{x}) \nabla T_{11}(\mathbf{x})\, d\mathbf{x} = -d\beta_d \mathbf{G} \cdot \int h(\mathbf{x}) \left[\beta_d V_a \nabla T_{10}(\mathbf{x}) + 2\nabla L_{20}(\mathbf{x})\right] d\mathbf{x}$$

$$= d\beta_d^3 V_a^2 G^2 - 2d\beta_d \mathbf{G} \cdot \int h(\mathbf{x}) \nabla L_{20}(\mathbf{x})\, d\mathbf{x}.$$

The second term in the last formula is connected with the c^2-value a_2 of the effective conductivity, see (2.23), so that

$$\int \nabla T_{10}(\mathbf{x}) \cdot \nabla T_{11}(\mathbf{x})\, d\mathbf{x} = \left(\beta_d^3 - \frac{d\kappa_m}{\kappa_f + (d-1)\kappa_m}\left(a_2 - d\beta_d^2\right)\right) dV_a^2 G^2. \tag{4.7}$$

The second integral in (4.4) is similarly simplified through integration by parts, and applying (2.13):

$$\int \nabla T_{10}(\mathbf{x}) \cdot \int \nabla_x T_{10}(\mathbf{x}-\mathbf{y}) R_0(\mathbf{y})\, d\mathbf{y} d\mathbf{x} = -d\beta_d \mathbf{G} \cdot \int\int \nabla T_{10}(\mathbf{x}) h(\mathbf{x}-\mathbf{y}) R_0(\mathbf{y})\, d\mathbf{y} d\mathbf{x}$$

$$= -d^2\beta_d^2\mathbf{G}\cdot\iint h(\mathbf{x}-\mathbf{y})R_0(\mathbf{y})\nabla\nabla\varphi(\mathbf{x})\,d\mathbf{y}d\mathbf{x}\cdot\mathbf{G}, \tag{4.8}$$

having used the representation (2.11b) of the single sphere solution $T_{10}(\mathbf{x})$ through the Newtonian potential of the sphere. But, due to the isotropy of the latter, the integral in the right-hand side of (4.8) represents a second rank isotropic tensor, so that

$$\iint h(\mathbf{x}-\mathbf{y})R_0(\mathbf{y})\nabla\nabla\varphi(\mathbf{x})\,d\mathbf{y}d\mathbf{x} = \gamma\mathbf{I} \tag{4.9}$$

and, upon contraction,

$$\gamma d = \iint h(\mathbf{x}-\mathbf{y})R_0(\mathbf{y})\Delta\varphi(\mathbf{x})\,d\mathbf{y}d\mathbf{x} = -\iint h(\mathbf{x}-\mathbf{y})h(\mathbf{x})R_0(\mathbf{y})\,d\mathbf{y}d\mathbf{x},$$

see (2.12). The product $h(\mathbf{x}-\mathbf{y})h(\mathbf{x})$ does not vanish however only in the sphere $|\mathbf{y}| \le 2a$, where $R_0(\mathbf{y}) = 1 - g_0(\mathbf{y}) = 1$, due to the nonoverlapping assumption ($g_0(\mathbf{y}) = 0$ if $|\mathbf{y}| \le 2a$, since the spheres are forbidden to intersect). Thus $\gamma d = -V_a^2$ and from (4.8) and (4.9) it follows that

$$\int \nabla T_{10}(\mathbf{x})\cdot\int \nabla_x T_{10}(\mathbf{x}-\mathbf{y})R_0(\mathbf{y})\,d\mathbf{y}d\mathbf{x} = -d\beta_d^2V_a^2G^2. \tag{4.10}$$

Combining (4.7) and (4.10) into (4.4) and using (3.1) and (2.5) give eventually

$$\mathcal{A}_{1\cdot2} = \left(d\beta_d^2 - 2(1-\beta_d)a_2\right)V_a^2G^2.$$

From (3.4), (4.4) and the last formula, as it was already discussed, immediately follows the lower bound

$$A_{12} \le A_2, \quad A_{12} = d\beta_d^2 - 2(1-\beta_d)a_2, \tag{4.11}$$

for the c^2-coefficient of the variance $\sigma^2_{\nabla\theta}$. Using (1.4) gives in turn the following upper bound for the respective coefficient of the flux variance σ_q^2, namely,

$$B_2 \le B_{12}, \quad B_{12} = d\beta_d^2\left[d(1-2\alpha+2\alpha\beta_d)-\alpha\right] + \left[2(1-\beta_d)\alpha+\alpha-1\right]a_2. \tag{4.12}$$

5. CONCLUDING REMARKS

The estimates (4.11) and (4.12) for the c^2-coefficients in the virial expansions (4.5) of the variances represent the central result of the present note. They have been obtained without using variational arguments — instead the full statistical solution of the problem (1.1) has been appropriately exploited. The estimates account for the statistics of the dispersion through the well-known and extensively studied in the literature c^2-coefficient a_2 of the effective conductivity. Moreover, they remain finite for high-contrast media, when the ratio $\alpha = \kappa_f/\kappa_m$ goes to 0 or ∞. A more detailed investigation of the estimates (4.11), (4.12) and of their implications will be performed elsewhere.

ACKNOWLEDGEMENTS. The partial support of the Fund "Scientific Investigations" at the "St. Kliment Ohridski" University of Sofia under Grant no. 286/97 is gratefully acknowledged.

REFERENCES

1. Beran, M. Statistical continuum theories. John Wiley, New York, 1968.
2. Beran, M. Bounds on field fluctuations in a random medium. *J. Appl. Phys.*, **39**, 1968, 5712–5714.
3. Beran, M. Field fluctuations in a two-phase random medium. *J. Math. Phys.*, **21**, 1980, 2583–2585.
4. Beran, M., J. J. McCoy. Mean field variation in random media. *Q. Appl. Math.*, **28**, 1970, 245–258.
5. Berdichevsky, V. Variational principles of continuum mechanics. Nauka, Moscow, 1983 (in Russian).
6. Bergman, D. J. The dielectric constant of a composite material — a problem in classical physics. *Phys. Reports*, **43**C, 1978, 377–407.
7. Bobeth, M., G. Diener. Field fluctuations in multicomponent mixtures. *J. Mech. Phys. Solids*, **34**, 1986, 1–17.
8. W. F. Brown. Solid mixture permittivities. *J. Chem. Phys.*, **23**, 1955, 1514–1517.
9. Christensen, R. C. Mechanics of composite materials. John Wiley, New York, 1979.
10. Christov, C. I., K. Z. Markov. Stochastic functional expansion for random media with perfectly disordered constitution. *SIAM J. Appl. Math.*, **45**, 1985, 289–311.
11. Golden, K., G. Papanicolaou. Bounds for effective properties of heterogeneous media by analytic continuation. *Comm. Math. Phys.*, **90**, 1983, 473–491.
12. Felderhof, B. U., G. W. Ford, E. G. D. Cohen. Two-particle cluster integral in the expansion of the dielectric constant. *J. Stat. Phys.*, **28**, 1982, 1649–1672.
13. Jeffrey, D. J. Conduction through a random suspension of spheres. *Proc. Roy. Soc. London*, **A335**, 1973, 355–367.
14. Landauer, R. Electrical conductivity in inhomogeneous media. In: *Electrical transport and optical properties of inhomogeneous media*, J. C. Garland, D. B. Tanner, eds., Am. Inst. Phys., New York, 1978, 2–43.
15. Matheron, G. Quelques inégalités pour la perméabilité effective d'un milieu poreux hélérogeène. *Cahiers de Géostatistique*, Fasc. 3, 1993, 1–20.
16. Markov, K. Z. On the heat propagation problem for random dispersions of spheres. *Math. Balkanica (New Series)*, **3**, 1989, 399–417.
17. Markov, K. Z. On the factorial functional series and their application to random media. *SIAM J. Appl. Math.*, **51**, 1991, 172–186.
18. Markov, K. Z., C. I. Christov. On the problem of heat conduction for random dispersions of spheres allowed to overlap. *Mathematical Models and Methods in Applied Sciences*, **2**, 1992, 249–269.
19. Markov, K. Z., K. S. Ilieva. A note on the c^2-term of the effective conductivity for random dispersions. *Ann. Univ. Sofia, Fac. Math. Méc., Livre 2*, **84**, 1993, 123–137.
20. Maxwell, J. C. A treatise on electricity and magnetism. Dover, New York, 1954 (republication of 3rd edition of 1891).

21. Nemat-Nasser, S., M. Hori. Micromechanics: Overall properties of heterogeneous solids. Elsevier, 1993.
22. Peterson, J. M., J. J. Hermans. The dielectric constant of nonconducting suspensions. *J. Composite Materials*, **3**, 1969, 338–354.
23. Stratonovich, R. L. Topics in theory of random noises. Vol. 1. Gordon and Breach, New York, 1963.

Received March 9, 1998
Revised May 28, 1998

Keranka S. ILIEVA
Faculty of Mathematics and Informatics
"K. Preslavski" University of Shumen
BG-9700 Shumen
Bulgaria

Konstantin Z. MARKOV
Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 Blvd J. Bourchier, P. O. Box 48
BG-1164 Sofia, Bulgaria
E-mail: kmarkov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 91, 1997

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 91, 1997

METHOD OF VARIATIONAL IMBEDDING FOR IDENTIFICATION OF HEAT-CONDUCTION COEFFICIENT FROM OVERPOSED BOUNDARY DATA

TCHAVDAR T. MARINOV, CHRISTO I. CHRISTOV

We consider the inverse problem of identifying a spatially varying coefficient in diffusion equation from overspecified boundary conditions. We make use of a technique called Method of Variational Imbedding (MVI) which consists in replacing the original inverse problem by the boundary value problem for the Euler-Lagrange equations presenting the necessary conditions for minimization of the quadratic functional of the original equations. The latter is well-posed for redundant data at boundaries. The equivalence of the two problems is demonstrated. In the recent authors' works difference scheme and algorithm have been created to apply MVI to the problem under consideration. In the present work we show that the number of boundary conditions can be decreased, replacing them with the so-called "natural conditions" for minimization of a functional. A difference scheme of splitting type is employed and featuring examples are elaborated numerically.

Keywords: inverse problem, coefficient identification, diffusion equation

1991/1995 Math. Subject Classification: 35N10, 35R30, 65R30, 65N06

1. INTRODUCTION

The attention attracted by the ill-posed (inverse, incorrect, etc.) problems constantly increases during the last decade because of their practical importance. The optimization of technological processes and identification of material properties yield as a rule mathematical problems in which initial or boundary conditions are missing (or overdetermined), while additional information is available for the needed solution (or additional unknown functions are present).

At the same time the incorrect problems have a great potential for inciting the development of the applied mathematics itself. According to [1]: "The analysis of inverse problems is of relevant importance for mathematical modelling and, in general, for applied mathematics. With this in mind, the applied mathematician should attempt the solution of problems without artificial simplification, which may obscure the information he hopes to obtain from the real system."

Naturally, the whole variety of the mentioned "non-standard" problems goes well beyond the framework of the Hadamard's [10] definition of incorrect problem. His definition does not cover all of them and is pertinent only to stability of a solution. For this reason, when we speak of "inverse problems," we mean the whole set of problems which are unusually or inconveniently posed. To distinguish from the problems for which Hadamard's definition applies, we shall call the latter "incorrect in the sense of Hadamard." In this instance we shall follow the classification from [1].

The work of Hadamard spurred significant activity for creating regularizing procedures (see, e.g., [15]) for the problems that are incorrect in the sense of Hadamard, e.g. for smoothing the data in order to evade the instability provoked by the pollution of the data. Such an approach has an important implication for the practical problems. At the same time the very notion of replacing the ill-formulated (e.g., ill-specified and inverse) or ill-posed by a well-formulated mathematical problem is of not lesser importance. Indeed, if one succeeds in doing so, one arrives at a problem that is also correct in the sense of Hadamard and then it is automatically regularizing the data if some pollution is present. To this end the Method of Variational Imbedding (MVI — for brevity) was proposed by the second author of the present paper. The idea of MVI is to replace an incorrect problem with the well-posed problem for minimization of quadratic functional of the original equations, i.e. we "embed" the original incorrect problem in a higher-order boundary value problem which is well-posed. For the latter a difference scheme and numerical algorithm for its implementation can easily be constructed.

The inverse problems for diffusion equation can be roughly separated into three principal classes. The first is the coefficient identification from over-posed data at the boundary; the second is the identification of the thermal regimes at one of the spatial boundaries from over-posed data at the other one (the parabolic version of the so-called analytical continuation); the third is the reversed-time problem for identification of initial temperature distribution from the known distribution at certain later moment of time. The second problem appears to be the most studied, due to the successful technique proposed in [14, 11], called "quasi-reversibility method" (see also [15]). Apart from being inverse, the second and the third problems are also incorrect in the sense of Hadamard (see [10]). The first one is merely inverse without being incorrect in the strict sense of amplifying the disturbances. The problem then is how to create the appropriate algorithm. This is the aim of the present paper. We make use of the above mentioned MVI technique, which consists in replacing the original inverse problem by the boundary value problem for the Euler-Lagrange equations for minimization of the quadratic functional of the original equations.

The first application of MVI was to the problem of identification of homoclinic trajectories as an inverse problem [2] (see also the ensuing works [8, 7]). The way to treat the classical inverse problems by means of MVI was sketched in [3–5]. In the recent authors' work [9] difference scheme and algorithm have been created to apply MVI to the problem under consideration. In the present work we show that the number of boundary conditions can be decreased replacing them with the so-called "natural conditions" for minimization of a functional. A similar case has already been treated in [12], where the identification of the boundary-layer thickness was done by means of MVI.

2. PROBLEM OF COEFFICIENT IDENTIFICATION

Consider the (1+1)-D equation of heat conduction

$$\mathcal{A}u \equiv -\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left[\lambda(x)\frac{\partial u}{\partial x}\right] = 0, \tag{2.1}$$

in the domain, shown in Fig. 1. The initial and boundary conditions are

$$u\Big|_{t=0} = u_0(x), \tag{2.2}$$

$$u(t,0) = f(t), \quad u(t,l) = g(t), \tag{2.3}$$

that are to match continuously, i.e.

$$f(0) = u_0(0), \quad g(0) = u_0(l). \tag{2.4}$$

The initial-boundary value problem (2.1)–(2.3) is correctly posed for the temperature $u(t,x)$, provided that the heat-conduction coefficient $\lambda(x)$ is a known positive function.

Suppose that the coefficient λ is unknown. In order to identify it, one needs more information. We consider here the case when a "terminal" condition is known:

$$u\Big|_{t=T} = u_1(x). \tag{2.5}$$

There can be different sources of such an information, e.g. the temperature in some interior point(s) as function of time, fluxes at the boundaries, etc. In the recent authors' work [9] we consider the case when the heat fluxes at boundaries are known functions of time, namely,

$$\lambda(0)\left.\frac{\partial u}{\partial x}\right|_{x=0} = \psi(t), \quad \lambda(l)\left.\frac{\partial u}{\partial x}\right|_{x=l} = \phi(t). \tag{2.6}$$

The goal of the present work is to show that the number of boundary conditions can be decreased as compared to (2.6). More precisely, we shall consider here the problem when only the values of the unknown coefficient $\lambda(x)$:

$$\lambda(0) = \lambda^0, \quad \lambda(l) = \lambda^l, \tag{2.7}$$

are prescribed in the boundary points.

3. METHOD OF VARIATIONAL IMBEDDING

We replace the original problem (2.1) by the problem of minimization of the following functional:

$$\mathcal{I} = \int_0^T \int_0^l [\mathcal{A}u]^2 \, \mathrm{d}x \, \mathrm{d}t \equiv \int_0^T \int_0^l \left[-\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\partial \lambda}{\partial x} + \lambda(x)\frac{\partial^2 u}{\partial x^2} \right]^2 \mathrm{d}x \, \mathrm{d}t = \min, \tag{3.1}$$

where u must satisfy the conditions (2.2), (2.3). The functional $\mathcal{I}$ is a quadratic and homogeneous function of $\mathcal{A}u$ and hence it attains its minimum if and only if $\mathcal{A}u \equiv 0$. In this sense there is one-to-one correspondence between the original equation (2.1) and the minimization problem (3.1).

The necessary conditions for minimization of (3.1) are the Euler-Lagrange equations for the functions $u(t,x)$ and $\lambda(x)$. The equation for u reads

$$-\frac{\partial^2 u}{\partial t^2} + \frac{\partial}{\partial x}\lambda(x)\frac{\partial^2}{\partial x^2}\lambda(x)\frac{\partial u}{\partial x} = 0. \tag{3.2}$$

This is an elliptic equation of second order with respect to time and hence it requires two conditions at the two ends of the time interval under consideration. These are the initial condition (2.2) at $t = 0$ and the "terminal" condition (2.5) at $t = T$. It is of fourth order with respect to the spatial variable x and its solution must satisfy the four conditions at the spatial boundaries — the original boundary conditions (2.3) and the so-called *natural conditions* for minimization of the functional $\mathcal{I}$:

$$\mathcal{A}u \equiv -\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left[\lambda(x)\frac{\partial u}{\partial x}\right] = 0 \ \text{ for } \ x = 0, l. \tag{3.3}$$

The problem is coupled by the Euler-Lagrange equation for λ, namely (see [4]):

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x)\frac{\mathrm{d}\lambda}{\mathrm{d}x} + G(x)\lambda = K(x)\,, \tag{3.4}$$

where

$$F(x) \equiv \int_0^T u_x^2 \, \mathrm{d}t\,, \quad G(x) \equiv \int_0^T u_x u_{xxx} \, \mathrm{d}t, \quad K(x) \equiv \int_0^T u_{tx} u_x \, \mathrm{d}t, \tag{3.5}$$

with the boundary conditions (2.7).

4. DIFFERENCE SCHEME

4.1. GRID PATTERN AND APPROXIMATIONS

In order to get second-order approximations of the boundary conditions, we employ a staggered mesh in the spatial direction, while the mesh in the temporal direction is standard (see Fig. 2). For the grid spacings we have $h = l/(N-3)$, $\tau = T/(M-1)$, where N is the total number of grid lines in the spatial direction, M — in the temporal direction, and the grid lines are defined as follows:

$$x_j = (j-2)h, \quad j = 1,\dots,N; \quad t_i = (i-1)\tau, \quad i = 1,\dots,M, \tag{4.1}$$

We employ symmetric central differences for the operators

$$\Lambda_{xx}u_{i,j} \overset{\text{def}}{=} \frac{\lambda_{j-1}}{h^2}u_{i,j-1} - \frac{\lambda_{j-1}+\lambda_j}{h^2}u_{i,j} + \frac{\lambda_j}{h^2}u_{i,j+1}$$
$$= \frac{\partial}{\partial x}\lambda(x)\frac{\partial}{\partial x}u(t,x) + O(h^2), \tag{4.2}$$

$$\Lambda_{xxxx}u_{i,j} \overset{\text{def}}{=} \frac{\lambda_{j-2}\lambda_{j-1}}{h^4}u_{i,j-2} - \frac{(\lambda_{j-2}+2\lambda_{j-1}+\lambda_j)\lambda_{j-1}}{h^4}u_{i,j-1}$$
$$+ \frac{(\lambda_{j-1}+\lambda_j)^2+\lambda_{j-1}^2+\lambda_j^2}{h^4}u_{i,j} - \frac{(\lambda_{j-1}+2\lambda_j+\lambda_{j+1})\lambda_{j+1}}{h^4}u_{i,j+1} + \frac{\lambda_{j+1}\lambda_j}{h^4}u_{i,j+2}$$
$$= \frac{\partial}{\partial x}\lambda(x)\frac{\partial^2}{\partial x^2}\lambda(x)\frac{\partial}{\partial x}u(t,x) + O(h^2)\ , \tag{4.3}$$

where $u_{i,j} = u(t_i, x_j)$ and $\lambda_j = \lambda(x_j + h/2)$.

The integrals, entering the equation for the diffusion coefficient, are approximated to the second order of accuracy as follows:

$$F_j \overset{\text{def}}{=} \tau\left[\frac{1}{2}\left(\frac{u_{1,j+2}-u_{1,j}}{2h}\right)^2 + \frac{1}{2}\left(\frac{u_{M,j+2}-u_{M,j}}{2h}\right)^2 + \sum_{i=2}^{M-1}\left(\frac{u_{i,j+2}-u_{i,j}}{2h}\right)^2\right]$$
$$= \int_0^T (u_x)^2 + O(\tau^2+h^2), \quad j = 1,2,\dots,N-2; \tag{4.4}$$

$$G_j \overset{\text{def}}{=} \tau\left[\frac{1}{2}\left(\frac{u_{1,j+1}-u_{1,j}}{h}\right)\left(\frac{u_{1,j+2}-3u_{1,j+1}+3u_{1,j}-u_{1,j-1}}{h^3}\right)\right.$$
$$+\frac{1}{2}\left(\frac{u_{M,j+1}-u_{M,j}}{h}\right)\left(\frac{u_{M,j+2}-3u_{M,j+1}+3u_{M,j}-u_{M,j-1}}{h^3}\right)$$
$$\left.+\sum_{i=2}^{M-1}\left(\frac{u_{i,j+1}-u_{i,j}}{h}\right)\left(\frac{u_{i,j+2}-3u_{i,j+1}+3u_{i,j}-u_{i,j-1}}{h^3}\right)\right]$$
$$= \int_0^T u_x u_{xxx} + O(\tau^2+h^2), \quad j = 2,3,\dots,N-2; \tag{4.5}$$

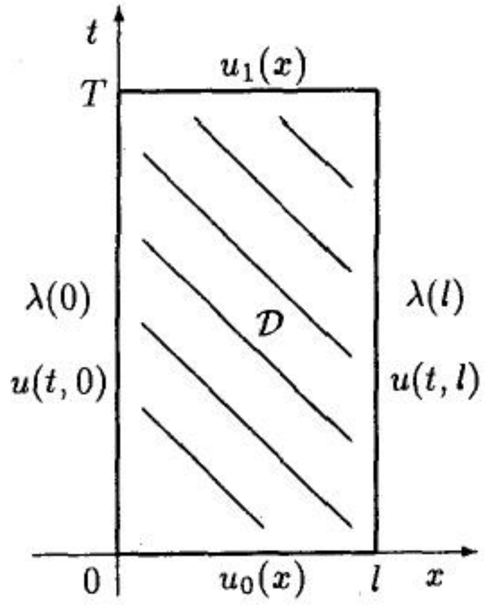

Fig. 1. Sketch of the domain

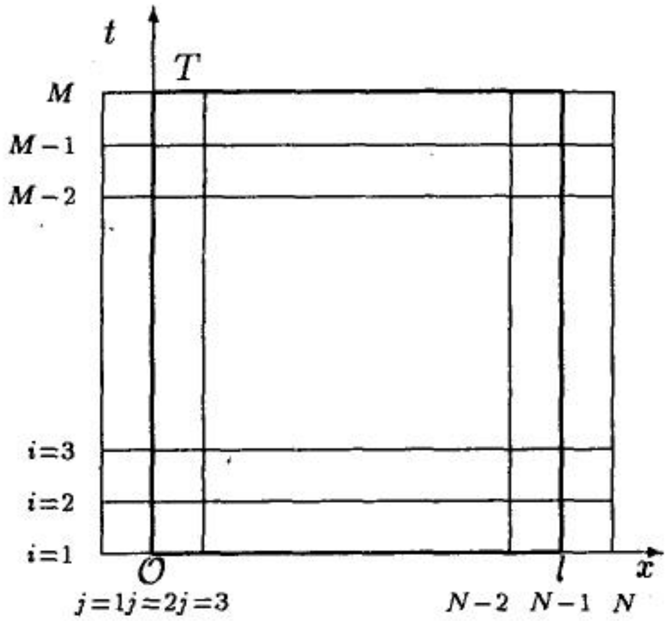

Fig. 2. Grid pattern

$$K_j \overset{\text{def}}{=} \frac{1}{2}\left[\left(\frac{u_{M,j+1}-u_{M,j}}{h}\right)^2 - \left(\frac{u_{1,j+1}-u_{1,j}}{h}\right)^2\right], \quad j=2,3,\ldots,N-2. \quad (4.6)$$

4.2. THE SCHEME FOR THE "DIRECT" PROBLEM

In order to gather "experimental" data for the "terminal" condition (2.5), we solve numerically the "direct" initial-boundary value problem (2.1)–(2.3). To this end we use a two-layer (Crank-Nicolson type) implicit difference scheme with second order of approximation in time and space, namely,

$$\frac{u_{i+1,j}-u_{i,j}}{\tau} = \frac{1}{2}\left(\Lambda_{xx}u_{i+1,j} + \Lambda_{xx}u_{i,j}\right) \quad (4.7)$$

for $i=1,\ldots,M-1$ and $j=2,\ldots,N-1$. The algebraic problem is coupled with the difference approximations of the initial and boundary conditions

$$u_{1,j} = u_0(x_j), \quad u_{i+1,2} = f(t_{i+1}), \quad u_{i+1,N-1} = g(t_{i+1}). \quad (4.8)$$

4.3. THE SPLITTING SCHEME FOR THE FOURTH-ORDER ELLIPTIC EQUATION

The particular choice of scheme for the fourth-order equation is not essential for the purposes of the present work. We use the iterative procedure based on the coordinate-splitting method because of its computational efficiency. The most straightforward approximation is the following:

$$-\frac{1}{\tau^2}\left(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right) + \Lambda_{xxxx}u_{i,j} = 0. \quad (4.9)$$

Upon introducing a fictitious time, the equation (4.9) adopts the form of a parabolic difference equation for which the implicit time stepping reads

$$\frac{u_{i,j}^{n+1}-u_{i,j}^{n}}{\sigma}=\Lambda_{tt}u_{i,j}^{n}-\Lambda_{xxxx}u_{i,j}^{n}, \tag{4.10}$$

where the notation Λ_{tt} stands for the second time difference, which enters (4.9). Then the splitting is enacted as follows:

$$\frac{\tilde{u}_{i,j}-u_{i,j}^{n}}{\sigma}=\Lambda_{tt}\tilde{u}_{i,j}-\Lambda_{xxxx}u_{i,j}^{n}, \quad \frac{u_{i,j}^{n+1}-\tilde{u}_{i,j}}{\sigma}=-\Lambda_{xxxx}\left[u_{i,j}^{n+1}-u_{i,j}^{n}\right], \tag{4.11}$$

where $\tilde{u}_{i,j}$ is called "half-time-step variable." The latter can be readily excluded, which yields the following $O(\sigma^2)$ approximation of (4.10):

$$B\frac{u_{i,j}^{n+1}-u_{i,j}^{n}}{\sigma}=\Lambda_{tt}u_{i,j}^{n}-\Lambda_{xxxx}u_{i,j}^{n}, \tag{4.12}$$

where $B=(E-\sigma^2\Lambda_{tt}\Lambda_{xxxx})$ is an operator whose norm is always greater than one. This means that the splitting scheme is even more stable than the general implicit scheme (4.10).

4.4. THE SCHEME FOR THE COEFFICIENT

If the solution $u_{i,j}$ of the imbedding problem is assumed known, then the coefficient can be computed on the base of the following second order scheme of approximation:

$$\frac{1}{h^2}\left[F_j\lambda_{j+1}-(F_j+F_{j-1})\lambda_j+F_{j-1}\lambda_{j-1}\right]+G_j\lambda_j=K_j, \tag{4.13}$$

where F_j, G_j and K_j, are defined in (4.4), (4.5) and (4.6), respectively.

4.5. GENERAL CONSEQUENCES OF THE ALGORITHM

(I) With given $\lambda(x)$, $u_0(x)$, $f(t)$ and $g(t)$, the "direct" problem (4.7), (4.8) is solved.

(II) With the obtained in (I) "experimentally observed" values of the $u_1(x)$, the fourth-order boundary value problem (4.11) is solved for the function u. The iterations with respect to the fictitious time are terminated when

$$\max_{i,j}\left|(u_{i,j}^{n+1}-u_{i,j}^{n})/u_{i,j}^{n}\right|<\varepsilon.$$

(III) The current iteration for the function $\lambda(x)$ is calculated from (4.13). If the difference between the new and the old $\lambda(x)$ is less than ε, then the calculations are terminated, otherwise one returns to (II).

5. NUMERICAL EXPERIMENTS

The first numerical experiment was to verify that the fourth-order elliptic problem for a given coefficient and consistent boundary data has the same solution as the "direct problem." We found that the iterative solution of the fourth-order problem does not depend on the magnitude of the increment σ of the artificial time. The optimal value turned out to be $\sigma = 0.05$. After the convergence of the "inner" iteration of the coordinate-splitting scheme, the obtained solution coincided with the "direct" solution within the truncation error of the scheme.

The second numerical experiment was to verify the approximation of the scheme for identification of the coefficient, with the field u considered as known from the solution of the "direct" problem. Once again the result was in a very good agreement within the truncation error.

Then the global iterative process can be started. The convergence of the "global" iterations does not necessarily follow from the correctness of the above discussed intermediate steps. For boundary data, which are not self-consistent, the "global" iteration can converge to a solution which has little in common with a solution of the heat-conduction equation.

To illustrate the numerical implementation of MVI, we solved the "direct" problem for a given diffusion coefficient and thus we obtained the self-consistent "experimental" over-posed terminal profile (2.5) at $t = T$.

The accuracy of the developed here difference scheme and algorithm were checked with the mandatory tests involving different grid spacing τ and h and different increments of the artificial time σ. We conducted a number of calculations with different values of mesh parameters and verified the practical convergence and the $O(\tau^2 + h^2)$ approximation of the difference scheme. The results confirmed the full approximation of the scheme (no dependence on σ) and the $O(h^2 + \tau^2)$ approximation.

To illustrate the accuracy and efficiency of the scheme, we considered the heat-conduction coefficient

$$\lambda(x) = x^2 + 1, \tag{5.1}$$

whose profile is shown in Fig. 3a. For smaller τ and h the differences are graphically indistinguishable. In Fig. 3b the ratio of the identified and "true" coefficient is shown, i.e.

$$r = \frac{\lambda_{\text{identified}}}{\lambda_{\text{true}}} \tag{5.2}$$

for different grids: $h = \tau = 1/64, 1/128, 1/256$.

A very serious test for the algorithm was the identification of a broken heat-conduction coefficient

$$\lambda(x) = \begin{cases} c_1 = \text{const} = 1 & \text{for} \quad 0 < x < 0.3, \\ c_2 = \text{const} = 1.1 & \text{for} \quad 0.3 < x < 0.7, \\ c_1 = \text{const} = 1 & \text{for} \quad 0.7 < x < 1. \end{cases} \tag{5.3}$$

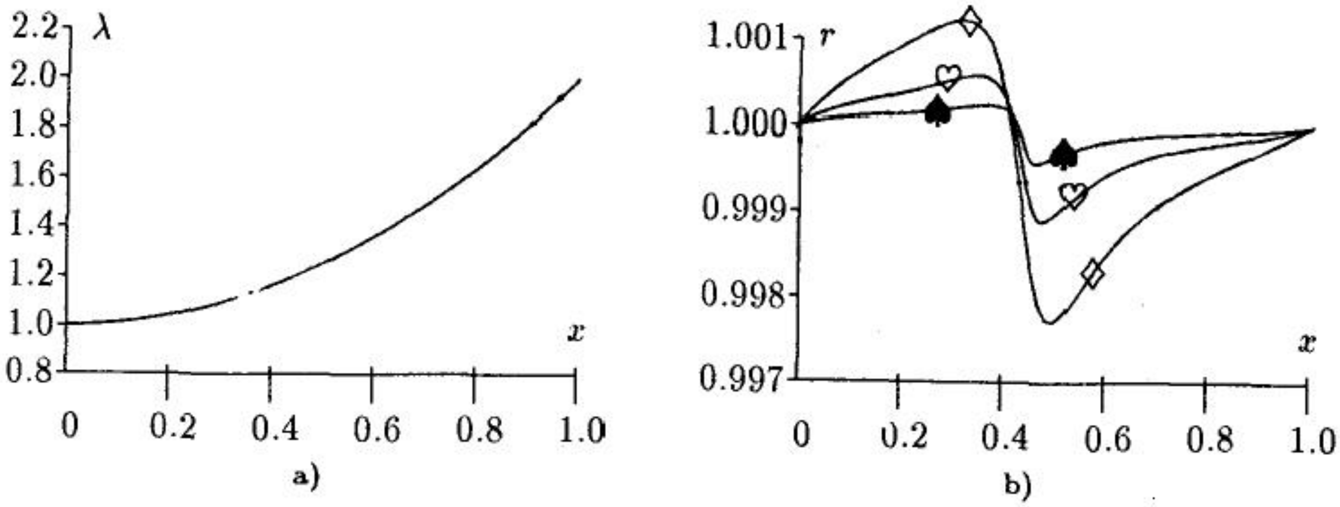

Fig. 3. Results of identification with $T = 1$, $l = 1$, $\varepsilon = 5 \cdot 10^{-8}$ for three different grid steps: a) the identified shape of the coefficient $\lambda(x)$; b) the ratio between the identified and the true coefficient: ◇ — $h = \tau = 1/64$, ♡ — $h = \tau = 1/128$, ♠ — $h = \tau = 1/256$

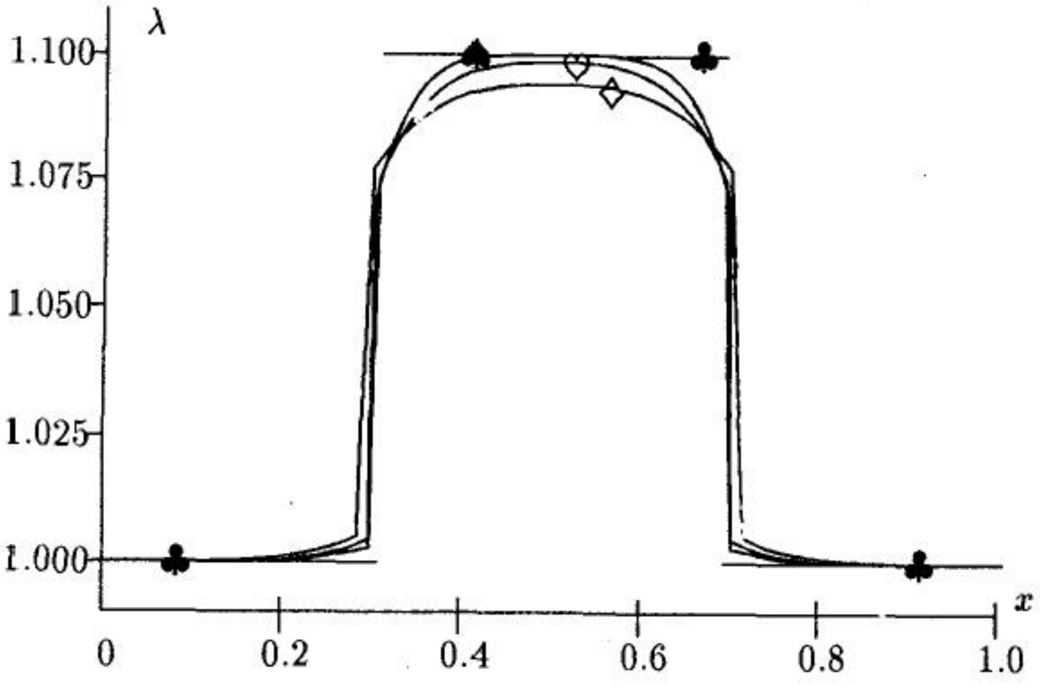

Fig. 4. Shapes of the identified and the true coefficient (5.3) for three different grid steps: ♣ – the true coefficient, ◇ – $\tau = 1/128$, $h = 1/64$, ♡ – $\tau = 1/256$, $h = 1/128$, ♠ – $\tau = 1/512$, $h = 1/256$

In Fig. 4 the shape of the "true" coefficient and the shapes of the three identified with different mesh-spaces coefficients are shown. The values of these coefficients are $\tau = 1/128$, $h = 1/64$, $\tau = 1/256$, $h = 1/128$ and $\tau = 1/512$, $h = 1/256$, respectively.

In Fig. 5 the ratios of the identified and "true" coefficient are shown.

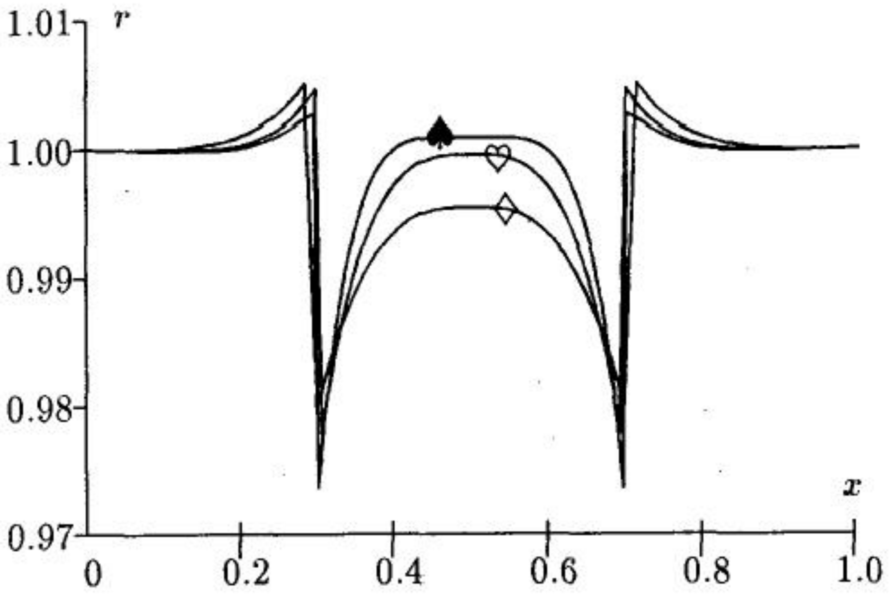

Fig. 5. Ratio between the identified and the true coefficient (5.3) for three different grid steps: ◇ – $\tau = 1/128, h = 1/64$, ♡ – $\tau = 1/256, h = 1/128$, ♠ – $\tau = 1/512, h = 1/256$

6. CONCLUSIONS

In the present paper we have displayed the performance of technique called Method of Variational Imbedding (MVI) for solving the inverse problem of coefficient identification in parabolic equation from over-posed data. The original inverse problem is replaced by the minimization problem for the quadratic functional of the original equation. The Euler-Lagrange equations for minimization comprise a fourth-order in space and second-order in time elliptic equation for the temperature and a second-order in space equation for the unknown coefficient. For this system the boundary data is not over-posed. It is shown that the solution of the original inverse problem is among the solutions of the variational problem, i.e. the inverse problem is imbedded into a higher-order but well posed elliptic boundary value problem ("imbedding problem"). In the present work we show that the number of boundary conditions can be decreased replacing them with the so-called "natural conditions" for minimization of a functional. Featuring examples are elaborated numerically with two different coefficients through solving the direct problem with a given coefficient and preparing the over-posed boundary data for the imbedding problem. The numerical results confirm that the solution of the imbedding problem coincides with the direct simulation of the original problem within the truncation error $O(\tau^2 + h^2)$.

ACKNOWLEDGEMENTS. The support of the Bulgarian Ministry of Education, Science and Technology under Grant No MM 610-96 is gratefully acknowledged.

REFERENCES

1. Bellomo, N., L. Preziosi. Modelling Mathematical Methods and Scientific Computation, CRC Press, 1995.
2. Christov, C. I. A method for identification of homoclinic trajectories. In: *Proc. 14-th Spring Conf. Union of Bulg. Mathematicians*, Sofia, Bulgaria, 1985, 571–577.
3. Christov, C. I. Method of variational imbedding for elliptic incorrect problems. *Comp. Rend. Acad. Buld. Sci.*, **39**(12), 1986, 23–26.
4. Christov, C. I. The method of variational imbedding for parabolic incorrect problems of coefficient identification. *Comp. Rend. Acad. Buld. Sci.*, **40**(2), 1987, 21–24.
5. Christov, C. I. The method of variational imbedding for time reversal incorrect parabolic problems. *Comp. Rend. Acad. Buld. Sci.*, **40**(6), 1987, 5–8.
6. Christov, C. I. Numerical implementation of the method of variational imbedding for coefficient identification for heat conduction equation. 1989 (unpublished manuscript).
7. Christov, C. I. Localized solutions for fluid interfaces via method of variational imbedding. In: *Fluid Physics*, M. G. Velarde and C. I. Christov, eds., World Scientific, 1995.
8. Christov, C. I., M. G. Velarde. On localized solutions of an equation governing Bénard-Marangoni convection. *Appl. Math. Modelling*, **17**, 1993, 311–320.
9. Christov, C. I., T. T. Marinov. Identification of heat-conduction coefficient via method of variational imbedding. *Math. Computer Modelling*, 1996 (in press).
10. Hadamard, J. Le probleme de Cauchy et les equations aux derivatives partielles lineares hyperboliques. Hermann, Paris, 1932.
11. Lattés, R., J.-L. Lions. Méthode de quasi-reversibilité et applicationes. Dunod, Paris, 1967.
12. Marinov, T. T., C. I. Christov. Boundary layer thickness as an inverse problem of coefficient identification. In: *Proc. HADMAR'91*, vol. 2, paper 57, Bulgarian Institute of Ship Hydrodynamics, Varna, 1991.
13. Christov, C. I., T. T. Marinov. Method of variational imbedding for the inverse problem of boundary-layer-thickness identification. *Math. Methods and Models in Applied Sciences*, **7**, 1997, 1005-1022.
14. Tikhonov, A. N. On the regularization of some ill-posed problems. *Doklady AN USSR*, **151**, 1963, 501–504; **153**, 1963, 49–52; **162**, 1965, 763–765 (in Russian).
15. Tikhonov, A. N., V. Ya. Arsenin. Methods for solving incorrect problems. Nauka, Moscow, 1974 (in Russian).

Received May 8, 1998

Christo I. Christov
National Institute of Meteorology and Hydrology
Bulgarian Academy of Sciences
BG-1184 Sofia, Bulgaria
E-mail: `christo.christov@meteo.bg`

Tchavdar T. Marinov
Department of Mathematics
Technical University of Varna
BG-9010 Varna, Bulgaria
E-mail: `marinov@ms3.tu-varna.acad.bg`