Rostocker Mathematisches Kolloquium

Heft 71

Universität Rostock

Institut für Mathematik

2017/18

Herausgeber: Universität Rostock
Mathematisch – Naturwissenschaftliche
Fakultät
Institut für Mathematik

Zitat–Kurztitel: Rostock. Math. Kolloq. **71** (2017/18)

ISSN 0138-3248

BEZUGSMÖGLICHKEITEN: Universität Rostock
Universitätsbibliothek, Schriftentausch
18051 Rostock
Tel.: +49-381-498 86 37
Fax: +49-381-498 86 32
e-mail: maria.schumacher@uni-rostock.de

Universität Rostock
Institut für Mathematik
18051 Rostock
Tel.: +49-381-498 6551
Fax: +49-381-498 6553
e-mail: romako@uni-rostock.de

DRUCK: Westarp & Partner Digitaldruck,
Hohenwarsleben

Rostock. Math. Kolloq. **71**, 3 – 13 (2017/18)

Subject Classification (AMS)
35J57, 35B50, 35B09

Bénédicte Alziary, Jacqueline Fleckinger

Sign of the solution to a non-cooperative system

ABSTRACT. Combining the results of a recent paper by Fleckinger-Hernández-de Thélin [14] for a non cooperative 2×2 system with the method of PhD Thesis of M. H. Lécureux we compute the sign of the solutions of a $n \times n$ non-cooperative systems when the parameter varies near the lowest principal eigenvalue of the system.

KEY WORDS. Maximum Principle, Antimaximum Principle, Elliptic Equations and Systems, Non Cooperative Systems, Principal Eigenvalue.

1 Introduction

Many results have been obtained since decades on Maximum Principle and Antimaximum principle for second order elliptic partial differential equations involving *e.g.* Laplacian, p-Laplacian, Schrödinger operator, ... or weighted equations. Then most of these results have been extended to systems.

The maximum principle (studied since centuries) has many applications in various domains as physic, chemistry, biology, ... Usually it shows that for positive data the solutions are positive (positivity is preserved). It is generally valid for a parameter below the "principal" eigenvalue (the smallest one). The Antimaximum principle, introduced in 1979 by Clément and Peletier ([8]), shows that, for one equation, as this parameter goes through this principal eigenvalue, the sign are reversed; this holds only for a small interval. The original proof relies on a decomposition into the groundstate (principal eigenfunction of the operator) and its orthogonal. It is the same idea which has been used in [14] (combined with a bootstrap method) to derive a precise estimate for the validity interval of the Antimaximum principle for one equation. By use of this result, Fleckinger-Hernández-de Thélin ([14]) deduce results on the sign of solution for some 2×2 non-cooperative systems. Indeed many papers have appeared for cooperative systems involving various elliptic operators: ([1], [2], [4], [9], [10],

[11], [12], [13], ...). Concerning non cooperative systems the literature is more restricted ([7], [14], ...).

In this paper we extend the results obtained in [14], valid for 2×2 non-cooperative systems involving Dirichlet Laplacian, to $n \times n$ ones. Recall that a system is said to be "cooperative" if all the terms outside the diagonal of the associated square matrix are positive.

For this aim we combine the precise estimate for the validity interval of the antimaximum principle obtained in [14] with the method used in [15], [1] for systems.

In Section 2 we are concerned with one equation. We first recall the precise estimate for the validity interval for the antimaximum principle ([14]); then we give some related results used in the study of systems.

In Section 3 we first state our main results for a $n \times n$ system (eventually non-cooperative) and then we prove them.

Finally, in Section 4, we compare our results with the ones of [14]. Our method, which uses the matricial calculus and in particular Jordan decomposition, allows us to have a more general point of view, even for a 2×2 system.

2 Results for one equation:

In [14], the authors consider a non-cooperative 2×2 system with constant coefficients. Before studying the system they consider one equation and establish a precise estimate of the validity interval for the antimaximum principle. We recall this result that we use later.

2.1 A precise Antimaximum for the equation [14]

Let Ω be a smooth bounded domain in $I\!R^N$. Consider the following Dirichlet boundary value problem

$$-\Delta z = \sigma z + h \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega, \tag{2.1}$$

where σ is a real parameter.

The associated eigenvalue problem is

$$-\Delta \phi = \lambda \phi \text{ in } \Omega, \quad \phi = 0 \text{ on } \partial\Omega. \tag{2.2}$$

As usual, denote by $0 < \lambda_1 < \lambda_2 \leq \ldots$ the eigenvalues of the Dirichlet Laplacian defined on Ω and by ϕ_k a set of orthonormal associated eigenfunctions, with $\phi_1 > 0$.

Hypothesis 1 *Assume $h \in L^q$, $q > N$ if $N \geq 2$ and $q = 2$ if $N = 1$.*

Hypothesis 2 *Assume* $h^1 := \int h\phi_1 > 0$.

Writing

$$h = h^1\phi_1 + h^\perp \tag{2.3}$$

where $\int_\Omega h^\perp\phi_1 = 0$ one has:

Lemma 2.1 [14] *We assume* $\lambda_1 < \sigma \leq \Lambda < \lambda_2$ *and* $h \in L^q$, $q > N \geq 2$. *We suppose that there exists a constant* C_1 *depending only on* Ω, q, *and* Λ *such that* z *satisfying* (2.1) *is such that*

$$\|z\|_{L^2} \leq C_1\|h\|_{L^2}. \tag{2.4}$$

Then there exist constants C_2 *and* C_3*, depending only on* Ω, q *and* Λ *such that*

$$\|z\|_{C^1} \leq C_2\|h\|_{L^q} \text{ and } \|z\|_{L^q} \leq C_3\|h\|_{L^q}. \tag{2.5}$$

Remark 2.1 The same result holds for $\Lambda < \sigma < \lambda_1$ where Λ is any given constant $< \lambda_1$, with the same proof.

Remark 2.2 Inequality (2.4) cannot hold, for all $\lambda_1 < \sigma \leq \Lambda$, unless h is orthogonal to ϕ_1.

Theorem 1 [14]*: Assume Hypotheses* 1 *and* 2*; fix* Λ *such that* $\lambda_1 < \sigma \leq \Lambda < \lambda_2$. *There exists a constant* K *depending only on* Ω, Λ *and* q *such that, for* $\lambda_1 < \sigma < \lambda_1 + \delta(h)$ *with*

$$\delta(h) = \frac{Kh^1}{\|h^\perp\|_{L^q}}, \tag{2.6}$$

the solution z *to* (2.1) *satisfies the antimaximum principle, that is*

$$z < 0 \text{ in } \Omega; \quad \partial z/\partial\nu > 0 \text{ on } \partial\Omega, \tag{2.7}$$

where $\partial/\partial\nu$ *denotes the outward normal derivative.*

2.2 Other remarks for one equation

Consider again Equation (2.1). For $\sigma \neq \lambda_k$, z solution to (2.1) is

$$z \;=\; z^1\phi_1 + z^\perp \;=\; \frac{h^1}{\lambda_1 - \sigma}\phi_1 \;+\; z^\perp, \tag{2.8}$$

with $z^\perp$ satisfying

$$-\Delta z^\perp \;=\; \sigma z^\perp + h^\perp \;\text{ in } \Omega\;; \quad z^\perp = 0 \text{ on } \partial\Omega. \tag{2.9}$$

In the next section, our proofs will use the following result.

Lemma 2.2 *We assume Hypothesis 1 and $\sigma < \lambda_1$. Then $z^\perp$ (and its first derivatives) is bounded: There exits a positive constant C_0, independent of σ such that*

$$\|z^\perp\|_{\mathcal{C}^1} \leq C_0 \|h\|_{L^q}. \tag{2.10}$$

Moreover, if $\sigma < \Lambda < \lambda_1$, where Λ is some given constant $< \lambda_1$, z is bounded and there exits a positive constant C_0', independent of σ such that

$$\|z\|_{\mathcal{C}^1} \leq C_0' \|h\|_{L^q}. \tag{2.11}$$

Proof: This is a simple consequence of the variational characterization of λ_2:

$$\lambda_2 \int_\Omega |z^\perp|^2 \leq \int_\Omega |\nabla z^\perp|^2 = \sigma \int_\Omega |z^\perp|^2 + \int_\Omega z^\perp h^\perp \leq \lambda_1 \int_\Omega |z^\perp|^2 + \int_\Omega z^\perp h^\perp.$$

By Cauchy-Schwarz we deduce

$$\|z^\perp\|_{L^2} \leq \frac{1}{\lambda_2 - \lambda_1} \|h^\perp\|_{L^2}. \tag{2.12}$$

This does not depend on $\sigma < \lambda_1$.

Then one can deduce (2.10), that is $z^\perp$ (and its derivatives) is bounded. This can be found *e.g.* in [6] (for $\sigma < \lambda_1$ and $\lambda_1 - \sigma$ small enough) or it can be derived exactly as in [14] (where the case $\sigma > \lambda_1$ and $\sigma - \lambda_1$ small enough is considered).

Finally we write $z = z_1 \phi_1 + z^\perp$ and deduce (2.11).

Remark 2.3 Note that in (2.8), since $h^1 > 0$, $\frac{h^1}{\lambda_1 - \sigma} \to +\infty$ as $\sigma \to \lambda_1$, $\sigma < \lambda_1$.

3 Results for a $n \times n$ system:

We consider now a $n \times n$ (eventually non-cooperative) system defined on Ω a smooth bounded domain in $I\!R^N$:

$$-\Delta U = AU + \mu U + F \text{ in } \Omega\,,\; U = 0 \text{ on } \partial\Omega, \tag{S}$$

where F is a column vector with components f_i, $1 \leq i \leq n$. Matrix A is not necessarily cooperative, that means that its terms outside the diagonal are not necessarily positive. First we introduce some notations concerning matrices. Then, with these notations we can state our results and prove them.

3.1 The matrix of the system and and the eigenvalues

Hypothesis 3 *A is a $n \times n$ matrix which has constant coefficients and has only real eigenvalues. Moreover, the largest one which is denoted by ξ_1 is positive and algebrically and geometrically simple. The associated eigenvectors X_1 has only non zero components.*

Of course some of the other eigenvalues can be equal. Therefore we write them in decreasing order

$$\xi_1 > \xi_2 \geq \ldots \geq \xi_n. \tag{3.13}$$

The eigenvalues of $A = (a_{ij})_{1 \leq i,j \leq n}$, denoted , ξ_1, ξ_2,. . . , ξ_n , are the roots of the associated characteristic polynomial

$$p_A(\xi) = det(\xi I_n - A) = \prod(\xi - \xi_k), \tag{3.14}$$

where I_n is the $n \times n$ identity matrix.

Remark 3.1 By above, $\xi > \xi_1 \Rightarrow p_A(\xi) > 0$.

Denote by $X_1 \ldots X_n$ the eigenvectors associated respectively to eigenvalue $\xi_1, \ldots, \xi_n$.

Jordan decomposition Matrix A can be expressed as $A = PJP^{-1}$, where $P = (p_{ij})$ is the change of basis matrix of A and J is the Jordan canonical form (lower triangular matrix) associated with A. The diagonal entries of J are the ordered eigenvalues of A and $p_A(\xi) = p_J(\xi)$.

Notation : In the following, set

$$U = P\tilde{U} \Leftrightarrow \tilde{U} = P^{-1}U, \; F = P\widetilde{F} \Leftrightarrow \tilde{F} = P^{-1}F. \tag{3.15}$$

Here $\widetilde{U}$ and $\widetilde{F}$ are column vectors with components $\widetilde{u}_i$ and $\widetilde{f}_i$.

Eigenvalues of the system: μ is an eigenvalue of the system if there exists a non zero solution U to

$$-\Delta U = AU + \mu U \text{ in}\,\Omega\,,\; U = 0 \text{ on } \partial\Omega. \tag{S_0}$$

We also say that μ is a "principal eigenvalue" of System (S) if it is an eigenvalue with components of the associated eigenvector which do not change sign. (Note that the components do not change sign but are not necessarily positive as claimed in [14]).

Then $\phi_j X_k$ is an eigenvector associated to eigenvalue

$$\mu_{jk} = \lambda_j - \xi_k. \tag{3.16}$$

3.2 Results for $|\mu - \mu_{11}| \to 0$

We study here the sign of the component of U as $\mu \to \mu_{11} = \lambda_1 - \xi_1$.

For this purpose we use the methods in [15] or [1] combined with [14]. Note that by (3.13), $\mu_{11} < \mu_{1k} = \lambda_1 - \xi_k$, for all $2 \le k \le n$.

Hypothesis 4 *F is with components $f_i \in L^q$, $q > N > 2$, $q = 2$ if $N = 1$, $1 \le i \le n$; moreover we assume that the first component $\tilde{f}_1$ of $\tilde{F} = P^{-1}F$ is $\geq 0, \not\equiv 0$.*

Theorem 2 *Assume Hypothesis 3 and 4. Assume also $\mu < \mu_{11}$. Then, there exists $\delta > 0$ independant of μ, such that for $\mu_{11} - \delta < \mu < \mu_{11}$, the components u_i of the solution U have the sign of p_{i1} and the outside normal derivatives $\frac{\partial u_i}{\partial \nu}$ have the sign of $-p_{i1}$.*

Theorem 3 *Assume Hypothesis 3 and 4 are satisfied; then, there exists $\delta > 0$ independant of μ such that for $\mu_{11} < \mu < \mu_{11} + \delta$ the components u_i of the solution U have the sign of $-p_{i1}$ and their outgoing normal derivatives have opposite sign.*

Remark 3.2 The results of Theorems 2 and 3 are still valid if we assume only $\int_\Omega \tilde{f}_1 \phi_1 > 0$ instead of $\tilde{f}_1 \geq 0 \not\equiv 0$.

3.3 Proofs

We start with the proof of Theorem 2 where $\mu < \mu_{11}$; assume Hypotheses 3 and 4.

3.3.1 Step 1: An equivalent system

We follow [15] or [1]. As above set $U = P\tilde{U}$ and $F = P\tilde{F}$.

Starting from

$$-\Delta U = AU + \mu U + F,$$

multiplying by P^{-1}, we obtain

$$-\Delta \tilde{U} = J\tilde{U} + \mu\tilde{U} + \tilde{F}.$$

Note that everywhere we have the homogeneous Dirichlet boundary conditions, but we do not write them for simplicity.

The Jordan matrix J has p Jordan blocks J_i $(1 \le i \le p \le n)$ which are $k_i \times k_i$ matrices of the form

$$J_i = \begin{pmatrix} \xi_i & 0 & \cdots & 0 \\ 1 & \xi_i & 0 & \cdots \\ \ddots & \ddots & \vdots & \\ 0 & \cdots 1 & \xi_i & 0 \\ 0 & \cdots & 1 & \xi_i \end{pmatrix}.$$

By Hypothesis 3, the first block is 1×1 : $J_1 = (\xi_1)$. Hence we obtain the first equation

$$-\Delta\widetilde{u_1} = \xi_1\widetilde{u_1} + \mu\widetilde{u_1} + \tilde{f}_1. \tag{3.17}$$

Since $\tilde{f}_1 \geq 0$, $\not\equiv 0$, $\xi_1 + \mu < \lambda_1$ and by Hypothesis 4, $\tilde{f}_1 \in L^2$, we have the maximum principle and

$$\widetilde{u_1} > 0 \ on \ \Omega. \ \frac{\widetilde{u_1}}{\partial\nu}|_{\partial\Omega} < 0. \tag{3.18}$$

Then we consider the second Jordan blocks J_2 which is a $k_2 \times k_2$ matrix with first line

$$\xi_2,\ 0,\ 0, \dots$$

The first equation of this second block is

$$-\Delta\tilde{u}_2 = \xi_2\tilde{u}_2 + \mu\tilde{u}_2 + \tilde{f}_2.$$

Since $\mu < \mu_{11} = \lambda_1 - \xi_1 < \lambda_1 - \xi_2 \leq \lambda_1 - \xi_k$, $k \geq 2$. Hence, by Lemma 2.2, $\widetilde{u_2}$ stays bounded as $\mu \to \mu_{11}$. and this holds for all the $\widetilde{u_k}$, $k > 1$. By induction $\widetilde{u_k}$ is bounded for all k.

3.3.2 Step 2: End of the proof of Theorem 2

Now we go back to the functions u_i: $U = P\tilde{U} = (u_i)$ implies that for each u_i, $1 \leq i \leq n$, we have

$$u_i = p_{i1}\widetilde{u_1} + \sum_{j=2}^{n} p_{ij}\widetilde{u_j}. \tag{3.19}$$

The last term in (3.19) stays bounded according to Lemma 2.2; indeed $\sum_{j=2}^{n} p_{ij}\widetilde{u_j}$ is bounded by a constant which does not depend on μ.

By Remark 2.3, $\widetilde{u_1} \to +\infty$ as $\mu \to \lambda_1 - \xi_1$. Hence, each u_i has the same sign than p_{i1} (the first coefficient of the $i-th$ line in matrix P which is also the i-th coefficient of the first eigenvector X_1) for $\lambda_1 - \xi_1 - \mu > 0$ small enough. Analogously, $\frac{\partial u_i}{\partial\nu}$ behaves as $p_{i1}\frac{\partial\tilde{u}_i}{\partial\nu}$ which has the sign of $-p_{i1}$.

It is noticeable that only $\widetilde{u_1}$ plays a role!! □

3.4 Proof of Theorem 3 ($\mu > \mu_{11}$)

Now $\mu_{11} < \mu < \mu_{11} + \epsilon$ where $\epsilon \leq min\{\xi_1 - \xi_2, \lambda_2 - \lambda_1\}$ and $f_i \in L^q, q > N$. We proceed as above but deduce immediately that for $\mu - \mu_{11}$ small enough ($\mu - \mu_{11} < \delta_1 := \delta(\widetilde{f_1}) < \frac{K\widetilde{f_1^1}}{\|f_1^\perp\|_{L^q}}$) defined in [14], Theorem 1), $\widetilde{u}_1 < 0$ by the antimaximum principle. From now on choose

$$\mu - \mu_{11} < \delta, \text{ with } \delta < min\{\epsilon, \delta_1\}. \tag{3.20}$$

For the other equations, by Lemma 2.1, $\widetilde{u}_k > 0$ is bounded as above.

We consider now U. We notice that $F = P\widetilde{F}$ which can also be written $f_i = \sum_{k=1}^n p_{ik}\widetilde{f}_k$ implies $f_i^\perp = \sum_{k=1}^n p_{ik}\widetilde{f}_k^\perp$. With the same argument as above, the components u_i of the solution U have the sign of $-p_{i1}$ for $\mu - \mu_{11}$ sufficiently small ($\mu - \mu_{11} < \delta$). The normal derivatives of the u_i are of opposite sign. □

4 Annex: The 2 × 2 non-cooperative system

We apply now our results to the 2 × 2 system, considered in [14]. Consider the 2 × 2 non-cooperative system depending on a real parameter μ

$$-\Delta U = AU + \mu U + F \text{ in}\,\Omega\,,\ U = 0 \text{ on } \partial\Omega, \tag{S}$$

which can also be written as

$$-\Delta u \ = \ au \ + \ bv \ + \ \mu u \ + \ f \ \text{ in } \Omega, \tag{S_1}$$

$$-\Delta v \ = \ cu \ + \ dv \ + \ \mu v \ + \ g \ \text{ in } \Omega, \tag{S_2}$$

$$u = v = 0 \text{ on } \partial\Omega. \tag{S_3}$$

Hypothesis 5 *Assume* $b > 0\,, c < 0,$ *and* $D := (a-d)^2 + 4bc > 0.$

Here System (S) has (at least) two principal eigenvalues μ_1^- and μ_1^+ where

$$\mu_1^- \ := \ \lambda_1 - \xi_1 \ < \ \mu_1^+ \ := \ \lambda_1 - \xi_2, \tag{4.21}$$

where ξ_1 and ξ_2. are the eigenvalues of Matrix A and we choose $\xi_1 > \xi_2$.

The main theorems in [14] are:

Theorem 4 ([14]) *Assume Hypothesis 5,* $\mu_1^- < \mu < \mu_1^+$ *and* $d < a$. *Assume also*

$$f \geq 0,\ g \geq 0,\ f, g \not\equiv 0, f, g \in L^q,\ q > N\,if\,N \geq 2\,;\ q = 2\,if\,N = 1.$$

Then there exists $\delta > 0$, *independent of* μ, *such that* $\mu < \mu_1^- + \delta$ *implies*

$$u < 0,\ v > 0\ in\ \Omega;\ \frac{\partial u}{\partial \nu} > 0,\ \frac{\partial v}{\partial \nu} < 0\ on\ \partial\Omega.$$

Theorem 5 ([14]) *Assume Hypothesis 5,* $\mu_1^- < \mu < \mu_1^+$ *and* $a < d$. *Assume also*

$$f \leq 0,\ g \geq 0,\ f, g \not\equiv 0\,, f, g \in L^q,\ q > N\,if\,N \geq 2\,;\ q = 2\,if\,N = 1.$$

Then there exists $\delta > 0$, *independent of* μ, *such that i* $\mu < \mu_1^- + \delta$ *implies*

$$u < 0,\ v < 0\ in\ \Omega;\ \frac{\partial u}{\partial \nu} > 0,\ \frac{\partial v}{\partial \nu} > 0\ on\ \partial\Omega.$$

Theorem 6 ([14]) *Assume Hypothesis 5 and $a < d$. Assume also that the parameter μ satisfies: $\mu < \mu_1^-$, and*

$$f \geq 0,\ g \geq 0,\ f, g \not\equiv 0,\ f, g \in L^2.$$

Assume also $t^*g - f \geq 0$, $t^*g - f \not\equiv 0$ with

$$t^* = \frac{d - a + \sqrt{D}}{-2c}.$$

Then

$$u > 0,\ v > 0 \text{ in } \Omega;\ \frac{\partial u}{\partial \nu} < 0,\ \frac{\partial v}{\partial \nu} < 0 \text{ on } \partial\Omega.$$

The matrix A is

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with eigenvalues $\xi_2 = \frac{a+d-\sqrt{D}}{2} < \xi_1 = \frac{a+d+\sqrt{D}}{2}$ where $D = (a-d)^2 + 4bc > 0$. The eigenvectors are

$$X_k = \begin{pmatrix} b \\ \xi_k - a \end{pmatrix},\quad P = \begin{pmatrix} b & b \\ \xi_1 - a & \xi_2 - a \end{pmatrix}.$$

Note that the characteristic polynomial is $\mathcal{P}(s) = (a - s)(d - s) - bc$. Since $\mathcal{P}(a) = \mathcal{P}(d) = -bc > 0$, a and d are outside $[\xi_2, \xi_1]$.

For $d > a$ both $p_{i1} > 0$ and for $d < a$ $p_{11} > 0$, $p_{21} < 0$.

$$P^{-1} = \frac{1}{b(\xi_1 - \xi_2)} \begin{pmatrix} a - \xi_2 & b \\ \xi_1 - a & -b \end{pmatrix}.$$

$$\tilde{f}_1 = \frac{1}{b(\xi_1 - \xi_2)}[(a - \xi_2)f + bg]. \tag{4.22}$$

In Theorem 2 of [14] $d < a$, $f, g \geq 0$ so that $\tilde{f}_1 > 0$ and u has the sign of $-p_{11} = -b < 0$; v has the sign of $-p_{21} = a - \xi_1 > 0$.

In Theorem 3 of [14] $d > a$, $f \leq 0$ and $g \geq 0$ implies $\tilde{f}_1 > 0$. So that u has the sign of $-p_{11} = -b < 0$; v has the sign of $-p_{12} = a - \xi_2 < 0$.

Finally the hypothesis $\tilde{f}_1 \geq 0$ is sufficient for having the sign of the solutions and the maximum principle holds (all $u_i > 0$) iff $p_{i1} > 0$.

Our results can conclude for other cases; *e.g*, as in Theorem 2, $d < a$, $f \geq 0$, but now $g < 0$ with $\tilde{f}_1 = \frac{1}{b(\xi_1 - \xi_2)}[(a - \xi_2)f + bg] > 0$.

Analogously, in Theorem 4, $f, g \geq 0$ and $\tilde{f}_1 > 0$ implies for having $u, v > 0$ that necessarily $\xi_2 - a > 0$ so that $a < d$. But again we can conclude for the sign in other cases (*e.g.* $a > d$) if only $\tilde{f}_1 > 0$, (which is precisely the added condition in Theorem 4). □

References

[1] **Alziary, B., Fleckinger, J., Lécureux, M. H.**, and **Wei, N. :** *Positivity and negativity of solutions to $n \times n$ weighted systems involving the Laplace operator defined on $\mathbb{R}^N$, $N \geq 3$.* Electron. J. Diff. Eqns, 101, 2012, p. 1 – 14

[2] **Alziary, B., Fleckinger, J.**, and **Takáč, P. :** *An extension of maximum and anti-maximum principles to a Schrödinger equation in $\mathbb{R}^N$.* Positivity, 5, (4), 2001, pp. 359 – 382

[3] **Amann, H. :** *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces.* SIAM Re. 18, 4, 1976, p. 620 – 709

[4] **Amann, H. :** *Maximum Principles and Principal Eigenvalues.* Ten Mathematical Essays on Approximation in Analysis and Topology, J. Ferrera, J. López-Gómez, F. R. Ruíz del Portal ed., Elsevier, 2005, 1 – 60

[5] **Arcoya, D.**, and **Gámez, J. :** *Bifurcation theory and related problems: anti-maximum principle and resonance.* Comm. Part. Diff. Equat., 26, 2001, p. 1879 – 1911

[6] **Brezis, H. :** *Analyse Fonctionnelle.* Masson, Paris, 1987

[7] **Caristi, G.**, and **Mitidieri, E. :** *Maximum principles for a class of non-cooperative elliptic systems.* Delft Progress Rep. 14, 1990, p. 33 – 56

[8] **Clément, P.**, and **Peletier, L. :** *An anti-maximum principle for second order elliptic operators.* J. Diff. Equ. 34, 1979, p. 218 – 229

[9] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *A Maximum Principle for an Elliptic System and Applications to semilinear Problems.* SIAM J. Math and Anal. N17, 1986, 836 – 849

[10] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *Maximum principles for cooperative elliptic systems.* C. R. Acad. Sci. Paris 310, 1990, p. 49 – 52

[11] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *Maximum principles for linear elliptic systems.* Quaterno Mat. 177, Trieste, 1988

[12] **Fleckinger, J., Gossez, J. P., Takáč, P.**, and **de Thélin, F. :** *Existence, nonexistence et principe de l'antimaximum pour le p-laplacien.* C. R. Acad. Sci. Paris 321, 1995, p. 731 – 734

[13] **Fleckinger, J., Hernández, J.**, and **de Thélin, F. :** *On maximum principles and existence of positive solutions for some cooperative elliptic systems.* Diff. Int. Eq. 8, 1, 1995, p. 69– 85

[14] **Fleckinger, J., Hernández, J.**, and **de Thélin, F. :** *Estimate of the validity interval for the Antimaximum Principle and application to a non-cooperative system.* Rostock Math. Kolloq. 69 (2014/15), p. 19 – 32

[15] **Lécureux, M. H. :** *Au-delà du principe du maximum pour des systèmes d'opérateurs elliptiques.* Thèse, Université de Toulouse, Toulouse 1, 13 juin 2008

[16] **Protter, M. H.**, and **Weinberger, H. :** *Maximum Principles in Differential Equations.* Springer-Verlag, 1984

received: April 4, 2016

Authors:

Bénédicte Alziary
TSE & IMT (UMR 5219) –
CEREMATH-UT1
Université de Toulouse,
31042 TOULOUSE Cedex,
France

e-mail: alziary@ut-capitole.fr

Jacqueline Fleckinger
IMT (UMR 5219) –
CEREMATH-UT1
Université de Toulouse,
31042 TOULOUSE Cedex,
France

e-mail: jfleckinger@gmail.com

Rostock. Math. Kolloq. **71**, 14 – 27 (2017/18)

Subject Classification (AMS)
03B05, 34D05, 34D20, 54C35, 54C50

Harry Poppe

Reference Stability for ODE

1 Introduction

We consider initial value problems for autonomous ODE, and we will study stability for these problems. The dignified definition of Ljapunow stability has two shortcomings. To overcome these difficulties we define the notion of reference stability. This notion has especially the advantage that we can characterize it topologically. We illustrate the procedere by simple examples.

2 Some simple but instructive examples

We consider the equations $\dot{x} = \pm x^n$, $n \geq 2$, $x(t_0) = x_0$. But these equations are autonomous and hence we let $t_0 = 0$.

These equations are of product type:

$$\dot{x} = g(t)h(x) = x^n \qquad (h(x) = \pm x^n)\,.$$

Since $h(x) = 0 \iff \pm x^n = 0 \iff x = 0$: $x \equiv 0$ is an equilibrium point of $\dot{x} = \pm x^n$, $x(0) = 0$. We will show that the zero solutions of our equations always are unique.

Proposition 2.1 *The zero solution of our equations always are unique.*

Proof: We have $h(x) = \pm x^n$, it is enough to consider $h(x) = x^n$.

Now

$$\int_y^1 \frac{1}{h(s)}\,ds = \int_y^1 s^{-n}\,ds = \left(\frac{1}{1-n}s^{1-n}\right)\Big|_y^1 = \frac{1}{1-n}\left(1 - \frac{1}{y^{n-1}}\right) \Longrightarrow$$

$$\left|\lim_{y\to 0}\left(\frac{1}{1-n}\right)\left(1 - \frac{1}{y^{n-1}}\right)\right| = +\infty\,,$$

since $n \geq 2$. Thus by a well-known criterion (see [4]) $x \equiv 0$ is a unique solution.

(a) $\dot{x} = x^n$; for $x \neq 0$, $x_0 \neq 0$ we find:

$$\begin{aligned} &\int_{x_0}^{x} s^{-n} ds = \int_0^t 1 ds = t\,, \\ &\frac{1}{1-n}\left(s^{1-u}\Big|_{x_0}^{x}\right) = t\,, \\ &x^{1-u} - x_0^{1-u} = (1-n)t \\ &x^{1-n} = x_0^{1-n} + (1-n)t \end{aligned} \tag{2.1}$$

(b) $\dot{x} = -x^n$; we get

$$\begin{aligned} x^{1-n} &= x_0^{1-n} + (1-n)(-t) \\ x^{1-n} &= x_0^{1-u} + (n-1)t \end{aligned} \tag{2.2}$$

Example 2.2 (a) $n = 3$: $\dot{x} = x^3$, $x(0) = x_0$:

$x^{1-3} = x^{-2} = x_0^{-2} - 2t = \dfrac{1}{x_0^2} - 2t = \dfrac{1-2x_0^2 t}{x_0^2} \Longrightarrow |x| = \dfrac{|x_0|}{\sqrt{1-2x_0^2 t}}$. We know that $x_0 \neq 0$ holds and hence x is defined on $\left(-\infty, \dfrac{1}{2x_0^2}\right)$.

We have two cases:

1. $x_0 > 0$, then by continuity and since $\dot{x} = x$ is autonomous: $\forall t \in \left(-\infty, \dfrac{1}{2x_0^2}\right)$: $x(t) > 0$, thus $|x| = x = \dfrac{x_0}{\sqrt{1-2x_0^2 t}}$.
2. $x_0 < 0$, by the same argument: $\forall t \in \left(-\infty, \dfrac{1}{1x_0^2}\right)$: $|x(t)| = -x(t) = \dfrac{-x_0}{\sqrt{1-2x_0^2 t}} \Longrightarrow x = x(t) = \dfrac{x_0}{\sqrt{1-2x_0^2 t}}$,

and here $x_0 < 0 \Longrightarrow x(t) < 0\, \forall t$.

Remarks 2.3 (a) Result concerning stability: $0 : \forall t \in [0, +\infty) : 0(t) = 0$ is defined on $[0, +\infty)$, but no other solution is defined on this interval.

(b) The sets of possible initial values of this equation are:

$$(0, +\infty), \quad (-\infty, 0)$$

(c) $\dfrac{1}{2x_0^2}$, $x_0 \neq 0$ is a pole:

$$\lim_{t \to \left(\frac{1}{2x_0^2}\right)} -\frac{x_0}{\sqrt{1-2x_0^2 t}} = \begin{cases} +\infty, & x_0 > 0 \\ -\infty, & x_0 < 0 \end{cases}$$

Example 2.4 (b) $n = 3$, $\dot{x} = -x^3$, $x(0) = x_0$, $x_0 \neq 0$. By (2.2) holds:

$$x^{1-n} = x_0^{1-n} + (n-1)t; n = 3 \Longrightarrow x^{-2} = x_0^{-2} + 2t,$$
$$x^{-2} = \frac{1 + 2x_0^2 t}{x_0^2} \Longrightarrow |x| = \frac{|x_0|}{\sqrt{1 + 2x_0^2 t}},$$

and $\forall x_0$, $x_0 \neq 0 : \forall t \in [0, +\infty) : 1 + 2x_0^2 t > 0$.

As above we get again

$$x_0 > 0 \Longrightarrow x = \frac{x_0}{\sqrt{1 + 2x_0^2 t}} > 0$$
$$x_0 < 0 \Longrightarrow x = \frac{x_0}{\sqrt{1 + 2x_0^2 t}} < 0$$

Remarks 2.5 1. Result concerning stability: $x \equiv 0$ is defined on $[0, +\infty)$ and all other solutions too.

2. The set of all possible initial values of this equation is $(-\infty, 0) \cup (0, +\infty)$.

More precisely we have

$$1 + 2x_0^2 t > 0 \Longleftrightarrow -\frac{1}{2x_0^2} < t, \text{ and } -\frac{1}{2x_0^2} < 0\,.$$

Remark 2.6 More example of ODE we consider to illustrate definitions or to apply the results of propositions.

3 Ljapunow – Stability

We consider an autonomous system of ordinary differential equations:

$$\dot{x} = f(x), \;\; f : G \to \mathbb{R}^n\,, \tag{3.1}$$

$G \subseteq \mathbb{R}^n$ is open and f is continuous. Let $t_0 \in \mathbb{R}$, $0 \leq t_0$ and let x be a solution, $x(t_0) = x_0 \in \mathbb{R}^n$ defined (at least) on $[t_0, +\infty)$.

We want to formulate that x is Ljapunow-stable in a precise way. But this is only possible, if we use the following definition:

Definition 3.1 *x is called Ljapunow-stable (L-stable) iff $\forall \varepsilon > 0\ \exists \delta > 0$, $\delta = \delta(\varepsilon)$, $0 < \delta \leq \varepsilon : \forall y$, where y solves the initial value problem $\dot{y} = f(y)$, $y(t_0) = y_0 \in G$ (on some intervall of $\mathbb{R}$):*

$$\|y_0 - x_0\| < \delta \Longrightarrow \; (y \textit{ is defined on } [t_0, +\infty) \textit{ and } \; \forall t \geq t_0 : \|y(t) - x(t)\| < \varepsilon).$$

Remarks 3.2 1. The very definition of Ljapunow stability is in some sense unclear: several authors use definition 3.1, see for instance [1], [3], [4].

Other do not mention at all the domain of the (reference) solutions y in definition 3.1, see for instance [5], [6], [7].

2. Definition 3.1 has two serious shortcomings.

First shortcoming.
The two statements of the conclusion of the implication:

y is defined on $[t_0, +\infty)$, $\forall\, t \geq t_0 : \|y(t) - x(t)\| < \varepsilon$ are not independent:

by the Ljapunow definition of stability we find a family of implications: $\forall\, \varepsilon > 0\; \exists\, \delta = \delta(\varepsilon) :$ $\|y(t_0) - x(t_0)\| < \delta \Longrightarrow$ (y exists on $[t_0, +\infty)$ and $\forall\, t \geq t_0 : \|y(t) - x(t)\| < \varepsilon$). This family depends on ε (and the associated $\delta(\varepsilon)$). Now we fix $\varepsilon = \overline{\varepsilon} > 0$ and we find $\delta = \delta(\overline{\varepsilon})$; indeed we now have one single implication: $\|y(t_0) - x(t_0)\| < \delta(\overline{\varepsilon}) \Longrightarrow$ (y is defined on $[t_0, +\infty)$ and $\forall\, t \geq t_0 : \|y(t) - x(t)\| < \overline{\varepsilon}$), in short: $A \Longrightarrow (B \wedge C)$.

But this implications is equivalent to

$$\neg(B \wedge C) \Longrightarrow \neg A\,, \text{ or}$$
$$\neg B \vee \neg C \Longrightarrow \neg A\,.$$

Now, if $\neg B$ is true, then there exists $t_1 \in (t_0, +\infty)$ such that y is not defined in $t_1 : y(t_1)$ does not exist.

If $\neg C$ is false, that is C is true, we have:

$$\forall\, t \geq t_0 : \|y(t) - x(t)\| < \overline{\varepsilon}\,,$$

which means $\|y(t) - x(t)\|$ is a (positive) real number and the assertion is: each of these numbers is smaller than $\overline{\varepsilon}$.

But here we find an error:

$$\|y(t_1) - x(t_1)\|$$

is no number, but a senseless symbol.

This senseless symbol also can occur if $\neg C$ is true. Then we find $t_2 \in (0, +\infty)$:

$$\|y(t_2) - x(t_2)\| \geq \overline{\varepsilon}\,.$$

and either $\|y(t_2) - x(t_2)\| \in \mathbb{R}$ or $\|y(t_2) - x(t_2)\|$ is a senseless symbol.

Second shortcoming.
If we have found a set of (explizite) solutions y, then we can often by the Ljapunow definition of stability easily, without starting to prove that x is stable or unstable, decide that the solution x is not stable. This we can conclude from the following proposition and its corollary.

Proposition 3.3 *We consider the initial value problem (3.1) and let $x : [t_0, +\infty) \to \mathbb{R}^n$ be a solution. If y is another solution $y : (a, b) \to \mathbb{R}^n$, $t_0 \in (a, b)$, we denote by $D(y)$ the domain (a, b) of y. Now we assume that there exists a sequence $(y_n)_{n\in\mathbb{N}}$ of solutions s. th. $y_n(t_0) \to x(t_0)$ and $\forall\, n \in \mathbb{N} : [t_0, +\infty) \nsubseteq D(y_n)$. Then x is not stable.*

Proof: We assume that x is stable: for $\varepsilon_0 = 1\ \exists\, \delta \in \mathbb{R},\ 0 < \delta \leq 1 : \forall\, y : \|y(t_0) - x(t_0)\| < \delta \Longrightarrow y$ is defined on $[t_0, +\infty)$ and $\forall\, t \geq t_0 : \|y(t) - x(t)\| < 1$; $\exists\, n_1 \in \mathbb{N} : y_{n_1}(t_0) \in U_\delta(x(t_0))$ and hence $\|y_{n_1}(t_0) - x(t_0)\| < \delta$. Thus y_{n_1} is defined on $[t_0, +\infty)$, yielding a constradiction since $[t_0, +\infty) \nsubseteq D(y_{n_1})$. Hence x is not stable.

Corollary 3.4 *Let S_0 be the set of all solutions of $\dot{y} = f(y)$, $y(t_0) = y_0 \in G$ which are not defined entirely on $(t_0, +\infty)$, hence $x \notin S_0$. Let S_0 be infinite and let $x(t_0)$ be a cluster point of $\{y(t_0) | y \in S_0\}$.*

Then x is not stable.

Example 3.5 We come back to example 2.2:

$$\dot{x} = x^3,\ x(0) = x_0 \in \mathbb{R}\,;$$

S_0 consists of all nontrivial solutions of the initial value problem and hence $\{y(0)|y \in S_0\} = (-\infty, 0) \cup (0, +\infty)$. Thus we can apply the corollary and since $x(0) = 0$ is a cluster point of $\{y(0)|y \in S_0\}$ we find that x is unstable.

But since we have no solution which we can compare with the zero function x on $[0, +\infty)$, the assertion "x is unstable" makes no sence.

4 The Reference-Stability

There exists a consequent and simple way out from the difficulties of the Ljapunow stability definition: we consider only the set of all solutions of the initial value problem which are defined (at least) on $[t_0, +\infty)$.

Definition 4.1 *Let x be defined on $[t_0, +\infty)$ and x is solution of the initial value problem (3.1)*

$$R = R(x) = \{y | y : [t_0, +\infty) \to \mathbb{R}^n,\ \dot{y} = f(y),\ y(t_0) = y_0 \in \mathbb{G} \text{ and } y \neq x\}\,;$$

$R(x)$ is called the set of reference solutions of the solution x. Of course, instead of $y : [t_0, +\infty) \to \mathbb{R}^n$ we can use: $[t_0, +\infty) \subseteq D(y)$.

Example 4.2 Let be $\dot{y} = f(y) = y$, $t_0 = 0$, $x : \forall\, t \geq 0 : x(t) = 0$, the zero solution: $x(0) = 0$. Then $R(x) = R(0) = \{y = y_0 e^t | y_0 \in \mathbb{R}\backslash\{0\}\}$ is the set of reference solutions of $x \equiv 0$.

Example 4.3 We consider example 2.4: $\dot{x} = -x^3$, $x(0) = x_0 \in \mathbb{R}$; then for the zero solution $x \equiv 0$, $x(0) = 0$, we find on

$$[0, +\infty) : R(x) = R(0) = \left\{ x = \frac{x_0}{\sqrt{1 + 2x_0^2 t}} \middle| x_0 \in \mathbb{R} \backslash \{0\} \right\} .$$

Definition 4.4 *We consider the initial value problem* (3.1) *and the solution* $x : [t_0, +\infty) \to \mathbb{R}^n$ *is to be investigated on stability; let* $R(x)$ *be the set of reference solutions of* x*; we assume:* $R(x) \neq \emptyset$*.* x *is called reference stable ,* R*-stable, iff* $\forall \, \varepsilon > 0 \, \exists \, \delta = \delta(\varepsilon)$, $0 < \delta \leq \varepsilon : \forall \, y \in R(x) :$

$$\|y(t_0) - x(t_0)\| < \delta \Longrightarrow \forall \, t \geq t_0 : \|y(t) - x(t)\| < \varepsilon \, .$$

Remarks 4.5 1. We emphasize insistently what was assumed in the definition: within reference stability we always assume $R(x) \neq \emptyset$. If $R(x) = \emptyset$ holds, we simply say that we have no stability problem. As an example for this situation we look at example 2.2: $\dot{x} = x^3$, $x(0) = 0$: here we have $R(x) = R(0) = \emptyset$.

2. As usual we still define: x is called to be asymptotically reference stable iff x is reference stable and $\exists \, \delta > 0 : \forall \, y \in R(x)$

$$\|y(t_0) - x(t_0)\| < \delta \Longrightarrow \lim_{t \to +\infty} \|y(t) - x(t)\| = 0$$

5 Topological characterization of the notion of reference stability

If x_0 is an equilibrium point of (3.1), then in [3] is defined: x_0 is called stable.iff for each neighborhood $V = V(x_0)$ there exists a neighborhood $U = U(x_0)$, $U \subseteq V$ and $U \subseteq G$ such that: for each solution y of (3.1), $y(t_0) = y_0 : y_0 \in U \Longrightarrow y$ is defined on $[t_0, +\infty)$ and $y([t_0, +\infty)) \subseteq V$.

In [8] the author considers only unique solutions and thus he can assign to each initial value $y(t_0)$ the solution y, $y(t_0) \to y$ and he assums that all y belong to the Banachspace $C_b([t_0, +\infty), \mathbb{R}^n)$ of all bounded continuous functions on $[t_0, +\infty)$ equipped with the sup-norm. Now he remarks that stability of a solution x is equivalent to the continuity of the map $y(t_0) \to y$ at the point $x(t_0)$. (See remark on page 137 of [8]). But the author has no precise domain of his map and the bounded continuous functions are not enough, since one wants for instance to consider instability too.

Best suited for topological characterization of stability is the notion of reference stability (see section 4).

Before we study such characterizations we will provide some facts from elementary general topology.

For topological spaces the notion of a neighborhood is important. Often a topology on a set is defined by open sets. But we also can start with neighborhoods. For a proof of the following propositions see [2], [9].

Proposition 5.1 *Let X be a set and for each $x \in X$ there exists a nonempty family $\underline{B}(x)$ of such subsets of X s. th. $B = (\underline{B}(x))_{x \in X}$ has the properties:*

(a) $B \in \underline{B}(x) \Longrightarrow x \in B$

(b) $B_1,\ B_2 \in \underline{B}(x)\, \exists\, B_3 \in \underline{B}(x) : B_3 \subseteq B_1 \cap B_2$

(c) $\forall\, V \in \underline{B}(x)\, \exists\, B \in \underline{B}(x)\, \forall\, y \in B\, \exists\, W \in \underline{B}(y) : W \subseteq V$

$G \subseteq X$ is called open iff

$$\forall\, x \in G\, \exists\, B \in \underline{B}(x) : B \subseteq G\,.$$

Then $\tau = \{G \subseteq X | G \text{ open}\}$ is a topology on X and τ is uniquely determined by the system $B = (\underline{B}(x))_{x \in X}$.

Moreover $\forall\, x \in X : \underline{B}(x)$ is a base of the τ-neighborhoodsystem $\underline{U}(x)$.

Hence we say that the base system B generates the topology τ.

Corollary 5.2 *Let be τ_1, τ_2 topologies on X which are generated by the base neighborhood systems $(\underline{B}^1(x))_{x \in X}$, $(\underline{B}^2(x))_{x \in X}$.*

If holds: $\forall\, x \in X\, \forall\, B_1 \in \underline{B}^1(x)\, \exists\, B_2 \in \underline{B}^2(x) : B_2 \subseteq B_1$ then we find: $\tau_1 \subseteq \tau_2$.

Proof: $\forall\, G \in \tau_1$, hence G is open w. r. t. τ_1 and we want to show that G is τ_2-open too: $\forall\, z \in G : G \in \tau_1 \Longrightarrow\ \exists B_1 \in \underline{B}^1(z) : z \in B_1 \subseteq G$; by assumption there exists $B_2 \in \underline{B}^2(z)$ s. th. $B_2(z) \subseteq B_1(z) \Longrightarrow z \in B_2 \subseteq G$ and hence G is open w. r. t, $\tau_2 : G \in \tau_2$.

Now we are looking for suitable topologies on $C([t_0, +\infty), \mathbb{R}^n)$; $[t_0, +\infty)$ (with Euclidian topology) is a locally compact Hausdorff space. Thus the compact-open topology for $C([t_0, +\infty), \mathbb{R}^u)$ has many open sets. But for applications to characterize stability we need "uniform topologies".

Remark 5.3 Algebraic operations in $C([t_0, +\infty), \mathbb{R}^n)$ and in $C([t_0, +\infty), \mathbb{R})$ we can define pointwise; we consider these spaces as vector spaces over $\mathbb{R}$.

Definition 5.4 *Let $M \subseteq C([t_0, +\infty), \mathbb{R})$, $M \neq 0$ and all functions from M are positive: $\forall\, (\alpha, t) \in M \times [t_0, +\infty) : \alpha(t) > 0$; now for $f \in C([t_0, +\infty), \mathbb{R}^u)$ we define α-neighborhoods of $f : B_\alpha(f) = \{g \in C([t_0, +\infty), \mathbb{R}^u) | \forall\, t \in [t_0, +\infty) : \|g(t) - f(t)\| < \alpha(t)\}$.*

Which properties M must have such that $B = (B_\alpha(f))_{(\alpha,f) \in M \times C([t_0,+\infty),\mathbb{R}^n)}$ is a base neighborhood system (see proposition 5.1).

Proposition 5.5 *We assume that holds:*

(1) $\alpha \in M \Longrightarrow \frac{1}{2}\alpha \in M$

(2) $\alpha, \beta \in M \Longrightarrow \min\{\alpha, \beta\} \in M.$

Then B is a base neighborhood system.

Proof: At first we remark that $\frac{1}{2}\alpha$, $\min\{\alpha, \beta\}$ are positive continuous functions. We will show that B fulfills the base neighborhood systems axioms (a), (b), (c) of proposition 5.1.

(a) $\forall\,(\alpha, f) : f \in B_\alpha(f)$, since $\forall\, t \in [t_0, +\infty)$:

$$\|(f(t) - f(t)\| = 0 < \alpha(t)$$

(b) $\forall\, f \in C([t_0, +\infty), \mathbb{R}^n)$
$\forall\, \alpha_1, \alpha_2 \in M$: let $\beta = \min\{\alpha_1, \alpha_2\}$, then $B_\beta(f) \subseteq B_{\alpha_1}(f) \cap B_{\alpha_2}(f)$, since

$$\forall\, t \geq t_0 : \min\{\alpha_1(t), \alpha_2(t)\} \leq \alpha_1(t),\ \min\{\alpha_1(t), \alpha_2(t)\} \leq \alpha_2(t)$$

(c) $\forall\, B_\alpha(f) \in (B_\beta(f))_{\beta\in M} : \alpha \in M \Longrightarrow \frac{1}{2}\alpha \in M \Longrightarrow B_{\frac{\alpha}{2}}(f) \in (B_\beta(f))_{\beta\in M}; \forall\, g \in B_{\frac{\alpha}{2}}(f)$: we will show that $B_{\frac{\alpha}{2}}(g) \subseteq B_{\frac{\alpha}{2}}(f)$ holds: $\forall\,(h, t) \in B_{\frac{\alpha}{2}}(g) \times [t_0, +\infty)$:

$$\begin{aligned}\|h(t) - f(t)\| &= \|h(t) - g(t) + g(t) - f(t)\| \leq \|h(t) - g(t)\| + \|g(t) - f(t)\| \\ &< \frac{1}{2}\alpha(t) + \frac{1}{2}\alpha(t) = \alpha(t)\,,\end{aligned}$$

hence $h \in B_\alpha(f)$.

Remark 5.6 If M fulfills (1), (2) then the α-base neighborhood system generates an unique topology $\tau = \tau_M$ for $C([t_0, +\infty), \mathbb{R}^u)$.

Lemma 5.7 *$M_1, M_2 \subseteq C([0, +\infty), \mathbb{R})$, M_1, M_2 generate the topologies τ_1, τ_2 respectively. Then holds:*

$$M_1 \subseteq M_2 \Longrightarrow \tau_1 \subseteq \tau_2$$

Proof: We show that the identity map id: $(C([t, +\infty), \mathbb{R}^u), \tau_2) \to (C([t_0, +\infty), \mathbb{R}^u), \tau_1)$ is continuous: let $G \in \tau_1$ be open $\Longrightarrow \mathrm{id}^{-1}(G) = G$; $G \in \tau_1 \Longrightarrow \forall\, h \in G\ \exists\, \alpha \in M_1 : B(h) = \{g \in C([t_0, +\infty), \mathbb{R}^n) | \forall\, t \geq t_0 : \|g(t) - h(t)\| < \alpha(t)\} \subseteq G$. But $\alpha \in M_1 \Longrightarrow \alpha \in M2$ and hence $G \in \tau_2$.

Definition 5.8 *Now we consider examples of the generating set $M \subseteq C([t_0, +\infty), \mathbb{R})$ and the corresponding topologies:*

1. *By ε we mean now the constant function:*

$$\varepsilon : \forall\, t \in [t_0, +\infty) : \varepsilon(t) = \varepsilon,\ \ M_\varepsilon = \{\varepsilon | \varepsilon > 0\}$$

   Of course:

$$\varepsilon > 0 \Longrightarrow \frac{1}{2}\varepsilon > 0;\ \ \varepsilon_1, \varepsilon_2 \in M_\varepsilon \Longrightarrow \min\{\varepsilon_1, \varepsilon_2\} \in M_\varepsilon$$

   As is well known, $B = (B_\varepsilon(f))_{(\varepsilon, f) \in M_\varepsilon \times C([t_0,+\infty),\mathbb{R}^n)}$ generates the uniform topology τ_u on $C([t_0, +\infty), \mathbb{R}^n)$

2. *We consider a subset of $M_\varepsilon : \forall\, n \in \mathbb{N},\ n \geq 1 : \varepsilon_n = \frac{1}{n}$: the constant functions now have the value $\frac{1}{n}$; $M_{\left(\frac{1}{n}\right)} = \left\{\frac{1}{n} | n \in \mathbb{N}, n \geq 1\right\}$. $M_{\left(\frac{1}{n}\right)}$generates a topology:*

$$\frac{1}{n} \in M_{\left(\frac{1}{n}\right)} \Longrightarrow \frac{1}{2}\frac{1}{n} = \frac{1}{2n} \in M_{\left(\frac{1}{n}\right)}, \quad \min\left\{\frac{1}{n}, \frac{1}{m}\right\} \in M_{\left(\frac{1}{n}\right)}. \tag{5.1}$$

3. *M_c, the symbol c, means: converging to zero; $M_c = \{\alpha \in M | \lim_{t\to+\infty} \alpha(t) = 0\}$. We denote the topology generated by M_c on $C([t_0, +\infty), \mathbb{R}^n)$ by τ_{pc}: positive – converging topology. Clearly:*

$$\alpha \in M_c \Longrightarrow \frac{1}{2}\alpha \in M_c, \alpha_1, \alpha_2 \in M_c \Longrightarrow \forall\, t \geq t_0 : 0 < \min\{\alpha_1(t), \alpha_2(t)\} \leq \alpha_1(t)$$
$$(\mathit{and}\ \leq \alpha_2(t))$$

   and thus $\lim\limits_{t\to+\infty} \min\{\alpha_1(t), \alpha_2(t)\} = 0$ showing $\min\{\alpha_1, \alpha_2\} \in M_c$.

4. *$M_a = \{\alpha \in C([t_0, +\infty), \mathbb{R}) | \forall\, t \geq t_0 : \alpha(t) > 0\}$; thus a means "all". Of course:*

$$\alpha \in M_a \Longrightarrow \frac{1}{2}\alpha \in M_a, \alpha_1, \alpha_2 \in M_a \Longrightarrow \min\{\alpha_1, \alpha_2\} \in M_a\,.$$

The topology generated by M_a we denote by τ_m, since this topology was used by Marston Morse; τ_m first was defined by E. Hewitt, it is also called Whitney – or fine topology.

As we have hoped, we can show: $\tau_u = \tau_{\left(\frac{1}{n}\right)}$.

Proposition 5.9 *On $C([t_0, +\infty), \mathbb{R}^n)$ holds $\tau_u = \tau_{\left(\frac{1}{n}\right)}$.*

Proof: $M_{\left(\frac{1}{n}\right)} \subseteq M_\varepsilon \Longrightarrow \tau_{\left(\frac{1}{n}\right)} \subseteq \tau_u$ by lemma 5.7. By corollary 5.2 we find $\tau_u \subseteq \tau_{\left(\frac{1}{n}\right)}$.

Corollary 5.10 *$(C([t_0, +\infty), \mathbb{R}^u)\tau_u)$ is a topological A_1-space. Hence we can use sequences instead of nets or filter.*

Proposition 5.11 *Moreover we have: $\tau_u \leq \tau_{pc}$.*

Proof: $\forall\,(f,\varepsilon)\in C([t_0,+\infty),\mathbb{R}^n)\times(0,+\infty): B_\varepsilon(f)\in\tau_u$; let $h=\frac{\varepsilon}{2}e^{-t}$, $t\in[t_0,+\infty)$; since $0\leq t_0$ we get for $0\leq t_0\leq t: e^{-t}\leq 1\Longrightarrow\frac{\varepsilon}{2}e^{-t}\leq\frac{\varepsilon}{2}<\varepsilon$, thus showing that $B_h(f)\subseteq B_\varepsilon(f)$ holds and $B_h(f)\in\tau_{pc}$. Hence by corollary 5.2 $\tau_u\subseteq\tau_{pc}$.

Corollary 5.12 *For our topologies τ_u, τ_{pc}, τ_m holds:*

$$\tau_u\leq\tau_{pc}\leq\tau_m$$

Now we come to the main point of this section: stability as continuity.

As already remarked the basic idea of stability is nothing else then the continuity of a natural map into the space of continuous functions. Using reference stability we can define this map in a clear and exact way:

We consider the initial value problem (3.1). Let x be a solution which is defined on $[t_0,+\infty)$ and $R(x)$ be the set of reference solutions of x (definition 4.1).

Let $\tilde{R}(x)=R(x)\cup\{x\}$ and we assume that all solution of $\tilde{R}(x)$ are unique; moreover V_{t_0} (V means “value”)$=\{y(t_0)|y\in\tilde{R}(x)\}$, $V_{t_0}\subseteq G\subseteq\mathbb{R}^n$ and for V_{t_0} we consider the Euclidian topology of $\mathbb{R}^n$, which can be generated by an arbitrary compatible norm of $\mathbb{R}^n$. Then the map F is well defined:

$$F:V_{t_0}\to C([t_0,+\infty),\mathbb{R}^n):\forall\,y(t_0)\in V_{t_0}:F(y(t_0))=y$$

$C([t_0,+\infty),\mathbb{R}^n)$ we provide with the uniform topology τ_u.

Remark 5.13 Since of course some $y\in\tilde{R}(x)$ may be unbounded we use $C([t_0,+\infty),\mathbb{R}^n)$ and not the space $C_b([t_0,+\infty),\mathbb{R}^n)$ of bounded continuous maps.

Now using the generation of τ_u by base ε-neighborhoods (see 5.8, 1.) and the characterization of the continuity of a map by (base) neighbourhoods it is not hard to prove the assertion of the following theorem:

Theorem 5.14 *Equivalent are:*

(1) *x is reference stable*

(2) *the map $F:V_{t_0}\to(C([t_0,+\infty),\mathbb{R}^n),\tau_u)$ is continuous in $x(t_0)$.*

Application of theorem 5.14 to concrete examples. We consider again Example 2.4: $\dot{x}=-x^3$, $t_0=0$, $x(0)=x_0$; on $[0,+\infty)$ are defined:

the zero solution 0 and the set of reference solutions

$$R(0)=\left\{x=\frac{x_0}{\sqrt{1+2x_0^2t}}|x_0\in\mathbb{R}\backslash\{0\}\right\},\text{ hence }\tilde{R}(0)=\left\{x=\frac{x_0}{\sqrt{1+2x_0^2t}}|x_0\in\mathbb{R}\right\};$$

for $x_0 = 0$ we obtain the zero function:

$$\frac{0}{\sqrt{1}} = 0\,, \quad V_{t_0} = V_0 = \left\{x(0) = x_0 | x_0 \in \tilde{R}(0)\right\} = \mathbb{R}\,.$$

Of course the solutions of $R(0)$ are unique solutions and by proposition 2.1 $x \equiv 0$ is unique. Thus all elements of $\tilde{R}(0)$ are unique solutions and we can apply theorem 5.14:

Now we show the continuity of the map F: we can use convergence too and especially we can use sequences here:

let (x_0^n) be a sequence from V_0 s. th. $(x_0^u) \to 0(0) = 0$. We will show:

$$F(x_0^n) = (x(t;0,x_0))_n \to F(0(0)) = 0 \text{ uniformly on } [0,+\infty) :$$

$$\begin{aligned}\forall\, t \geq 0,\ x_0 \neq 0,\ 1 \leq 1 + 2x_0^2 t \Longrightarrow 1 \leq \sqrt{1+2x_0^2 t} \Longrightarrow 0 < \frac{1}{\sqrt{1+2x_0^2 t}} \leq 1 \\ \Longrightarrow 0 < \frac{|x_0|}{\sqrt{1+2x_0^2 t}} \leq |x_0|\,.\end{aligned}$$

Hence $|F(x_0^n) - 0| = |F(x_0^u)| \leq |x_0^n|$, but $x_0^u \to 0 \Longrightarrow |x_0^u| \to 0 \Longrightarrow F(x_0^u) \to 0$ uniformly on $[0,+\infty)$, since $|x_0^u|$ does not depend on t.

Thus from theorem 5.14 follows that the zero solution 0 is reference stable.

Remarks 5.15 1. Using the continuity – arguments we were able to avoid any epsilontics.

2. Let $\delta = 1 : \forall\, x_0 \in \mathbb{R}$, $x_0 \neq 0$, $|x_0 - 0(0)| = |x_0| < 1$ we get

$$\lim_{t\to+\infty} |x(t)| = \lim_{t\to+\infty} \frac{|x_0|}{\sqrt{1+2x_0^2 t}} = 0\,,$$

hence the zero solution 0 is even asymptotically reference stable.

Example 5.16 We consider the equation $\dot{x} = x^2$, $x(0) = x_0$.

For $x_0 \neq 0$ by (2.1) we find for the solutions:

$$x^{1-n} = x_0^{1-n} + (1-n)t$$

and

$$n = 2 \Longrightarrow x^{-1} = x_0^{-1} - t \Longrightarrow x = \frac{1}{x_0^{-1} - t} = \frac{x_0}{1 - x_0 t}\,;\ x_0 \neq 0 :$$

Since we look for solutions which are defined at least on $[0,+\infty)$, we find here:

$$\frac{1}{x_0} \notin (0,+\infty) \Longleftrightarrow x_0 < 0 \Longleftrightarrow [0,+\infty) \subseteq \left(\frac{1}{x_0}, +\infty\right) \Longleftrightarrow [0,+\infty) \subseteq D(x(t;0,x_0))$$

and equivalently:

$$\frac{1}{x_0} \in (0,+\infty) \Longleftrightarrow x_0 > 0 \Longleftrightarrow [0,+\infty) \nsubseteq D(x(t;0,x_0))\,.$$

Now we study the stability of the zero solution $x \equiv 0$ on $[0,+\infty)$.

$\{x_0 \big| [0,+\infty) \nsubseteq D(x(t;0,x_0)) = \{x_0 | x_0 > 0\} = (0,+\infty)\}$. But $0 = 0(0)$ is clusterpoint of $(0,+\infty)$ yielding by corollary 3.4 that $x \equiv 0$ is unstable in the sense of Ljapunow.

Since we know that only for $x(0) = x_0 < 0$ holds: $[0,+\infty) \subseteq D(x(t;0,x_0))$ we get as set of all reference solution of $x \equiv 0$:

$$\begin{aligned} R(0) &= \left\{ x = \frac{x_0}{1-x_0t} | x_0 < 0 \right\} \Longrightarrow \tilde{R}(0) = \left\{ x = \frac{x_0}{1-x_0t} | x_0 \le 0 \right\}; \\ V_0 &= \left\{ x(0) | x \in \tilde{R}(0) \right\} = (-\infty, 0]\,. \end{aligned}$$

We show that $x \equiv 0$ is reference stable: By the same arguments as above we find too: all elements of $\tilde{R}(0)$ are unique solutions.

Now:

$$\begin{aligned} & x_0 < 0 \Longrightarrow -x_0 > 0 \Longrightarrow -tx_0 \ge 0,\ \text{since } t \ge 0; \\ & -tx_0 \ge 0 \Longrightarrow 1 - tx_0 \ge 1 \Longrightarrow \frac{1}{1-x_0t} \le 1 \Longrightarrow \frac{|x_0|}{1-x_0t} \le |x_0|\,; \end{aligned}$$

now let (x_0^n) be a sequence from $V_0 \backslash \{0\}$, $x_0^n = (x(t;0,x_0))_n$;

$$\begin{aligned} & |F(x_0^n) - 0(0)| = |F(x_0^n)| = \left| \frac{x_0^n}{1-x_0^n t} \right| \le |x_0^n|\,; \\ & (x_0^n) \to 0 \Longrightarrow |x_0^n| \to 0 \Longrightarrow F(x_0^n) \to 0 = 0(0) = F(0) : (F(x_0^n))_n \end{aligned}$$

converges uniformly on $[0,+\infty)$ to $F(0)$, yielding that F is continuous in $0(0) = 0$. Hence by theorem 5.14 $0 = x \equiv 0$ is reference stable.

$$\forall\, x_0 < 0 : |x_0| < 1 \Longrightarrow \lim_{t\to+\infty} |x(t;0,x_0)| = \lim_{t\to+\infty} \frac{|x_0|}{1-x_0t} = 0\,,$$

meaning that 0 is asymptotically reference stable.

We need a simple lemma.

Lemma 5.17 *Let be (h_n) a sequence from $C([0,+\infty),\mathbb{R}^n)$ and let $h \in ([0,+\infty),\mathbb{R}^n)$ be bounded: $\forall t \in [0,+\infty)\ \|h(t)\| \le a$, $a \in \mathbb{R}$, $a > 0$. If (h_n) converges uniformly to h on $[0,+\infty)$ then almost all members of the sequence (h_n) are bounded too.*

Proof: $h_n \xrightarrow{\tau_u} h \Longrightarrow \exists\, n_1 \in \mathbb{N} : \forall\,(t,n) \in [0,+\infty) \times \{n \in \mathbb{N} | n \geq n_1\}$: $\|h_n(t) - h(t)\| < 1$; now

$$\|h_n(t)\| - \|h(t)\| \leq \|h_n(t) - h(t)\| \Longrightarrow \|h_n(t)\| \leq \|h_n(t) - h(t)\| + \|h(t)\| < 1 + a\,,$$

hence h_n is bounded $\forall\, n \geq n_1$.

We will apply this lemma and consider example 4.2: $\dot{x} = x$, $x(0) = x_0$; $x \equiv 0$ on $[0,+\infty)$ is solution: $0(0) = 0$.

$$R(0) = \{x_0 e^t | x_0 \in \mathbb{R}\backslash\{0\}\} \Longrightarrow \tilde{R}(0) = \{x_0 e^t | x_0 \in \mathbb{R}\}, V_0 = \{x(0) = x_0 | x \in \tilde{R}(0)\}\,.$$

$x \equiv 0$ is not reference stable.

Proof: We consider the sequence $(x_0^n) = \left(\frac{1}{n}\right)$ from V_0; $\frac{1}{n} \to 0$ but all $F\left(\frac{1}{n}\right) = x\left(t; 0, \frac{1}{n}\right) = \frac{1}{n}e^t$ are unbounded and hence by the lemma 5.17 $F\left(\frac{1}{n}\right)$ does not converges uniformly to 0.

Thus by theorem 5.14 $x \equiv 0$ is not reference stable.

We still consider the positive – converging topology τ_{pc}.

Proposition 5.18 *Under the assumptions of theorem 5.14 holds:*

If the solution x is τ_{pc}-stable then x is asymptotically reference stable.

Proof: We consider the map

$$F : V_{t_0} \to C([t_0,+\infty), \mathbb{R}^n),\;\; x(t_0) \in V_{t_0} \Longrightarrow F(x(t_0) = x \in C([t_0,+\infty), \mathbb{R}^n)$$

and x is τ_{pc}-stable means that

$$F : V_{t_0} \to (C([t_0,+\infty), \mathbb{R}^n), \tau_{pc})$$

is continuous in $x(t_0)$; 5.11 shows that $\tau_u \subseteq \tau_{pc}$, yielding that $F : V_{t_0} \to (C([t_0,+\infty), \mathbb{R}^n), \tau_u)$ is continuous in $x(t_0)$ too. Hence by theorem 5.14 x is R-stable.

Let $\alpha \in M_c$: $\forall\, t \in [t_0,+\infty)$: $\alpha(t) > 0$ and $\lim_{t\to\infty} \alpha(t) = 0$. By the τ_{pc}-continuity of F in $x(t_0)$ we find $\delta > 0$, $\delta = \delta(\alpha) : \forall\, y(t_0) \in U_\delta(x(t_0)) \cap V_{t_0} \Longrightarrow F(y(t_0)) = y \in U_\alpha(x(t_0)) \Longrightarrow \forall\, t \geq t_0$:

$$\|y(t) - x(t)\| < \alpha(t),\;\; \alpha(t) \to 0 \Longrightarrow \|y(t) - x(t)\| \to 0$$

for $t \to +\infty$, showing that x is asymptotically R-stable, since we finally have: $\forall\, y \in R(x)$: $\|y(t_0) - x(t_0)\| < \delta \Longrightarrow \|y(t) - x(t)\| \to 0$ for $t \to +\infty$.

Remark 5.19 If we want to define the basics of reference stability by means of topologies for the function space $C([0,+\infty), \mathbb{R}^n)$, then we can proceed:

1. reference stability by the uniform topology τ_u
2. asymptotic reference stability by τ_{pc}.

References

[1] **Barreira, L.**, and **Valls, C.** : *Ordinary Differential Equations.* Amer. Math. Soc., vol. 137, 2012

[2] **Engelking, R.** : *General Topology.* Heldermann Verlag Berlin, 1989

[3] **Hirsch, M. W., Smale, S.**, and **Devaney, R. L.** : *Differential Equations, Dynamical Systems, and an Introduction to Chaos.* Sec. Ed., Academic Press, 2004

[4] **Kabelka, B.** : *Gewöhnliche Differentialgleichungen.* http://fsmath.at/~bkabelka/math/analysis/diff-gl/16.htm

[5] **Krasnov, M. L., Kislyov, A. I.**, and **Makarenko, G. I.** : *A Book of Problems in Ordinary Differential Equations.* Mir Publishers Moscow, 1981

[6] **Rouche, N., Habets, P.**, and **Laloy, M.** : *Stability Theory by Ljapunows Direct Method.* Springer Verlag, 1973

[7] **Verhulst, F.** : *Nonlinear Differential Equation and Dynamical Systems.* Springer Verlag, 1996

[8] **Vidyasagar, M.** : *Nonlinear Systems Analysis.* Sec. Edit., Prentice Hall, 1993

[9] **Willard, S.** : *General Topology.* Addison-Wesley Publ. Comp., 2004

received: April 28, 2017

Author:

Harry Poppe
Universität Rostock
Institut für Mathematik
e-mail: harry.poppe@uni-rostock.de

Rostock. Math. Kolloq. **71**, 28 – 40 (2017/18)

Subject Classification (MSC)
Primary 20L05, 46C50;
Secondary 39B32, 39B62

Zoltán Boros, Árpád Száz

A weak Schwarz inequality for semi-inner products on groupoids[*]

ABSTRACT. By introducing appropriate notions of semi-inner products and their induced generalized seminorms on groupoids, we shall prove a weak form of the famous Schwarz inequality.

In case of groups, this will be sufficient to prove the subadditivity of the induced generalized seminorms. Thus, some of the results of the theory of inner product spaces can be extended to inner product groups.

However, in the near future, we shall only be interested in the corresponding extensions of some fundamental theorems of Gy. Maksa, P. Volkmann, A. Gilányi, J. Rätz and W. Fechner on additive and quadratic functions.

KEY WORDS AND PHRASES. Groupoids, additive functions, semi-inner products, generalized seminorms, Schwarz inequality, triangle inequality.

1 Introduction

By introducing appropriate notions of semi-inner products and their induced generalized seminorms on groupoids, we shall prove a weak form of the famous Schwarz inequality.

More concretely, if X is an additively written groupoid and P is a function of X^2 to $\mathbb{C}$ such that

$$P(x, x) \geq 0, \qquad P(y, x) = \overline{P(x, y)}, \qquad P(x+y, z) = P(x, z) + P(y, z)$$

for all $x, y, z \in X$, then by using the notation

$$p(x) = \sqrt{P(x, x)}$$

[*] The work of the authors has been supported by the Hungarian Scientific Research Fund (OTKA) Grant K-111651.

with $x \in X$, we shall prove that

$$-P_1(x, y) \leq p(x)p(y)$$

for all $x, y \in X$, where P_1 denotes the real part, i. e., the first coordinate function of P.

If in particular X is a group, then this weak Schwarz inequality already implies that $P_1(x, y) \leq p(x)p(y)$ also holds for all $x, y \in X$. Therefore, in this important particular case, the generalized seminorm p can be proved to be a seminorm on X in the sense it is an even, $\mathbb{N}$–homogeneous, subadditive function of X to $\mathbb{R}$.

Thus, some of the results of the theory of inner product spaces can be naturally extended to inner product groups. However, in the near future, we shall only be interested in the corresponding extensions of some fundamental theorems of Maksa and Volkmann [14], Gilányi [8], Rätz [15] and Fechner [6] on additive and quadratic functions.

2 Additive functions of groupoids

If X is a set, then a function $+$ of X^2 to X is called an operation on X, and the ordered pair $X(+) = (X, +)$ is called a groupoid.

In the sequel, as is customary, we shall simply write X in place of $X(+)$. And, for any $x, y \in X$, we shall write $x + y$ in place of the value $+(x, y)$.

Moreover, for any $x \in X$ and $n \in \mathbb{N}$, with $n > 1$, we define

$$1\,x = x \qquad \text{and} \qquad n\,x = (n-1)\,x + x\,.$$

If in particular, X is group, then for any $x \in X$ and $n \in \mathbb{N}$ we may also naturally define

$$0\,x = 0 \qquad \text{and} \qquad (-n)\,x = n\,(-x)\,.$$

A function f of one groupoid X to another Y is called additive if

$$f(x+y) = f(x) + f(y)$$

for all $x, y \in X$.

Moreover, the function f may be naturally called $\mathbb{N}$–homogeneous if it is n–homogeneous for all $n \in \mathbb{N}$ in the sense that $f(n\,x) = n\,f(x)$ for all $x \in X$.

Additive functions were first studied only on $\mathbb{R}$ or $\mathbb{R}^n$ (see Kuczma [12]). However, later they have also been intensively investigated on arbitrary groups (see Stetkaer [21]).

Some of the results obtained in groups can be naturally extended to monoids and semigroups. In [17] and [10], additive functions and relations were considered on groupoids too.

For instance, by induction, we can easily prove the following

Theorem 2.1 *If f is an additive function of a groupoid X to another Y, then f is $\mathbb{N}$–homogeneous.*

Proof. To check this, note that if $f\,(n\,x) = n\,f(x)$ holds for some $x \in X$ and $n \in \mathbb{N}$, then we also have

$$f\,\big((n+1)\,x\big) = f\,(n\,x + x) = f\,(n\,x) + f\,(x) = n\,f\,(x) + f\,(x) = (n+1)\,f(x)\,. \qquad \square$$

Remark 2.2 If f is an additive function of a groupoid X, with zero, to a group Y, then f is 0–homogeneous too.

Namely, in this case, we have

$$f\,(0) + f\,(0) = f\,(\,0+0\,) = f\,(0)\,,$$

and thus $f\,(0) = 0$. Therefore,

$$f\,(0\,x) = f\,(0) = 0 = 0\,f\,(x)$$

also holds for all $x \in X$.

Now, by using the above observations and the corresponding definitions, we can also easily prove the following

Theorem 2.3 *If f is an additive function of a group X to another Y, then f is $\mathbb{Z}$–homogeneous.*

Proof. If $x \in X$, then by using Remark 2.2 we can see that

$$f\,(-x) + f\,(x) = f\,(-x+x) = f\,(0) = 0\,,$$

and thus $f\,(-x) = -f\,(x)$. Now, if $n \in \mathbb{N}$, then by using Theorem 2.1 we can also see that

$$f\,\big((-n)\,x\big) = f\,\big(n\,(-x)\big) = n\,f\,(-x) = n\,\big(-f\,(x)\big) = (-n)\,f\,(x)\,.$$

Therefore, f is also $-N$–homogeneous. Thus, by Theorem 2.1, the required assertion is also true. □

In addition to the above theorems, sometimes we shall also need the following

Theorem 2.4 *If f is an additive function of an arbitrary groupoid X to a commutative one Y, then for any $x,\,y \in X$ we have*

$$f\,(y+x) = f\,(x+y)\,.$$

Proof. By the above assumptions, we evidently have

$$f\,(y+x) = f\,(y) + f\,(x) = f\,(x) + f\,(y) = f\,(x+y)\,. \qquad \square$$

Remark 2.5 In this case, in contrast to the termilogy of Stetkaer [21, p. 315], we would rather say that f is commutative.

3 Semi-inner products on groupoids

The following definition is a straightforward generalization of that introduced in [19] and [3].

Notation 3.1 Suppose that X is a groupoid and P is a function of X^2 to $\mathbb{C}$ such that, for any $x, y, z \in X$, we have

(a) $P(x, x) \geq 0$,

(b) $P(y, x) = \overline{P(x, y)}$,

(c) $P(x+y, z) = P(x, z) + P(y, z)$.

Remark 3.2 In this case, the function P will be called a *semi-inner product* on X.

Moreover, if in particular X has a zero, then the semi-inner product P will be called an inner product if

(d) $P(x, x) = 0$ implies $x = 0$ for all $x \in X$.

Remark 3.3 Thus, our present definition is in accordance with that of [16], but differs from that used by Lumer [11] and Giles [9]. (See also Dragomir [4, p. 19] for some further developments.)

The definition and results of the above mentioned authors allowed to carry over some arguments in inner product spaces to those in normed spaces. While, our ones will only allow of a similar transition from inner product spaces to inner product groups.

Example 3.4 If a is an additive function of X to an inner product space H and

$$Q(x, y) = \langle a(x), a(y) \rangle$$

for all $x, y \in X$, then Q is a semi-inner product on X. Moreover, if if in particular X is a group, then Q is an inner product if and only if a is injective.

Despite this, Q may be a rather curious function even if $X = \mathbb{R}^n$ and $H = \mathbb{R}$. Namely, by Kuczma [12, p. 292], there exist discontinuous, injective additive functions of $\mathbb{R}^n$ to $\mathbb{R}$. In the case $n = 1$, by Makai [13], Kuczma [12, p. 293] and Baron [1], we can say even more.

The most basic properties of the semi-inner product P can be listed in the next

Theorem 3.5 *For any $x, y, z \in X$ and $n \in \mathbb{N}$, we have*

(1) $P(y+x, z) = P(x+y, z)$,

(2) $P(x, z+y) = P(x, y+z)$,

(3) $P(x, y+z) = P(x, y) + P(x, z)$,

(4) $P(nx, y) = nP(x, y) = P(x, ny)$.

Proof. By using (b) and (c), and the additivity of complex conjugation, we can see that (3) is true.

Thus, P is actually a biadditve function of X^2 to $\mathbb{C}$. Hence, by Theorem 2.1, it is clear that (4) is also true.

Moreover, by using (c) and (3) and the commutativity of the addition in $\mathbb{C}$, we can see that (1) and (2) are also true. □

Remark 3.6 Note that if in particular X has a zero, then by Remark 2.2 we have $P(x, 0) = 0$ and $P(0, y) = 0$, and thus also

$$P(0x, y) = 0\,P(x, y) = P(x, 0y)$$

for all $x, y \in X$.

Moreover, if more specially X is a group, then by Theorem 2.3 we have

$$P(kx, y) = k\,P(x, y) = P(x, ky)$$

for all $k \in \mathbb{Z}$ and $x, y \in X$.

Remark 3.7 Note that the first and second coordinate functions P_1 and P_2 of P also have the same commutativity and bilinearity properties as P.

Furthermore, by properties (a) and (b), for any $x, y \in X$ we have

(1) $P_1(x, x) = P(x, x)$ and $P_2(x, x) = 0\,,$

(2) $P_1(y, x) = P_1(x, y)$ and $P_2(y, x) = -P_2(x, y)\,.$

Thus, in particular P_1 is also a semi-inner product on X. However, because of its skew-symmetry, P_2 cannot be a semi-inner product on X whenever $P_2 \neq 0$.

More exactly, one can easily prove the following

Theorem 3.8 *A function Q of X^2 to $\mathbb{C}$ is a semi-inner product if and only if for any $x, y \in X$ we have*

(1) $Q_1(x, x) \geq 0$ *and* $Q_2(x, x) = 0\,,$

(2) $Q_1(y, x) = Q_1(x, y)$ *and* $Q_2(y, x) = -Q_2(x, y)\,,$

(3) $Q_i(x + y, z) = Q_i(x, z) + Q_i(y, z)$ *for* $i = 1$ *and* $i = 2\,.$

Remark 3.9 Note that the second part of (2) implies that of (1). Moreover, the second parts of (2) and (3) imply that Q_2 is additive in its second variable too.

Therefore, by the above theorem, we can also state that a function Q of X^2 to $\mathbb{C}$ is a semi-inner product if and only if Q_1 is a semi-inner product and Q_2 is a skew-symmetric and biadditive.

4 The induced generalized norm

Definition 4.1 *For any $x \in X$, we define*

$$p(x) = \sqrt{P(x, x)}\,.$$

Example 4.2 If in particular Q is as in Example 3.4, then

$$q(x) = \sqrt{Q(x, x)} = \|\, a(x)\, \|$$

for all $x \in X$.

The most immediate properties of the function p can be listed in the following

Theorem 4.3 *For any $x, y \in X$ and $n \in \mathbb{N}$, we have*

(1) $p(x) \geq 0\,,$

(2) $p(nx) = n\,p(x)\,,$

(3) $p(x+y) = p(y+x)\,,$

(4) $p\big(n(x+y)\big) = p(nx+ny)\,,$

(5) $p(x+y)^2 = P_1(x+y, x) + P_1(x+y, y)\,,$

(6) $p(x+y)^2 = p(x)^2 + p(y)^2 + 2\,P_1(x, y)\,.$

Proof. To prove (5) and (6), note that by the Definition 4.1 and Remark 3.7 we have

$$p(x) = \sqrt{P_1(x, x)}$$

and

$$\begin{aligned} p(x+y)^2 &= P_1(x+y,\ x+y) = P_1(x+y, x) + P_1(x+y, y) \\ &= P_1(x, x) + P_1(y, x) + P_1(x, y) + P_1(y, y) = p(x)^2 + 2\,P_1(x, y) + p(y)^2. \end{aligned}$$

Hence, by the symmetry of P_1 and the commutativity of the addition in $\mathbb{R}$, it is clear that (3) is also true.

Moreover, by using (2), (6) and Theorem 3.5, we can see that

$$p\big(n(x+y)\big)^2 = n^2 p(x+y)^2 = n^2 p(x)^2 + n^2 p(y)^2 + 2n^2 P_1(x, y)$$

and

$$\begin{aligned} p(nx+ny)^2 &= p(nx)^2 + p(nx)^2 + 2\,P_1(nx, ny) \\ &= n^2 p(x)^2 + n^2 p(y)^2 + 2n^2 P_1(x, y)\,. \end{aligned}$$

Therefore, $p\big(n(x+y)\big)^2 = p(nx+ny)^2$, and thus by the nonnegativity of p (4) also holds. □

Remark 4.4 If in particular X has a zero, the by Remark 3.6 we have $p(0)=0$, and thus also

$$p(0x) = |0|\,p(x) \qquad \text{and} \qquad p\big(0(x+y)\big) = p(0x+0y)$$

for all $x, y \in X$.

Moreover, if more specially X is a group, then by Remark 3.6 we have

$$p(kx) = |k|\,p(x) \qquad \text{and} \qquad p\big(k(x+y)\big) = p(kx+ky)$$

for all $k \in \mathbb{Z}$ and $x, y \in X$.

5 A weak Schwarz inequality

To prove a Schwarz type inequality for P, it is convenient to start with

Lemma 5.1 *For any $n, m \in \mathbb{N}$ and $x, y \in X$, we have*

$$p(nx+my)^2 = n^2\,p(x)^2 + m^2\,p(y)^2 + 2\,n\,m\,P_1(x, y)\,.$$

Proof. By Theorem 4.3 and Remark 3.7, we have

$$\begin{aligned} p(nx+my)^2 &= p(nx)^2 + p(my)^2 + 2\,P_1(nx,\ my) \\ &= n^2\,p(x)^2 + m^2\,p(y)^2 + 2\,n\,m\,P_1(x, y)\,. \end{aligned} \qquad \square$$

Now, by using this simple lemma, we can give two different proofs for the following theorem. The first one is more novel than the second one.

Theorem 5.2 *For any $x, y \in X$, we have*

$$-P_1(x, y) \le p(x)\,p(y)\,.$$

Proof 1. From Lemma 5.1, we can see that

$$-2\,P_1(x, y) \le (n/m)\,p(x)^2 + (m/n)\,p(y)^2\,.$$

for all $n, m \in \mathbb{N}$.

Therefore, by the definition of rational numbers, we actually have

$$-2\,P_1(x, y) \le r\,p(x)^2 + r^{-1}\,p(y)^2$$

for all $r \in \mathbb{Q}$ with $r > 0$.

Hence, by using that each real number is a limit of a sequence of rational numbers and the operation in $\mathbb{R}$ are continuous, we can already infer that

$$-2\,P_1(x,\,y) \leq \,\lambda\,p\,(x)^2 + \,\lambda^{-1}\,p\,(y)^2$$

for all $\lambda \in \mathbb{R}$ with $\lambda > 0$.

Now, by defining

$$f\,(\lambda) = \lambda\,p\,(x)^2 + \,\lambda^{-1}\,p\,(y)^2$$

for all $\lambda > 0$, we can state that

$$-2\,P_1(x,\,y) \leq \inf_{\lambda>0}\,f\,(\lambda)\,.$$

Moreover, if $p\,(x) \neq 0$ and $p\,(y) \neq 0$, then by taking

$$\lambda_0 = p\,(y)/p\,(x)$$

we can note that $\lambda_0 > 0$ such that

$$f\,(\lambda_0) = 2\,p\,(x)\,p\,(y)\,.$$

Therefore,

$$\inf_{\lambda>0}\,f\,(\lambda) \leq \,2\,p\,(x)\,p\,(y)\,, \qquad \text{and thus} \qquad -\,2\,P_1(x,\,y) \leq \,2\,p\,(x)\,p\,(y)\,.$$

Hence, the required inequality follows.

While, if either $p\,(x) = 0$ or $p\,(y) = 0$, then from the definition of f we can see that

$$\inf_{\lambda>0}\,f\,(\lambda) = 0\,, \qquad \text{and thus} \qquad -\,2\,P_1(x,\,y) \leq \,0\,.$$

Therefore, $-P_1(x,\,y) \leq 0$, and thus the required inequality trivially holds. □

Remark 5.3 If $p\,(x) \neq 0$ and $p\,(y) \neq 0$, then by computing $f'(\lambda)$ for all $\lambda > 0$, we can prove that $f\,(\lambda_0) < f\,(\lambda)$ for all $\lambda > 0$ with $\lambda \neq \lambda_0$.

Proof 2. From Lemma 5.1, we can also see that

$$0 \leq p\,(x)^2 + \,(m/n)^2\,p\,(y)^2 + \,2\,(m/n)\,P_1(x,\,y)$$

for all $n,\,m \in \mathbb{N}$.

Therefore, by using a similar argument as in Proof 1, we can state that

$$0 \leq p\,(x)^2 + \,\lambda^2\,p\,(y)^2 + \,2\,\lambda\,P_1(x,\,y)\,,$$

and thus

$$0 \leq p\,(x)^2 + \lambda\, P_1(x,\, y) + \lambda\,\big(\,\lambda\, p\,(y)^2 + P_1(x,\, y)\big)$$

for all $\lambda > 0$.

Hence, if $p\,(y) > 0$ and $P_1(x,\, y) < 0$, then by taking $\lambda = -P_1(x,\, y)/p\,(y)^2$ we can see that

$$0 \leq p\,(x)^2 - P_1(x,\, y)^2/\,p\,(y)^2\,, \qquad \text{and thus} \qquad P_1(x,\, y)^2 \leq \big(\,p\,(x)\,p\,(y)\big)^2.$$

Therefore, because of $|\,P_1(x,\, y)\,| = -P_1(x,\, y)$, the required inequality is also true.

While, if $p\,(y) = 0$ and $P_1(x,\, y) < 0$, then by taking $\lambda = -n\,P_1(x,\, y)$ for some $n \in \mathbb{N}$ we can see that

$$0 \leq p\,(x)^2 - 2\,n\,P_1(x,\, y)^2, \qquad \text{and thus} \qquad P_1(x,\, y)^2 \leq p\,(x)^2/\,2\,n\,.$$

Hence, by taking the limit $n \to \infty$, we can infer that $P_1(x,\, y) = 0$. Therefore, the required inequality trivially holds.

Now, to complete the proof, it remains only to note that if $P_1(x,\, y) \geq 0$, then the required inequality is also trivially true. □

From Theorem 5.2, we can easily infer the following

Corollary 5.4 *If in particular X is a group, then for any $x,\, y \in X$, we have*

$$|\,P_1(x,\, y)\,| \;\leq\; p\,(x)\,p\,(y)\,.$$

Proof. By Theorem 5.2 and Remarks 3.6 and 4.4, now we also have

$$P_1(x,\, y) = -P_1(-x,\, y) \leq p\,(-x)\,p\,(y) = p\,(x)\,p\,(y)\,.$$

Therefore, the required inequality is also true. □

Remark 5.5 Note that if $x,\, y \in X$ such that $|\,P\,(x,\, y)\,| \;\leq\; p\,(x)\,p\,(y)$ holds, then we also have $|\,P_i(x,\, y)\,| \;\leq\; p\,(x)\,p\,(y)$ and hence $P_i(x,\, y) \;\leq\; p\,(x)\,p\,(y)$ and $-P_i(x,\, y) \;\leq\; p\,(x)\,p\,(y)$ for $i = 1,\, 2$.

The following example shows that if in particular $X = \mathbb{R}^2$ and P is an $\mathbb{R}$–bihomogeneous semi-inner product on X, then even the weak Scwarz inequality $-P_2(x,\, y) \;\leq\; p\,(x)\,p\,(y)$ need not be true for all $x,\, y \in X$.

Example 5.6 For any $x,\, y \in \mathbb{R}^2$, define

$$a(x) = x \qquad \text{and} \qquad b(y) = (\, y_2,\, -y_1)\,,$$

and moreover

$$Q_1(x,\, y) = x_1\, y_1 \qquad \text{and} \qquad Q_2(x,\, y) = \langle\, a(x)\,,\; b(y)\,\rangle\,.$$

Then, $Q = (\, Q_1,\, Q_2)$ is an $\mathbb{R}$–bihogeneous semi-inner product on $\mathbb{R}^2$ such that, under the notation

$$q(x) = \sqrt{Q(x,\, x)}$$

with $x \in \mathbb{R}^2$, even the inequality

$$-\,Q_2(x,\, y) \;\leq\; q(x)\, q(y)$$

fails to hold for all $x,\, y \in \mathbb{R}^2$.

It is clear that Q_1 is a symmetric, bilinear function of $\big(\mathbb{R}^2\big)^2$ to $\mathbb{R}$. Moreover, we can easily see that a and b are linear functions of $\mathbb{R}^2$ to itself. Therefore, Q_2 is also a bilinear function of $\big(\mathbb{R}^2\big)^2$ to $\mathbb{R}$. Hence, it is clear that Q is a bilinear function of $\mathbb{R}^2$ to itself.

Moreover, since

$$Q_2(x,\, y) = \langle\, a(x)\,,\; b(y)\,\rangle = \langle\, (x_1,\, x_2)\,,\; (\, y_2,\, -y_1)\,\rangle = x_1 y_2 - x_2\, y_1$$

for all $x,\, y\, \in \mathbb{R}^2$, we can note that

$$Q_2(x,\, x) = 0 \qquad \text{and} \qquad Q_2(y,\, x\,) = -Q_2(y,\, x\,)$$

for all $x,\, y \in \mathbb{R}^2$. Hence, it is clear that Q is an $\mathbb{R}$–bihogeneous semi-inner product on $\mathbb{R}^2$.

On the other hand, for instance, by taking

$$u = (0,\, 1) \qquad \text{and} \qquad v = (1,\, 0)\,,$$

we can see that

$$q(u)\, q(v) = \,|\, u_1\,|\,|\, v_1\,| = 0\,, \qquad \text{but} \qquad -\,Q_2(u,\, v) = u_2\, v_1 - u_1\, v_2 = 1\,.$$

Remark 5.7 Note that, by making the plausible change

$$Q_1(x,\, y) = \langle\, x,\, y\,\rangle$$

for all $x,\, y \in \mathbb{R}^2$, we could get

$$\begin{aligned} |\, Q(x,\, y)\,|^2 = \; Q_1(x,\, y)^2 + Q_2(x,\, y)^2 &= \,(\, x_1\, y_1 + \, x_2\, y_2\,)^2 + \,(\, x_1 y_2 - x_2\, y_1\,)^2 \\ &= \,(\, x_1^2 + \, x_2^2\,)\,(\, y_1^2 + \, y_2^2\,) = \,|\, x\,|^2\,|\, y\,|^2 = q(x)^2\, q(y)^2\,, \end{aligned}$$

and thus $|\, Q(x,\, y)\,| \,=\, q(x)\, q(y)$ for all $x,\, y \in \mathbb{R}^2$.

However, it is now more important to note that, by using Corollary 5.4, we can give two different proofs for the subadditivity of p. The first one is more novel than the second one.

Theorem 5.8 *If in particular X is a group, then for any $x, y \in X$, we have*

(1) $p(x+y) \leq p(x) + p(y)$,

(2) $|p(x) - p(y)| \leq p(x-y)$.

Proof 1. By using Theorem 4.3 and the inequality $P_1(x, y) \leq p(x)\,p(y)$, we can see that

$$p(x+y)^2 = P_1(x+y, x) + P_1(x+y, y) \leq p(x+y)\,p(x) + p(x+y)\,p(y) .$$

Therefore, by the nonnegativity of p, inequality (1) is also true. □

Proof 2. By using Theorem 4.3 and the inequality $P_1(x, y) \leq p(x)\,p(y)$, we can also see that

$$\begin{aligned} p(x+y)^2 &= p(x)^2 + p(y)^2 + 2\,P_1(x, y) \\ &\leq p(x)^2 + p(y)^2 + 2\,p(x)\,p(y) = \big(p(x) + p(y)\big)^2. \end{aligned}$$

Therefore, by the nonnegativity of p, inequality (1) is also true. □

Remark 5.9 Theorems 4.3 and 5.8, together with Remark 4.4, show that if in particular X is a group, then p is already a seminorm on X in the sense it is an even, $\mathbb{N}$–homogeneous, subadditive function of X to $\mathbb{R}$.

Hence, it can be easily seen that, in this case, the function d, defined by

$$d(x, y) = p(-x+y)$$

for all $x, y \in X$, is a both left and right translation invariant semimetric on X.

In an improved and enlarged version of [3], we shall show that, analogously to seminorms and semimetrics derived from the usual semi-inner products on vector spaces, the generalized seminorms and semimetrics derived from semi-inner products on groupoids and groups also have several useful additional properties.

Acknowledgement. The authors are indebted to Professors Gyula Maksa and Jens Schwaiger for some helpful conversations and encouragements.

References

[1] **Baron, K. :** *On additive involutions and Hamel bases.* Aquationes Math. **87**, 159 – 163 (2014)

[2] **Boros, Z. :** *Schwarz inequality over groups.* Talk held at the Conference on Inequalities and Applications, Hajdúszoboszló, Hungary, 2016

[3] **Boros, Z.**, and **Száz, Á. :** *Semi-inner products and their induced seminorms and semimetrics on groups.* Tech. Rep., Inst. Math., Univ. Debrecen 2016/6, 11 pp.

[4] **Dragomir, S. S. :** *Semi-Inner Products and Applications.* Nova Science Publishers, Hauppauge, NY, 2004

[5] **Drygas, H. :** *Quasi-inner product and their applications.* In: Gupta, A. K. (Ed.), Advances in Multivariate Statistical Analysis, Theory Decis. Ser. B, Math. Statis. Methods, Reidel, Dordrecht, 1987, 13 – 30

[6] **Fechner, W. :** *Stability of a functional inequality associated with the Jordan-von Neumann functional equation.* Aequationes Math. **71**, 149 – 161 (2006)

[7] **Ger, R. :** *On a problem of Navid Safaei.* Talk held at the Conference on Inequalities and Applications, Hajdúszoboszló, Hungary, 2016

[8] **Gilányi, A. :** *Eine zur Parallelogrammgleichung äquivalente Ungleichung.* Aequationes Math. **62**, 303 – 309 (2001)

[9] **Giles, J. R. :** *Classes of semi-inner-product spaces.* Trans. Amer. Math, Soc. **129**, 436 – 446 (1967)

[10] **Glavosits, T.**, and **Száz, Á. :** *Constructions and extensions of free and controlled additive relations.* In: Rassias, Th. M. (Ed.), Handbook of Functional Equations: Functional Inequalities, Springer Optimization and Its Applications **95**, 161 – 208 (2014)

[11] **Lumer, G. :** *Semi-inner-product spaces.* Trans. Amer. Math, Soc. **100**, 29 – 43 (1961)

[12] **Kuczma, M. :** *An Introduction to the Theory of Functional Equations and Inequalities.* Państwowe Wydawnictwo Naukowe, Warszawa 1985

[13] **Makai, I. :** *Über invertierbare Lösungen der additive Cauchy-Functionalgleichung.* Publ. Math. Debrecen **16**, 239 – 243 (1969)

[14] **Maksa, Gy.**, and **Volkmann, P. :** *Characterizations of group homomorphisms having values in an inner product space.* Publ. Math. Debrecen **56**, 197 – 200 (2000)

[15] **Rätz, J. :** *On inequalities associated with the Jordan-von Neumann functional equation.* Aequationes Math. **66**, 191 – 200 (2003)

[16] **Száz, Á. :** *An instructive treatment of convergence, closure and orthogonality in semi-inner product spaces.* Tech. Rep., Inst. Math., Univ. Debrecen 2006/2, 29 pp.

[17] **Száz, Á. :** *Applications of fat and dense sets in the theory of additive functions.* Tech. Rep., Inst. Math., Univ. Debrecen 2007/3, 29 pp.

[18] **Száz, Á. :** *A natural Galois connection between generalized norms and metrics.* Tech. Rep., Inst. Math., Univ. Debrecen 2016/4, 9 pp.

[19] **Száz, Á. :** *Generalizations of a theorem of Maksa and Volkmann on additive functions.* Tech. Rep., Inst. Math., Univ. Debrecen 2016/5, 6 pp. (An improved and enlarged version is available from the author.)

[20] **Száz, Á. :** *Remarks and problems at the Conference on Inequalities and Applications, Hajdúszoboszló, Hungary, 2016.* Tech. Rep., Inst. Math., Univ. Debrecen 2016/9, 34 pp.

[21] **Stetkaer, H. :** *Functional Equations on Groups.* World Scientific, New Jersey 2013

received: September 27, 2016

Authors:

Zoltán Boros
Department of Mathematics,
University of Debrecen,
H-4002 Debrecen,
Pf. 400, Hungary

e-mail: zboros@science.unideb.hu

Árpád Száz
Department of Mathematics,
University of Debrecen,
H-4002 Debrecen,
Pf. 400, Hungary

e-mail: szaz@science.unideb.hu

Rostock. Math. Kolloq. **71**, 41 – 56 (2017/18)

Subject Classification (AMS)
35J61, 35J10

Bénédicte Alziary, Jacqueline Fleckinger

Semi-linear cooperative elliptic systems involving Schrödinger operators: Groundstate positivity or negativity

ABSTRACT. We study here the behavior of the solutions to a 2×2 semi-linear cooperative system involving Schrödinger operators (considered in its variational form):

$$LU := (-\Delta + q(x))U = AU + \mu U + F(x, U) \text{ in } \mathbb{R}^N$$

$$U(x)_{|x| \to \infty} \to 0$$

where q is a continuous positive potential tending to $+\infty$ at infinity; μ is a real parameter varying near the principal eigenvalue of the system; U is a column vector with components u_1 and u_2 and A is a square cooperative matrix with constant coefficient. F is a column vector with components f_1 and f_2 depending eventually on U.

1 Introduction

We study here the behaviour of the solutions to a 2×2 semi-linear cooperative system involving Schrödinger operators (considered in its variational form):

$$LU := (-\Delta + q(x))U = AU + \mu U + F(x, U) \text{ in } \mathbb{R}^N$$

$$U(x)_{|x| \to \infty} \to 0$$

where q is a continuous positive potential tending to $+\infty$ at infinity; U is a column vector with components u_1 and u_2 and A is a square matrix with constant coefficients; moreover A is a cooperative matrix (which means that its coefficients outside the diagonal are non negative). F is a column vector with components f_1 and f_2 depending eventually on U. The real parameter μ varies near the principal eigenvalue of the system and plays a key role. According to its position it determines not only the sign of the solutions but also their position w.r.t. the groundstate.

Such systems have been intensively studied (very often for $\mu = 0$) and mainly for Dirichlet problems defined on bounded domains ([16], [17], [18], [21], [20], [25], [12], [4]). When the whole $\mathbb{R}^N$ is considered, as here, 2 cases are generally studied: either "Schrödinger systems" ([1], [2], [3], [7]), that is system involving Schrödinger operators, as here, or systems with a weight tending to 0 ([23], [6]). It is also possible to consider a combination of these 2 problems with a potential q and a weight g :

$$LU := (-\Delta + q(x))U = g(x)AU + \mu g(x)U + F(x, U) \ \text{ in } \ \mathbb{R}^N$$

as far as $\dfrac{g}{q}$ tends to 0 at infinity which is the condition for having some compactness and therefore a discrete spectrum.

The first results on Schrödinger systems, when F does not depend on U (linear systems) deal with cooperative systems and with the Maximum Principle (**MP**) that is:

"If the data F is non negative, $\neq 0$, then, any solution U is non negative".

As for the case of one equation, this Maximum Principle holds for a parameter $\mu < \Lambda^*$, where Λ^* is the principal eigenvalue of the system, which means that $LU - AU - \Lambda^* U = 0$ has a non zero solution which does not change sign.

For the classical case of an equation defined on a bounded domain with zero boundary conditions, $-\Delta u = \mu u + f(x)$, $f > 0$, Clément and Peletier [14] have shown that the solution u changes sign as soon as μ goes over λ_1, the first eigenvalue of the Dirichlet Laplacian defined on Ω. More precisely there exists a small positive δ, depending on f, such that for all $\mu \in (\lambda_1, \lambda_1 + \delta)$, $u < 0$. This phenomenon is known as "Anti-maximum Principle" (**AMP**).

In our present case, where we have no boundary, we have improved these results giving not only the sign of the solutions but also comparing the solutions with the groundstate (principal eigenfunction); it is what we call "groundstate positivity"(**GSP**) (resp. negativity) (resp. **GSN**). We extend in particular previous results established in [5] for linear systems to some semi-linear cooperative systems. For being not excessively technical, we limit our study to radial potentials and cooperative systems. Extensions to more general cases will appear somewhere else.

Our paper is organized as follows:

We recall first some previous results of the linear case that we use. Then we study a semi-linear equation. Finally we study a cooperative semi-linear system.

2 Linear Case: one equation

We shortly recall the case of a linear equation with a parameter μ varying near the principal eigenvalue of the operator.

$$(E) \qquad Lu := (-\Delta + q(x))u \;=\; \mu u + f(x) \text{ in } \mathbb{R}^N,$$

$$\lim_{|x|\to+\infty} u(x) \;=\; 0.$$

(H_q) q is a positive continuous potential tending to $+\infty$ at infinity.

We seek u in V where

$$V := \left\{ u \in L^2(\mathbb{R}^N)\, s.t.\ \|u\|_V = \left(\int |\nabla u|^2 + q(x)u^2 \right)^{1/2} < \infty \right\}.$$

If (H_q) is satisfied, the embedding of V into $L^2(\mathbb{R}^N)$ is compact (see *e.g.* [19], [15]). Hence L possesses an infinity of eigenvalues tending to $+\infty$:

$$0 < \lambda_1 < \lambda_2 \leq .... \leq \lambda_k \leq ... ,\ \lambda_k \to +\infty \text{ as } k \to \infty.$$

Notation (Λ, ϕ)**:** We set from now on $\Lambda := \lambda_1$ the smallest one (which is positive and simple) and ϕ the associated eigenfunction, positive and with L^2-norm $\|\phi\| = 1$.

It is classical (see *e.g.* [24]) that if $f \geq 0, \neq 0$, and $\mu < \Lambda$, there exists exactly one solution which is positive: the positivity is "improved", or in other words, the (strong) maximum principle **(MP)** is satisfied:

$$(MP) \qquad f \geq 0, \not\equiv 0 \;\Rightarrow\; u > 0.$$

Lately, as said above, another notion has been defined ([8], [10], [22]) the "groundstate positivity" (**GSP**) (resp. "negativity" (**GSN**)) which means that, there exists $k > 0$ such that the solution $u > k\phi$ (GSP) (resp. $u < -k\phi$ (GSN)).

We also say shortly "fundamental positivity" or "negativity", or also "ϕ-positivity" or "negativity". Indeed these properties are more precise than MP or AMP. But for proving them, it is necessary to have a potential growing fast enough, a potential with a super quadratic growth.

In [10] a class $\mathcal{P}$ of radial potentials is defined:

$$\mathcal{P} := \left\{ Q \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}_+^*) / \exists R_0 > 0, Q' > 0\, a.e. \text{ on } [R_0, \infty),\ \int_{R_0}^{\infty} Q(r)^{-1/2} < \infty \right\}. \tag{1}$$

The last inequality holds precisely if Q is growing sufficiently fast, indeed faster than r^2 (the harmonic oscillator). In this paper we consider only a radial potential $q \in \mathcal{P}$. Note that our proof is valid for more general potentials, in particular for perturbations of radial potential [9] or [10] . We assume here

(H'_q) q is radial and is in $\mathcal{P}$

Remark 1 Note that since q is in $\mathcal{P}$ it satisfies (H_q).

On f we assume

(H^*_f) $f \in L^2(\mathbb{R}^N), \quad f^1 = \int f\phi > 0.$

For having more precise estimates on u, in particular the "groundstate negativity" **(GSN)**, we have to define another set X in which f varies, the set of "groundstate bounded functions":

$$X := \{h \in L^2(\mathbb{R}^N) : |h|/\phi \in L^\infty(\mathbb{R}^N)\}, \tag{2}$$

equipped with the norm $\|h\|_X = ess \sup_{\mathbb{R}^n}(|h|/\phi)$.

Theorem 1 *Assume (H'_q) and (H^*_f), $f \in X$. For $\mu < \Lambda$ or $\Lambda < \mu < \lambda_2$ there exists $\delta > 0$ (defined below) depending on f and a positive constant C, depending on f such that if $0 < |\Lambda - \mu| < \delta$,*

$$\Lambda - \delta < \mu < \Lambda \;\Rightarrow\; u \geq \frac{C}{\Lambda - \mu}\phi > 0,$$

$$\Lambda < \mu < \Lambda + \delta \;\Rightarrow\; u \leq \frac{C}{\Lambda - \mu}\phi < 0.$$

Proof of Theorem 1 Decompose now u and f in (E) on ϕ and its orthogonal:

$$u = u^1\phi + u^\perp\,;\; f = f^1\phi + f^\perp;\; u^1 = \int u\phi,\; \int u^\perp\phi = \int f^\perp\phi, = 0;$$

we derive from Equation (E)

$$(L-\mu)u^1\phi = (\Lambda - \mu)u^1\phi = f^1\phi\,,\; Lu^\perp = \mu u^\perp + f^\perp. \tag{3}$$

Choose $\mu < \Lambda$ or $\Lambda < \mu < \lambda_2$. From the first equation we derive

$$u^1 = \frac{f^1}{(\Lambda - \mu)} \to \pm\infty \;\text{ as }\; (\Lambda - \mu) \to 0.$$

By use of Theorem 3.2 (c) in [9] or [10], we know that the restriction of the resolvent $(L-\mu)^{-1}$ to X is bounded from X into itself. The following lemma is a direct consequence of this result as it is shown in the proof of the Theorem 3.4 in [9].

Lemma 1 *There exists δ_0 small enough and there exists a constant c_0 (depending on δ_0) such that for all μ with $\Lambda - \delta_0 < \mu < \Lambda$ or $\Lambda < \mu < \Lambda + \delta_0 < \lambda_2$,*

$$-c_0 \|f^\perp\|_X \leq \|u^\perp\|_X \leq c_0 \|f^\perp\|_X.$$

Finally we take in account Lemma 1 and (3):

$$\|u^\perp\|_X \leq c_0 \|f^\perp\|_X \text{ and } u = \frac{f^1}{\Lambda - \mu}\phi + u^\perp;$$

for $|\Lambda - \mu| \to 0$, $\frac{f^1}{\Lambda-\mu}\phi \to \pm\infty$ when $u^\perp$ stays bounded. Hence, for $|\Lambda - \mu|$ small enough, more precisely for $|\Lambda - \mu| < \delta_1(f) := \frac{f^1}{c_0\|f^\perp\|_X}$, we have

$$\frac{f^1}{|\Lambda - \mu|} > c_0 \|f^\perp\|_X.$$

We deduce that Theorem 1 is valid for $\delta := \min\{\delta_0, \delta_1(f)\}$.

3 Semi-linear Schrödinger equation

We study now the case of a semi-linear equation. We first obtain bounds for the solutions, if they exist and then we show their existence via the method of "sub-super solutions". Finally, with additional assumptions, we prove the uniqueness of them.

Consider the semi-linear Schrödinger equation (SLSE)

$$(SLSE) \qquad Lu := (-\Delta + q(x))u \; = \; \mu u + f(x,u) \text{ in } \mathbb{R}^N,$$

$$\lim_{|x|\to+\infty} u(x) \; = \; 0.$$

We assume that the potential q satisfies (H_q') and we denote as above by (Λ, ϕ) the principal eigenpair with $\phi > 0$.

We work in $L^2(\mathbb{R}^N)$ and we consider the problem in its variational formulation. We seek u in V for a suitable f.

We assume that f satisfies :

(H_f) $f : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ is a Caratheodory function *i.e.* the function $f(\bullet, u)$ is Lebesgue measurable in $\mathbb{R}^N$, for every $u(x) \in \mathbb{R}$ and the function $f(x, \bullet)$ is continuous in $\mathbb{R}$ for almost every $x \in \mathbb{R}^N$. Moreover, f is such that

$$(i) \qquad \forall u \in L^2(\mathbb{R}^N), \;\; f(.,u) \in L^2(\mathbb{R}^N),$$

$$(ii) \qquad \exists\, \kappa > 0 \quad s.t. \quad \forall u \, \in \, V, \;\; f(x,u) \, \geq \, \kappa\phi(x) > 0$$

$$(iii) \qquad \exists K > \kappa > 0 \quad s.t. \quad \forall u \in V, \;\; f(x,u) \leq K\phi(x).$$

Later we also suppose

$$(H'_f) \qquad \forall x \in \mathbb{R}^N, \; u \to \frac{f(x,u)}{|u|} \text{ is strictly decreasing}$$

Remark 2 Note that, by (ii) and (iii), for any $u \in V$, $f(.,u) \in X$ and hence the solutions, if they exist, are in X.

Let a parameter μ be given, with $|\mu - \Lambda|$ "small enough". In this section we prove groundstate positivity and negativity for the semi-linear Schrödinger equation.

Theorem 2 *If (H'_q) and (H_f) are satisfied , then there exists $\delta(f) > 0$ ($\delta = \delta(f) :=$ $\min\{\delta_0, \delta'_1(f) := \frac{\kappa}{c_0 K}\}$ where δ_0 and c_0 are given in Lemma 1) such that, for $0 < |\mu - \Lambda| < \delta$ there exists a solution u to $(SLES)$ such that*

$$\|u\|_X \leq \frac{K}{|\Lambda - \mu|} + 2c_0 K.$$

Also

- for $\Lambda - \delta < \mu < \Lambda$, $u > \frac{\kappa}{\Lambda - \mu}\phi > 0$,

- for $\Lambda < \mu < \Lambda + \delta < \lambda_2$, $u < \frac{K}{\Lambda - \mu}\phi < 0$.

Moreover if (H'_f) is satisfied, the solution to $(SLSE)$ is unique.

Remark 3 If (ii) does not hold, for $\mu < \Lambda$, there exists a solution u such that

$$\|u\|_X \leq \frac{K}{|\Lambda - \mu|} + 2c_0 K.$$

The existence is classical ($e.g.$ [3]) and the estimate follows from the proof below.

Proof of Theorem 2

We do the proof in 3 steps: first maximun and anti-maximum principles, secondly existence of the solution such that $u > \frac{\kappa}{\Lambda - \mu}\phi > 0$ for $\Lambda - \delta < \mu < \Lambda$ and such that $u < \frac{K}{\Lambda - \mu}\phi < 0$, for $\Lambda < \mu < \Lambda + \delta$, and thirdly the uniqueness.

Step 1. Maximun and anti-maximum principles

We prove the positivity or negativity of the solutions exactly as for the linear case, but, since f depends on u we have to show that δ (which depends on f in the linear case) is now uniform. This follows from hypotheses (ii) and (iii).

Let u be a solution to $Lu = \mu u + f(x,u)$. For this u, set

$$f^1(u) = \int f(x,u)\phi(x)dx\,,\ f^{\perp}(x,u) = f(x,u) - f^1(u)\phi(x).$$

Also $u^1 = \int u\phi(x)dx$ and $u^{\perp} = u - u^1\phi$.

Note that, always by (ii) and (iii), $0 < \kappa \leq f^1(u) \leq K$.

With this decomposition, reporting in $(SLSE)$, we obtain 2 equations:

$$(L-\mu)u^1\phi = (\Lambda - \mu)u^1\phi = f^1\phi\,,\ \ Lu^{\perp} = \mu u^{\perp} + f^{\perp}.$$

Choose $\mu < \Lambda$ or $\Lambda < \mu < \lambda_2$. From the first equation we derive

$$u^1 = \frac{f^1}{(\Lambda-\mu)} \to \pm\infty\, as\, (\Lambda-\mu) \to 0.$$

Now we proceed exactly as for the linear case. By use of Theorem 3.2 (c) in [9] or [10], we know that the restriction of the resolvent $(L-\mu)^{-1}$ to X is bounded from X into itself. So by (iii) and by Lemma 1 there exists a δ_0 small enough and there exists a constant c_0 (depending on δ_0) such that for all μ with $|\Lambda - \mu| < \delta_0$,

$$\|u^{\perp}\|_X \leq c_0\|f^{\perp}(x,u)\|_X \leq c_0\|f(x,u) - f^1(u)\phi(x)\|_X \leq 2c_0K.$$

Write now

$$u = \frac{f^1(u)}{\Lambda-\mu}\phi + u^{\perp}$$

Hence $\|u\|_X \leq \frac{f^1(u)}{|\Lambda-\mu|} + \|u^{\perp}\|_X \leq \frac{K}{|\Lambda-\mu|} + 2c_0K$. For $|\Lambda-\mu| \to 0$, $\frac{f^1}{\Lambda-\mu}\phi \to \pm\infty$ when $u^{\perp}$ stays bounded. For $|\Lambda-\mu|$ small enough, that is here $|\Lambda-\mu| < \delta_1'(f) := \frac{\kappa}{2c_0K}$, we get (since $f^1 > 0$)

$$\frac{f^1}{|\Lambda-\mu|} \geq \frac{\kappa}{|\Lambda-\mu|} > 2c_0K \geq c_0\|f^{\perp}\|_X.$$

Finally Maximum and anti-maximum principles are valid for
$\delta(f) := \min\{\delta_0, \delta_1'(f)\}$.

Step 2. Existence of solutions

We prove the existence of solutions by Schauder fixed point theory; for this purpose we need some classical elements: a set $\mathcal{K}^{\pm}$ constructed with the help of sub-super solutions and a compact operator T acting in $\mathcal{K}^{\pm}$ such that $\mathcal{K}^{\pm}$ stays invariant by T: $T(\mathcal{K}^{\pm}) \subset \mathcal{K}^{\pm}$.

1: "Sub-super solution" :

- Case $\Lambda - \delta < \mu < \Lambda$.

Obviously, by (ii), $u_0 = \frac{\kappa}{\Lambda-\mu}\phi > 0$ is a subsolution:

$$L(u-u_0) = \mu(u-u_0) + f - (\Lambda-\mu)u_0 = \mu(u-u_0) + f - \kappa\phi$$

and by (ii) and GSP, $u - u_0 \geq 0$.

Analogously $u^0 = \frac{K}{\Lambda-\mu}\phi > 0$ (K given in (iii)) is a supersolution :

$$Lu^0 = \frac{\Lambda}{\Lambda-\mu}K\phi = \Lambda u^0 = \mu u^0 + (\Lambda-\mu)u^0.$$

Remark 4 The sub- and supersolutions tend to $+\infty$ as $\mu \nearrow \Lambda$.

- Case $\Lambda < \mu < \Lambda + \delta < \lambda_2$. $v^0 = \frac{\kappa}{\Lambda-\mu}\phi < 0$ is a supersolution. Indeed

$$L(v^0-u) = \mu(v^0-u) + \kappa\phi - f$$

and by (H_f) and the anti-maximum $0 > v^0 \geq u$.

Analogously, $v_0 = \frac{K}{\Lambda-\mu}\phi < 0$ is a subsolution.

Remark 5 The sub- and supersolutions tend to $-\infty$ as $\mu \searrow \Lambda$.

Remark 6 Obviously, $u_0 < u^0$ for $\Lambda - \delta < \mu < \Lambda$ (resp. $v_0 < v^0$ for $\Lambda < \mu < \Lambda + \delta$).

2: The operator T

We define $T : u \in L^2 \longrightarrow w = Tu \in V$, where $w \in X$ is the unique solution to $Lw = \mu w + f(x,u)$.

3: The invariant set $\mathcal{K}^+ := [u_0, u^0]$ for $\Lambda - \delta < \mu < \Lambda$ (resp. $\mathcal{K}^- := [v_0, v^0]$ for $\Lambda < \mu < \Lambda + \delta$).

If $\mu < \Lambda$, by the maximum principle and the hypothesis (iii) , $u \leq u^0$ implies $w \leq u^0$. Indeed,

$$L(u^0-w) = \mu(u^0-w) + (\Lambda-\mu)u^0 - f(x,u) = \mu(u^0-w) + K\phi - f(x,u);$$

since, by (iii), $K\phi - f(x,u) \geq 0$, we apply the maximum principle and hence $w \leq u^0$. The 3 other cases lead to analogous calculation.

4: T is compact in X.

First note that $\mathcal{K}^+ \subset X$ (resp. $\mathcal{K}^- \subset X$). $Lw - \mu w = f(x,u)$ can also be written $w = (L-\mu I)^{-1}f(x,u) = T(u)$. Since by [10], [9], the resolvent $R(\mu) := (L-\mu I)^{-1}$ is compact in X for $\mu \in (\Lambda-\delta, \Lambda)$ or $(\Lambda, \Lambda+\delta)$, and since $F : u \to f(x,u)$ is continuous, $T = R(\mu)F$ is compact.

We deduce from Schauder fixed point theory that there exists a solution to $(SLSE)$ in $\mathcal{K}^+$, (resp. in $\mathcal{K}^-$).

Step 3. Uniqueness

For proving uniqueness we follow [13], p. 57. First we assume not only (H_f) but also (H'_f). Assume that u and v are two solutions:

$$Lu = \mu u + f(x,u)\ ,\ \ Lv = \mu v + f(x,v)$$

The solutions are in X and we have shown that $u, v > u_0 > 0$ for $\Lambda - \delta < \mu < \Lambda$ (resp. $u, v < v^0 < 0$ for $\Lambda < \mu < \Lambda + \delta$). Hence we can write

$$\frac{Lu}{u} = \mu + \frac{f(x,u)}{u}\ ;\ \frac{Lv}{v} = \mu + \frac{f(x,v)}{v}.$$

By subtraction $q(x)$ and μ disappear. Multiply by $u^2 - v^2$ and integrate.

$$\int \left[\frac{-\Delta u}{u} + \frac{\Delta v}{v}\right][u^2 - v^2] \ = \int \left[\frac{f(x,u)}{u} - \frac{f(x,v)}{v}\right][u^2 - v^2];$$

the last term is non positive by (H'_f).

We transform exactly as in [13] the first term.

$$\int \left[\frac{-\Delta u}{u} + \frac{\Delta v}{v}\right][u^2 - v^2] \ = \int \left|\nabla u - \frac{u}{v}\nabla v\right|^2 + \left|\nabla v - \frac{v}{u}\nabla u\right|^2 =$$

$$\int \left|v\nabla(\frac{u}{v})\right|^2 + \left|u\nabla(\frac{v}{u})\right|^2 \geq 0; \tag{4}$$

therefore both terms are equal to 0 and

$$u^2 - v^2 = 0 \ \Rightarrow \ u = v\, a.e.;$$

by regularity, $u = v$.

4 Semi-linear cooperative system

We extend here to a class of semi-linear systems previous results shown in [5] where linear systems of the form $LU = \mu U + AU + F(x)$ are studied.

We study for $a > 0$, $b > 0$, $c > 0$

$$(S) \qquad \left\{ \begin{array}{rcl} Lu_1 &=& (\mu + a)u_1 + bu_2 + f_1(x,u_1) \\ Lu_2 &=& cu_1 + (\mu + d)u_2 + f_2(x,u_2) \end{array} \right. \quad in\ \mathbb{R}^N, .$$

$$u_1(x), u_2(x)_{|x|\to\infty} \to 0.$$

We write shortly $LU = \mu U + AU + F(x,U)$, where A is the cooperative matrix with components a, b, c, d:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Notation (ξ_1, Y)**:** Denote ξ_1 the largest eigenvalue of A (the other one being denoted by ξ_2); Y is the eigenvector associated with ξ_1:

$$AY = \xi_1 Y.$$

$$\xi_1 = \frac{a+d+\sqrt{(a-d)^2+4bc}}{2}.$$

An easy calculation shows that $(L-A)(Y\phi) = (\Lambda - \xi_1)Y\phi$; moreover here $Y\phi$ is with components which do not change sign: we choose both components of Y positive:

$$y_1 = b > 0\,,\ y_2 = \frac{d-a+\sqrt{(a-d)^2+4bc}}{2} > 0.$$

Notation Λ^***:** $\Lambda^* := \Lambda - \xi_1$ is the principal eigenvalue of System (S) with associated eigenvector $Y\phi$:

$$(L-A)(Y\phi) = (\Lambda - \xi_1)Y\phi = \Lambda^* Y\phi.$$

Hypotheses We assume

(H_A) A is a 2×2 cooperative matrix with positive coefficients outside the diagonal.

(H_F) : $f_1, f_2 : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ are Caratheodory function *i.e.* the functions $f_1(\bullet, u_1)$ or $f_2(\bullet, u_2)$ are Lebesgue measurable in $\mathbb{R}^N$, for every $u_1(x)$ or $u_2(x)$ in $\mathbb{R}$ and the functions $f_1(x,\bullet)$, $f_2(x,\bullet)$ are continuous in $\mathbb{R}$ for almost every $x \in \mathbb{R}^N$. Moreover, f_1, f_2 are such that

(i) $$\forall u_1, u_2 \in L^2(\mathbb{R}^N),\ f_1(x,u_1), f_2(x,u_2) \in L^2(R^N),$$

(ii) $$\exists\, \kappa > 0 \ s.t.\ f_1(x,u_1), f_2(x,u_2) \geq \kappa\phi(x)\ \forall u_1, u_2 \in L^2(\mathbb{R}^N),$$

(iii) $$\exists K > \kappa > 0 \ s.t.\ f_1(x,u_1), f_2(x,u_2) \leq K\phi(x)\ \forall u_1, u_2 \in L^2(\mathbb{R}^N).$$

(H'_F) : $\frac{f_1(x,u_1)}{|u_1|}$ and $\frac{f_2(x,u_2)}{|u_2|}$ are decreasing w.r.t. u_1 and u_2.

We introduce 2 sets :

$$\mathcal{K_S}^+ := \left\{ (u_1,u_2) \in X^2 \,/\, u_1 \in \left(\frac{\kappa y_1 \phi}{\max(y_1,y_2)(\Lambda^*-\mu)}\,, \frac{K y_1 \phi}{\min(y_1,y_2)(\Lambda^*-\mu)} \right), \right.$$
$$\left. u_2 \in \left(\frac{\kappa y_2 \phi}{\max(y_1,y_2)(\Lambda^*-\mu)}\,, \frac{K y_2 \phi}{\min(y_1,y_2)(\Lambda^*-\mu)} \right) \right\}$$

for $\mu < \Lambda^*$, and

$$\mathcal{K}_{\mathcal{S}}{}^{-} := \left\{ (u_1, u_2) \in X^2 \,/\, u_1 \in \left(\frac{K y_1 \phi}{\min(y_1,y_2)(\Lambda^* - \mu)} , \frac{\kappa y_1 \phi}{\max(y_1,y_2)(\Lambda^* - \mu)} \right), \right.$$
$$\left. u_2 \in \left(\frac{K y_2 \phi}{\min(y_1,y_2)(\Lambda^* - \mu)} , \frac{\kappa y_2 \phi}{\max(y_1,y_2)(\Lambda^* - \mu)} \right) \right\}$$

for $\Lambda^* < \mu$.

Theorem 3 *If (H_A) and (H_F) are satisfied there exists $\delta > 0$, depending on f_1 and f_2 such that if $\Lambda^* - \delta < \mu < \Lambda^*$ (resp. $\Lambda^* < \mu < \Lambda^* + \delta$), (with $\delta < min\{\frac{\xi_2 - \xi_1}{2}, \lambda_2 - \Lambda\}$) System (S) has a solution which is in $\mathcal{K}_{\mathcal{S}}^{+}$, (resp. in $\mathcal{K}_{\mathcal{S}}^{-}$). Moreover, if (H'_F) is satisfied, the solution is unique.*

Proof of Theorem 3 We use of course the results above as well as previous results for linear systems obtained in [5] where Theorem 3 is shown for suitable assumptions on f_1 and f_2 (independent on u).

1. Maximun and anti-maximum principles

We diagonalize System(S) thanks to the change of basis matrix P, and we get a system of 2 equations. Here

$$P = \begin{pmatrix} b & b \\ \xi_1 - a & \xi_2 - a \end{pmatrix}, \; P^{-1} = \frac{1}{b(\xi_1 - \xi_2)} \begin{pmatrix} a - \xi_2 & b \\ \xi_1 - a & -b \end{pmatrix},$$

Set

$$D := P^{-1}AP = \begin{pmatrix} \xi_1 & 0 \\ 0 & \xi_2 \end{pmatrix} ; \; U = PV \,; \; G := P^{-1}F. \tag{5}$$

We obtain

$$LV = DV + \mu V + G \tag{6}$$

which is a system of 2 equations (with obvious notation):

$$Lv_1 = (\xi_1 + \mu)v_1 + g_1(u_1, u_2);$$

$$Lv_2 = (\xi_2 + \mu)v_2 + g_2(u_1, u_2).$$

Note that g_1 and g_2 are in X.

The second equation, where the parameter $\xi_2 + \mu$ stays away (below) from Λ, has a ϕ bounded solution v_2. Concerning the first equation, we apply Theorem 2 above. We compute g_1, g_2 and get *(ii′)* $$\exists\,\kappa' > 0 \; s.t.\; g_1(x,u_1,u_2) \;\geq\; \kappa'\phi(x) \;\forall u_1,u_2 \;\in\; L^2(\mathbb{R}^N),$$

(iii′) $$\exists K' > \kappa' > 0 \; s.t.\; g_1(x,u_1,u_2),\, |g_2(x,u_1,u_2)| \;\leq\; K'\phi(x) \;\forall u_1,u_2 \;\in\; L^2(\mathbb{R}^N),$$

where κ' and K' are 2 positive constants depending on κ, K and on the coefficients of A. This follows from $\xi_1-\xi_2>0$ and $(a-\xi_2)=\frac{a-d}{2}+\frac{\sqrt{(a-d)^2+4bc}}{2}$ with $(a-d)^2+4bc>(a-d)^2$, so that

$$g_1=\frac{1}{\xi_1-\xi_2}[(a-\xi_2)f_1+bf_2]>\kappa'\phi>0.$$

Analoguously we have $g_1<K'\phi$. Therefore Theorem 2 holds here with $\delta=\min(\delta_0,\frac{\kappa'}{c_0K'},\frac{\xi_1-\xi_2}{2})$. Finally we deduce from the maximum principle for $\Lambda^*-\delta<\mu<\Lambda^*$ that $v_1>\frac{\kappa'}{\Lambda^*-\mu}\phi>0$.

If $\Lambda^*<\mu<\Lambda^*+\delta$, reasoning similarly, we deduce $v_1<\frac{K'}{\Lambda^*-\mu}\phi<0$. As $\mu\to\Lambda^*$, v_1 tends to ∞ when v_2 stays bounded. Indeed, by Remark 3,

$$\|v_2\|_X\leq\frac{K'}{|\Lambda-\xi_2-\mu|}+2c_0K'<\frac{2K'}{\xi_1-\xi_2}+2c_0K';$$

the last inequality follows from $\delta<\frac{\xi_1-\xi_2}{2}$.

Now we go back to $U=PV$.

$$u_1=av_1+bv_2\,,\;\; u_2=(\xi_1-a)v_1+(\xi_2-a)v_2.$$

Combining the estimates above on v_1 and v_2, we conclude that, as $|\Lambda^*-\mu|\to0$, there exists δ^*, depending only on L,A,κ,K such that as $\mu\nearrow\Lambda^*$, u_1 has the sign of $a>0$ and $u_2>0$. If $\mu\searrow\Lambda^*$, u_1 has the sign of $-a<0$ and $u_2<0$.

2. Existence of the solution in $\mathcal{K}_{\mathcal{S}}^+$, (resp. in $\mathcal{K}_{\mathcal{S}}^-$)

Sub-supersolutions:

1. Case $\Lambda^*-\delta^*<\mu<\Lambda^*$. Recall that Y has positive components y_1 and y_2, and the principal eigenvector $\Phi=Y\phi$ satisfies

$$L\Phi-\mu\Phi-A\Phi=(\Lambda^*-\mu)\Phi.$$

Inspired by the case of one equation, we seek a subsolution U_0 of the form $cY\Phi$.

$$L(U-U_0)=A(U-U_0)+\mu(U-U_0)+(F(x,U)-(\Lambda^*-\mu)c\Phi).$$

For c such that $F(x,U)-(\Lambda^*-\mu)cY\phi(x)>0$, for $\mu<\Lambda^*$, we get $U-U_0>0$ by the maximum principle. Finally, since $F(x,U)-\dfrac{\kappa}{\max(y_1,y_2)}Y\phi>0$, a subsolution is

$$U_0=\frac{\kappa}{\max(y_1,y_2)}\frac{1}{(\Lambda^*-\mu)}Y\phi.$$

Analogously $U^0 = \dfrac{K}{\min(y_1, y_2)(\Lambda^* - \mu)} Y\phi$ is a supersolution.

2. Case $\Lambda^* < \mu < \Lambda^* + \delta^*$. We have similar results with change of sign and replacing K by κ.

$$V_0 = \frac{K}{\min(y_1, y_2)(\Lambda^* - \mu)} Y\phi$$
$$V^0 = \frac{\kappa}{\max(y_1, y_2)} \frac{1}{(\Lambda^* - \mu)} Y\phi$$

The operator T: We define $T : (u_1, u_2) \longrightarrow (w_1, w_2)$ where (w_1, w_2) is the solution to the linear system

$$(S') \qquad \begin{cases} Lw_1 &= (a+\mu)w_1 + bw_2 + f_1(x, u_1) \\ Lw_2 &= cw_1 + (d+\mu)w_2 + f_2(x, u_2) \end{cases} \quad in\ \mathbb{R}^N, .$$

$$w_1(x), w_2(x)_{|x| \to \infty} \to 0.$$

The rectangle: If $(u_1, u_2) \in \mathcal{K}^+_{\mathcal{S}}$ for $\Lambda^* - \delta^* < \mu < \Lambda^*$ (resp. $(u_1, u_2) \in \mathcal{K}^-_{\mathcal{S}}$ for $\Lambda^* < \mu < \Lambda^* + \delta^*$) then $(w_1, w_2) \in \mathcal{K}^+_{\mathcal{S}}$ (resp $\mathcal{K}^-_{\mathcal{S}}$). Indeed, for $\Lambda^* - \delta^* \mu < \Lambda^*$, this can be written with obvious notations

$$L(W - U_0) = (\mu + A)(W - U_0) + F;$$

for $\mu < \Lambda^*$, since F has non negative components, $F \not\equiv 0$, then $W - U_0 > 0$. Analogously, we obtain the supersolution $U^0 - W > 0$.

We argue exactly as for one equation: $\mathcal{K}^+_{\mathcal{S}}$ or $\mathcal{K}^-_{\mathcal{S}}$ is invariant by T and $LW = (A + \mu)W + F(x, U)$ can be written $W = (L - A - \mu I)^{-1} \hat{F}(x, u) = T(U)$. Since by [10], [9], the resolvent $R(\mu) := (L - \mu I)^{-1}$ is compact in X for $\mu \in (\Lambda^* - \delta^*, \Lambda^*)$ or $(\Lambda^*, \Lambda^* + \delta^*)$, and since $\hat{F} : u \to F(x, u)$ is continuous, $T = R(\mu)\hat{F}$ is compact.

We apply the fixed point theorem. There exists a solution U.

3. Uniqueness

We assume now (H'_F). assume there are 2 positive solutions (u_1, u_2) and (v_1, v_2) to (S); for the first equation we have $Lu_1 = (\mu + a)u_1 + bu_2 + f_1(x, u_1)$ and $Lv_1 = (\mu + a)v_1 + bv_2 + f_1(x, v_2)$. Since we are in $\mathcal{K}^+$ (resp. $\mathcal{K}^-$), divide by bu_1 the first equation and by bv_1 the second one and subtract:

$$\frac{-\Delta u_1}{bu_1} + \frac{\Delta v_1}{bv_1} = \frac{u_2}{u_1} - \frac{v_2}{v_1} + \frac{f_1(x, u_1)}{bu_1} - \frac{f_1(x, v_1)}{bv_1}. \tag{7}$$

Exactly as in [13] multiply by $(u_1^2 - v_1^2)$ and integrate; hence

$$\int \left(\frac{-\Delta u_1}{bu_1} + \frac{\Delta v_1}{bv_1} \right) (u_1^2 - v_1^2) = \int \left(\frac{u_2}{u_1} - \frac{v_2}{v_1} + \frac{f_1(x, u_1)}{bu_1} - \frac{f_1(x, v_1)}{bv_1} \right) (u_1^2 - v_1^2).$$

The first terme is non-negative by (4):

$$\int \left(\frac{-\Delta u_1}{bu_1} + \frac{\Delta v_1}{bv_1} \right) (u_1^2 - v_1^2) > 0.$$

Then do exactly the same calculus with the second equation in (S) and add these two lines: we derive from (7) that $T_1 = T_2$ with

$$T_1 = \int \left(\frac{-\Delta u_1}{bu_1} + \frac{\Delta v_1}{bv_1} \right) (u_1^2 - v_1^2) + \int \left(\frac{-\Delta u_2}{cu_2} + \frac{\Delta v_2}{cv_2} \right) (u_2^2 - v_2^2).$$

$$T_2 = \int \left(\frac{u_2}{u_1} - \frac{v_2}{v_1} + \frac{f_1(x,u_1)}{bu_1} - \frac{f_1(x,v_1)}{bv_1} \right) (u_1^2 - v_1^2) + \int \left(\frac{u_1}{u_2} - \frac{v_1}{v_2} + \frac{f_2(x,u_2)}{cu_2} - \frac{f_2(x,v_2)}{cv_2} \right) (u_2^2 - v_2^2).$$

Of course the 1st term T_1 is non-negative by (4). By (H'_F),

$$\int \left(\frac{f(x,u_1)}{bu_1} - \frac{f_1(x,v_1)}{bv_1} \right) (u_1^2 - v_1^2) + \int \left(\frac{f_2(x,u_2)}{cu_1} - \frac{f_2(x,v_2)}{cv_1} \right) (u_2^2 - v_2^2) < 0.$$

We develop what is left and get

$$\int \left(\frac{u_2}{u_1} - \frac{v_2}{v_1} \right) (u_1^2 - v_1^2) + \int \left(\frac{u_1}{u_2} - \frac{v_1}{v_2} \right) (u_2^2 - v_2^2) = -\int \left(\sqrt{\frac{u_2 v_1^2}{u_1}} - \sqrt{\frac{u_1 v_2^2}{u_2}} \right)^2 - \int \left(\sqrt{\frac{v_2 u_1^2}{v_1}} - \sqrt{\frac{v_1 u_2^2}{v_2}} \right)^2 < 0$$

Hence $T_1 = T_2 = 0$ and $u_1 = v_1$,$u_2 = v_2$. The solution is unique.

References

[1] **Abakhti-Mchachti, A.**, and **Fleckinger, J. :** *Existence of positive solutions for non cooperative semilinear elliptic systems defined on an unbounded domain.* Pitman Research Notes in Math. **255**, (1992), p. 92 – 106

[2] **Alziary, B.**, and **Besbas, N. :** *Anti-Maximum Principle for a Schrödinger Equation in $\mathbb{R}^N$, with a non radial potential.* Rostock. Math. Kolloq. **59** (2005), p. 51 – 62

[3] **Alziary, B., Cardoulis, L.**, and **Fleckinger, J. :** *Maximum principle and existence of solutions for elliptic systems involving Schrödinger operators.* Rev. R. Acad. Cienc. Exact. Fis. Nat. **91 (1)** (1997), p. 47 – 52

[4] **Alziary, B.**, and **Fleckinger, J. :** *Sign of the solutions to a non cooperative system.* Rostock. Math. Kolloq. **71**, (2016), p. 3 – 13

[5] **Alziary, B.**, and **Fleckinger, J. :** *Blow up of the solutions to a linear elliptic system involving Schrödinger operators.* to appear

[6] **Alziary, B., Fleckinger, J., Lécureux, M. H.**, and **Wei, N. :** *Positivity and negativity of solutions to $n \times n$ weighted systems involving the Laplace operator defined on $\mathbb{R}^N$, $N \geq 3$.* Electron. J. Diff. Eqns. **101**, 2012, p. 1 – 14

[7] **Alziary, B., Fleckinger, J.**, and **Takáč, P. :** *Maximum and anti-maximum principles for some systems involving Schrödinger operator.* Operator Theory: Advances and applications **110**, 1999, p. 13 – 21

[8] **Alziary, B., Fleckinger, J.**, and **Takáč, P. :** *An extension of maximum and anti-maximum principles to a Schrödinger equation in $\mathbb{R}^N$.* Positivity **5**, (4), 2001, p. 359 – 382

[9] **Alziary, B., Fleckinger, J.**, and **Takáč, P. :** *Groundstate positivity, negativity, and compactness for a Schrödinger operator in $\mathbb{R}^N$.* J. Funct. Anal., **245** (2007), p. 213 – 248. *Online*: doi: 10.1016/j.jfa.2006.12.007

[10] **Alziary, B.**, and **Takáč, P. :** *Compactness for a Schrödinger operator in the groundstate space over $\mathbb{R}^N$.* Electr. J. Diff. Eq., Conf. 16, (2007) p. 35 – 58

[11] **Amann, H. :** *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces.* SIAM Re. **18**, 4, 1976, p. 620-709

[12] **Amann, H. :** *Maximum Principles and Principal Eigenvalues.* Ten Mathematical Essays on Approximation in Analysis and Topology, J. Ferrera, J. López-Gómez, F. R. Ruíz del Portal ed., Elsevier, 2005, p. 1 – 60

[13] **Brezis, H.**, and **Oswald, L. :** *Remarks on sublinear elliptic equations.* Nonlinear Anal., T.M.A. **10**, 1986, p. 55 – 64

[14] **Clément, P.**, and **Peletier, L. :** *An anti-maximum principle for second order elliptic operators.* J. Diff. Equ. **34**, 1979, p. 218 – 229

[15] **Edmunds, D. E.**, and **Evans, W. D. :** *Spectral Theory and Differential Operators.* Oxford Science Publications, 1987

[16] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *A maximum principle for an elliptic system and applications to semilinear problems.* S.I.A.M., J. Math. Anal. **17**, 1986, p. 836 – 849

[17] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *Maximum principle for linear elliptic systems.* Quaterno Matematico 177, Dip. Sc. Mat., Univ. Trieste, 1988

[18] **de Figueiredo, D. G.**, and **Mitidieri, E. :** *Maximum principle for cooperative elliptic systems.* Comptes Rendus Acad. Sc. Paris **310**, 1990, p. 49 – 52

[19] **Fleckinger, J. :** *Estimate of the number of eigenvalues for an operator of Schrödinger type.* Proceedings of the Royal Society of Edinburgh **89 A**, p. 355 – 361, 1981

[20] **Fleckinger, J., Hernandez, J.**, and **de Thélin, F. :** *On maximum principles and existence of positive solutions for some cooperative elliptic systems.* Diff and Int Eq. **V. 8**, N.1, p. 69 – 85, 1995

[21] **Hernández, J. :** *Maximum Principles and Decoupling for Positive Solutions of Reaction Diffusion Systems.* In: K. J. Brown, A. A. Lacey, eds. Reaction Diffusion Equations, Oxford, Clarendon Press, p. 199 – 224, 1990

[22] **Lécureux, M. H. :** *Comparison with groundstate for solutions of non cooperative systems for Schrödinger operators in $\mathbb{R}^N$.* Rostock. Math. Kolloq. **65**, 2010, p. 51 – 69

[23] **Fleckinger, J.**, and **Serag, H. :** *Semilinear cooperative elliptic systems on $\mathbb{R}^n$.* Rendiconti Mat. Ser. VII, **V. 15**, (1995) p. 89 – 108

[24] **Reed, M.**, and **Simon, B. :** *Methods of modern mathematical physics IV. Analysis of operators.* Acad. Press, New York, 1978

[25] **Sweers, G. :** *A Strong Maximum Principle for a Non-Cooperative Elliptic Systems.* S.I.A.M. J. Math. Anal. **Vol. 20**, p. 367 – 371, 1989

received: May 22, 2018

Authors:

Bénédicte Alziary
Institut de Mathématiques
de Toulouse CeReMath, TSE,
Univ. Toulouse 1,

e-mail: alziary@ut-capitole.fr

Jacqueline Fleckinger
Institut de Mathématiques
de Toulouse CeReMath,
Univ. Toulouse 1,
address: 21 allée de Brienne
31042 Toulouse Cedex

e-mail: jfleckinger@gmail.com

Rostock. Math. Kolloq. **71**, 57 – 67 (2017/18)

Subject Classification (AMS)
62 F07

Dieter Schott

Some Remarks on a Statistical Selection Procedure of Bechhofer for Expectations

ABSTRACT. Following a Bechhofer statistical selection procedure we discuss from an analytical and from a probabilistic point of view why the real function

$$F(x) = \int_{-\infty}^{+\infty} \Phi^{a-t}\left(z + r\sqrt{x}\right) \cdot t\,(1 - \Phi(z))^{t-1}\,\varphi(z)\,dz$$

is for $x \geq 0$ and fixed integer parameters $a > 0$, $t \in]0, a[$ as well as real parameters $r > 0$ and $\beta \in]0, 1[$ strictly monotone increasing and bounded by 1. Here φ and Φ denote the p.d.f. and c.d.f. of the standard normal distribution. Numerical procedures are described to determine the minimal natural n satisfying the inequality $F(n) \geq K$ where $0 < K < 1$. The dependence of n on the parameters a, t and r is investigated, too. Some simulation results are given and discussed for $t > 1$.

KEY WORDS. Monotone Functions, Inequalities, Selection Procedures for Expectations, Bechhofer Selection Problem, Indifference Zone Selection

Acknowledgement. I thank the referee of the journal supplying valuable suggestions for improvement.

1 Introduction

In the textbooks [2], p. 489f., and [3], p. 513f., statistical selection procedures are described to separate the t in some sense best populations from a collection of a normally distributed populations with unknown expectations and the same known variance σ^2 for a given risk $\beta \in]0, 1[$ of wrong decision. More precisely, the populations G_i $(i = 1, ..., a)$ in the collection

$$L = \{G_1,\, G_2,\, ...,\, G_a\}$$

with characteristics $\mathbf{u}_i$ are assumed to be ranked according to increasing (not decreasing) expectations (means) μ as follows:

$$\left(G_{(1)}, \mu_{(1)}\right),\ \left(G_{(2)}, \mu_{(2)}\right),\ ...,\ \left(G_{(a)}, \mu_{(a)}\right)\ .$$

Then L is to partition into two sets

$$M = \left\{G_{(a)}, ..., G_{(a-t+1)}\right\}, \; N = \left\{G_{(a-t)}, ..., G_{(1)}\right\}$$

where M is the top set with the t best populations. As a prerequisite we assume that there is a gap

$$d(M, N) = \left|\, \mu_{(a-t+1)} - \mu_{(a-t)} \,\right| \geq \delta > 0 \tag{1.1}$$

between the top populations and the remaining ones. The populations are selected taking stochastically independent random samples

$$(\mathbf{u}_{i1},\, \mathbf{u}_{i2},\, ...,\, \mathbf{u}_{in})$$

of G_i $(i = 1, ..., a)$ with constant size n and ranking G_i according to their sample means $\bar{\mathbf{u}}_i$ obtaining a set $\mathbf{M}_s$ of t populations G_i. The question is if

$$P(\mathbf{M}_s = M |\, d(M, N) \geq \delta) \geq 1 - \beta.$$

The number n has to be great enough such that $\mathbf{M}_s$ satisfies this condition, but not too great to reduce the effort of ranking. Hence, we look for a minimal or at least nearly minimal such n.

Following [1] this problem is solved under the gap assumption (1.1) and the natural condition

$$\frac{1}{\binom{a}{t}} < 1 - \beta \tag{1.2}$$

if n fulfils the inequality

$$\int_{-\infty}^{+\infty} \Phi^{a-t}\left(z + \frac{\delta}{\sigma}\sqrt{n}\right) \cdot t\,(1 - \Phi(z))^{t-1}\, \varphi(z)\, dz \geq 1 - \beta \tag{1.3}$$

(see also [2], p. 494 and [3], p. 517, respectively). Here φ is as usual the probability density function (or p.d.f) and Φ the cumulative distribution function (or c.d.f.) of the standard normal distribution $N(0, 1)$. For an analytical investigation it is useful to introduce the real function

$$F(x) := \int_{-\infty}^{+\infty} \Phi^{a-t}\left(z + r\sqrt{x}\right) \cdot t\,(1 - \Phi(z))^{t-1}\, \varphi(z)\, dz, \quad x \geq 0 \tag{1.4}$$

and the constants

$$r := \frac{\delta}{\sigma}, \quad K := 1 - \beta.$$

It is easy to see that the improper integral in (1.4) exists and $F(x)$ is defined. Besides $F(x)$ is a real extension of the left-hand side of the inequality in (1.3). Hence, a reformulation of condition (1.3) reads

$$F(n) \geq K. \tag{1.5}$$

The function F in (1.4) is called *Bechhofer function* in the following. Observe that in the paper [5] the function $f(x) = \frac{1}{t}F(x)$ is used instead of $F(x)$.

2 Problem Analysis and Probabilistic Interpretation

The condition (1.1) is chosen from a practical point of view to get a reasonable separation of the two sets M and N. On the other hand: in the case $\delta = 0$, i.e. $r = 0$, we would get in (1.4) simply a constant $F(x) = F(0)$. Thus (1.5) would be fulfilled for all natural n or for no natural n.

The condition (1.2) has also an important practical meaning. If it is violated, then no selection process is necessary. Without sampling one could denote any of the $\binom{a}{t}$ subsets of size t by $\mathbf{M}_s$ satisfying the above probability condition.

The condition (1.3) can be given a simple probabilistic interpretation. Taking the extreme cases with

$$\mu_{(a)} = \ldots = \mu_{(a-t+1)} = m + \delta, \quad \mu_{(a-t)} = \ldots = \mu_{(1)} = m$$

for some $m \in R$ into consideration then the probability of a correct decision is just

$$P\left(\mathbf{V}_M + r\sqrt{n} \geq \mathbf{V}_N\right) = F(n), \tag{2.1}$$

where

$$\mathbf{V}_M = \min_{i=a-t+1,\ldots,a} \mathbf{v}_{(i)}, \quad \mathbf{V}_N = \max_{i=1,\ldots,a-t} \mathbf{v}_{(i)}$$

and the $\mathbf{v}_{(i)}$ are the independently and identically distributed (i.i.d.) standard normal random variables obtained from the corresponding sample means $\bar{\mathbf{u}}_{(i)}$. Replacing $r\sqrt{n}$ by the real variable y and the statistic $\mathbf{V}_N - \mathbf{V}_M$ by $\mathbf{D}$ this can be rewritten in the generalized form

$$G(y) = P(\mathbf{D} \leq y) = \int_{-\infty}^{+\infty} F_N(z+y)\, f_M(z)\, dz \tag{2.2}$$

with the functions

$$F_N(z) := \Phi^{a-t}(z), \quad f_M(z) := t\,(1 - \Phi(z))^{t-1}\,\varphi(z).$$

Here F_N is the c.d.f. of $\mathbf{V}_N$ and f_M is the p.d.f. of $\mathbf{V}_M$. Besides, we have the relation

$$F(x) = G(r\sqrt{x}), \quad x \geq 0. \tag{2.3}$$

In [1], [2] and [3] tables for $r\sqrt{n}$ can be found for some special parameters though only for small values of a. We look for a general solution of the problem to support an extensive simulation study [4] and further experiments.

The Bechhofer problem simplifies for $t = 1$. Then an effective formula for a minimal n exists using β-quantiles of an $(a-1)$-dimensional normal distribution (see again [2] and [3]). Hence, we treat especially the cases $t > 1$.

3 Global Properties of the Bechhofer Function

The behavior of the Bechhofer function is crucial for the solution of the Bechhofer selection problem. It is interesting to compare different approaches for proving statements about this function. We use here on the one hand basic facts of analysis and on the other hand basic facts of probability calculus given as supplements.

We start with the integrand in (1.4) denoted by

$$I(x,z) = I(x,z;a,t,r) := \Phi^{a-t}\left(z + r\sqrt{x}\right) \cdot t\,(1-\Phi(z))^{t-1}\,\varphi(z), \tag{3.1}$$

where $a \in \mathbb{N}$, $t \in \mathbb{N}$, $t < a$, $r \in \mathbb{R}$, $r > 0$ are parameters and $z \in \mathbb{R}$, $x \in \mathbb{R}_+$ are variables. By the way, another representation is

$$I(x,z) = -\Phi^{a-t}\left(z + r\sqrt{x}\right) \cdot \frac{d}{d\,z}\,(1-\Phi(z))^{t}\,.$$

In the paper [5] modified integrands $i(x,z) := \frac{1}{t} \cdot I(x,z)$ are plotted for $a = 10$, $t = 3$, $r = 1$ and $x = n = 0, 1, \ldots, 5$ using MATLAB. These integrands turn out to be unimodal for fixed n. The global maxima $i(n, z_{max})$, and also $I(n, z_{max})$, increase and their positions z_{max} walk to the left with increasing n.

We consider the two functions

$$F(x) = F(x;a,t,r) = \int_{-\infty}^{+\infty} I(x,z)\,dz,$$
$$F_N(x) = F_N(x;a,t,r) = \int_{-N}^{+N} I(x,z)\,dz$$

where N is an appropriate positive number. The cut function F_N of the Bechhofer function F comes into play if numerical integration is used to compute the improper integral. In the paper [5] a modified cut function $f_N(x) := \frac{1}{t} \cdot F_N(x)$ is plotted for $a = 10$, $t = 3$, $r = 1$ and $N = 5$ using MATLAB. We state some general properties of these functions.

Proposition 3.1 *The Bechhover function $F(x)$ and its cut version $F_N(x)$ are continuous, strictly monotone increasing and bounded for all $x \geq 0$ as well as smooth for all $x > 0$. The difference of both functions satisfying the relation $0 < F_N(x) < F(x)$ for all $x \geq 0$ can be made (uniformly in x) arbitrary small for sufficiently large N.*

Proof: The integrand $I(x,z)$ is composed of continuous functions and is itself continuous with respect to z and x. Therefore both $F(x)$ and $F_N(x)$ are continuous.

The integrand $I(x,z)$ is strictly monotone increasing with respect to x (for fixed z) taking the strict monotony of $\sqrt{x}$ and the powers of $\Phi(z)$ into account. Consequently, both $F(x)$ and $F_N(x)$ are strictly monotone increasing. Further, the integrand satisfies the estimations

$$0 < I(x,z) < t\,\left[1-\Phi(z)\right]^{t-1}\varphi(z) < t\,\varphi(z)$$

since the values of Φ are contained in the interval $]0,1[$. Thus we have

$$0 < F_N(x) < F(x) < t \int_{-\infty}^{+\infty} \varphi(z)\, dz = t.$$

Hence, F and F_N are bounded (e.g. by t).
The functions in the integrand $I(x,z)$ are arbitrarily often differentiable for all arguments with the exception of $\sqrt{x}$ where $x > 0$ is necessary. Since

$$F^{(k)}(x) = \int_{-\infty}^{+\infty} \frac{\partial^k}{\partial x^k} I(x,z)\, dz, \quad x > 0$$

holds for $k \in \mathbb{N}$, the smoothness follows for $x > 0$ using rules of differential calculus.
Finally we get

$$\begin{aligned} F(x) &= F_N(x) + R_N(x), \\ 0 < R_N(x) &= \int_{-\infty}^{-N} I(x,z)\, dz + \int_{N}^{+\infty} I(x,z)\, dz < 2t \int_{N}^{+\infty} \varphi(z)\, dz \\ &= 2t(1 - \Phi(N)) \end{aligned}$$

Consequently, the difference $R_N(x)$ of both functions can be made smaller than ε for $N > \Phi^{-1}\left(1 - \frac{\varepsilon}{2t}\right)$. ■

Supplement 3.1 The continuity and the monotony of $F(x)$ can also be derived from (2.2) and (2.3) since G is a continuous c.d.f. It follows also that $F(x)$ is even bounded by 1.

4 Local Properties of the Bechhofer Function

First we present an interesting statement for improper integration with respect to functions involving p.d.f. and c.d.f. of the normal distribution.

Lemma 4.1 *For nonnegative integers l and m it holds*

$$\begin{aligned} J_{l,m} := \int_{-\infty}^{+\infty} \Phi^l(z) \cdot (1 - \Phi(z))^m \varphi(z)\, dz &= \sum_{i=0}^{m} \binom{m}{i} \cdot \frac{(-1)^i}{l+i+1} \\ &= \frac{m!}{(l+1)\cdot(l+2)\cdot\ldots\cdot(l+m+1)} = \frac{l! \cdot m!}{(l+m+1)!} \\ &= \frac{1}{(m+1)\cdot\binom{l+m+1}{m+1}}\,. \end{aligned}$$

Proof: The first assertion can be shown by partial integration and the other ones by mathematical induction and simple transformations. In [6] a corresponding result is proved in detail with a more general integrand. ■

Remark 4.1 The final result in Lemma 4.1 can also be proved by applying the substitution $u = \Phi(z)$ in the integral. Then $du = \varphi(z)\,dz$, and the integral is reduced to the Euler beta function and gamma function, respectively. Namely, it is

$$\begin{aligned} J_{l,m} &= \int_0^1 u^l (1-u)^m \, du = B(l+1, m+1) = \frac{\Gamma(l+1)\cdot\Gamma(m+1)}{\Gamma(l+m+2)} \\ &= \frac{l!\cdot m!}{(l+m+1)!}. \end{aligned}$$

Proposition 4.1 *The Bechhofer function $F(x)$ starts with the value*

$$F(0) = \frac{1}{\binom{a}{t}}$$

and tends from below to

$$F(\infty) = \lim_{x\to+\infty} F(x) = 1.$$

Its range is

$$R(F) = \left[\frac{1}{\binom{a}{t}}, 1 \right[.$$

Further, the ascent (gradient) of $F(x)$ starts vertical and ends horizontal, that is

$$F'(0+0) = \lim_{x\to+0} F'(x) = +\infty, \quad F'(\infty) = \lim_{x\to+\infty} F'(x) = 0.$$

Proof: We get from Lemma 4.1 putting $l = a - t$ and $m = t - 1$

$$F(0) = t \cdot J_{a-t,t-1} = t \cdot \frac{1}{t\binom{a}{t}} = \frac{1}{\binom{a}{t}}$$

and putting $l = 0$ and $m = t - 1$

$$F(\infty) = t \cdot J_{0,m-1} = t \cdot \frac{1}{t} = 1$$

since then the continuous function Φ in the integrand tends to 1. As F is strictly monotone increasing and continuous by Proposition 3.1 the limit is reached from below and the range $R(F)$ is as asserted.

Now we calculate under observation of (3.1)

$$\frac{\partial}{\partial x} I(x,z) = \frac{rt(a-t)}{2} \frac{\Phi^{a-t-1}\left(z + r\sqrt{x}\right)\cdot\varphi\left(z + r\sqrt{x}\right)}{\sqrt{x}} \cdot \left(1 - \Phi(z)\right)^{t-1} \varphi(z)$$

which is defined for all $x > 0$. This implies

$$\begin{aligned} F'(0+0) &= \lim_{x\to+0} F'(x) = \lim_{x\to+0} \int_{-\infty}^{+\infty} \frac{\partial}{\partial x} I(x,z)\,dz \\ &= \lim_{x\to+0} \frac{rt(a-t)}{2\sqrt{x}} \cdot \int_{-\infty}^{+\infty} \Phi^{a-t-1}(z)\left(1-\Phi(z)\right)^{t-1} \cdot \varphi^2(z)\,dz \\ &= +\infty. \end{aligned}$$

Here we remark that the improper integral exists and is finite because

$$\begin{aligned}&\int_{-\infty}^{+\infty} \Phi^{a-t-1}(z)\,(1-\Phi(z))^{t-1}\cdot\varphi^2(z)\,dz\\ &< \int_{-\infty}^{+\infty} \Phi^{a-t-1}(z)\,(1-\Phi(z))^{t-1}\cdot\varphi(z)\,dz\\ &= J_{a-t-1,t-1} = \frac{1}{(a-t)\cdot\binom{a-1}{t}}\end{aligned}$$

considering $0 < \varphi(z) < 1$ for all z and Lemma 4.1. Finally, we have

$$\begin{aligned}F'(\infty) &= \lim_{x\to+\infty}\int_{-\infty}^{+\infty}\frac{\partial}{\partial x}I(x,z)\,dz\\ &= \lim_{x\to+\infty}\frac{rt(a-t)}{2\sqrt{x}}\cdot\int_{-\infty}^{+\infty}0\,dz = 0. \quad\blacksquare\end{aligned}$$

Supplement 4.1 The value $F(0)$ in Proposition 4.1 is also obvious from the probabilistic point of view. Consider that F is continuous and $x = n = 0$ corresponds to the case of no sampling where each of the $\binom{a}{t}$ selections is equiprobable.
The value $F(\infty) = 1$ is intuitively clear because for sampling sizes $n \to \infty$ the means $\bar{u}_i$ converge to the expectations μ_i for all $i \in \{1,\dots,a\}$. Another argument is based on (2.2) and (2.3). G is a c.d.f. and $F(\infty) = G(\infty) = 1$. Besides, we have also $F'(\infty) = G'(\infty) = 0$ because of

$$F'(x) = G'(r\sqrt{x})\cdot\frac{r}{2\sqrt{x}}.$$

5 About the solution of the Bechhofer Problem

Theorem 5.1 *For each constant* $K \in \left]\frac{1}{\binom{a}{t}}, 1\right[$ *there is a unique argument* $x = x_K > 0$ *of the Bechhofer function* F *satisfying*

$$F(x) = K.$$

This x_K *can be written as* $x_K = F^{-1}(K)$. *The solution set of*

$$F(x) \geq K$$

is given by the elements $x \geq x_K$.

Proof: The assertions are direct consequences of Proposition 3.1 and Proposition 4.1. Besides, Proposition 3.1 ensures that the inverse function of F exists. $\blacksquare$

Principally x_K can also be derived from the c.d.f. G given in (2.2) taking (2.3) into account. Then we get

$$x_K = \left(\frac{G^{-1}(K)}{r} \right)^2 .$$

An analogue statement exists for the cut function F_N. The approximative solution $\tilde{x}_K = F_N^{-1}(K)$ of $F_N(x) = K$ satisfies $\tilde{x}_K > x_K$.

Corollary 5.1 *Under the conditions* (1.1) *and* (1.2) *there is a minimal natural* $n = n_B$ *such that* (1.3) *is fulfilled for all natural* $n \geq n_B$*.*

Proof: We consider the reformulation (1.5) of (1.3) using the Bechhofer function (1.4) which involves already the gap number δ from (1.1). Here it is $K = 1 - \beta$. If (1.2) holds we choose $\frac{1}{\binom{a}{t}} < K < 1$ and can apply Theorem 5.1. Then $F(x) \geq K$ is fulfilled just for all $x \geq x_K = F^{-1}(K)$ and (1.5) just for all natural $n \geq n_B := \lceil x_K \rceil$. ■

Remark 5.1 The Corollary shows that the number n_B indicates the minimal sampling size for the Bechhofer selection problem.

Let us now turn to practical solution methods of our problem. In statistically relevant cases condition (1.2) is fulfilled such that we can concentrate on condition (1.3), i. e. the Bechhofer function $F(x)$. Using mathematical software the p.d.f. φ and the c.d.f. Φ of the standard normal distribution are predefined or can be easily defined. We can work with natural arguments n or real arguments x to solve the Bechhofer selection problem. We start with the first possibility which is realized by a MATLAB program in [5].
Then we declare the integrand $I(n, z)$. The cut number N is chosen great enough, $K = 1-\beta$ is determined and a numerical procedure to calculate the definite integrals $F_N(n)$ is used (trapezoidal rule, Simpson's rule or some more sophisticated method) such that a certain accuracy is realized for integration. It is not necessary to use numerical top methods since F_N has a very simple behavior. Starting with $n = 0$ the natural number n is increased step by step as long until $F_N(n)$ jumps over K (e.g. by using a while-loop). We know that $F(n)$ is than over K, too. The program should print this last $n = n_B$ as the one we looked for. Considering the numerical errors it can happen if we are not careful enough that n_B is chosen one unit too large. But this is meaningless for practical applications.
Another possibility is to replace n by real $x \geq 0$. Then F_N becomes a continuous, strongly monotone function $F_N(x)$ which takes the value K exactly once, say at $x = \tilde{x}_K \approx x_K$. Then numerical methods can be used to determine the zero of $F_N(x) - K = 0$ for sufficient large N (bisection, Newton's method or other ones). In this case we get n_B by rounding $\tilde{x}_K$ off to the next integer. This approach will be used in this paper on the basis of MATLAB and in the simulation study [4] on the basis of the statistical software R.

6 About Simulation Results

In the paper [5] a MATLAB program is presented implementing the solution method mentioned above for natural n and using the function '*quad*' for numerical quadrature (adaptive Simpson's rule). For

$$a = 30,\ 50,\ 100,\ 200; \quad t = 2,\ 3$$
$$r = 0.5,\ 1; \quad \beta = 0.05$$

the minimal sample sizes $n = n_B$ are calculated. The simulation shows that n_B increases for decreasing r, increasing t and increasing $a - t$ assuming the other parameters are fixed. This can also easily be seen by theoretical considerations.
Considering the expression $r\sqrt{x}$ in (1.4) we have

$$cr \cdot \sqrt{\frac{x}{c^2}} = r\sqrt{x}, \quad c > 0.$$

Consequently, multiplying r with the factor c means dividing $x = x_K$ by c^2. Thus we can restrict ourselves to $r = 1$.
In [5] the relation between the calculated a and n_B is fit for $t = 2$, $r = 1$ and the mentioned values of a very well by a logarithmic term. This fact can be explained at least for large a and moderate fixed t. It is well-known that for independent standard normal random variables $\mathbf{v}_{(i)}$ $(i = 1, ..., a - t)$ the relation

$$\mathbf{V}_N = \max_{i=1,\dots,a-t} \mathbf{v}_{(i)} \approx \sqrt{2\ln(a-t)}$$

holds as $a - t$ gets large. Using this in (2.1) with $r = 1$ we have

$$P(\mathbf{V}_M + \sqrt{n_B} \geq \sqrt{2\ln a}) \approx F(n_B) \approx K$$

replacing $\ln(a - t)$ by $\ln a$ for relatively small t and taking in (1.5) the limit case. Since $\mathbf{V}_M$ is fixed for fixed t we obtain

$$n_B \approx 2\ln a + c_{t,K}.$$

with a constant $c_{t,K}$ depending only on t and K.
This asymptotic estimate can also be used with x_K instead of n_B.
By Proposition 4.1 and because of $K = 1 - \beta = 0.95$ we get for certain $a > 2$ and $t = 2$

$$F(0) = \frac{2}{a \cdot (a-1)}, \quad F(\infty) = 1$$

and for certain $a > 3$ and $t = 3$

$$F(0) = \frac{3}{a \cdot (a-1) \cdot (a-2)}, \quad F(\infty) = 1.$$

Thus, the Bechhofer function starts with small or very small values $F(0)$. The critical lower bound K and the supremum $F(\infty)$ of F form a narrow strip.

As already mentioned, in this paper we use a MATLAB program based on a zero method for real arguments x to determine first $x = x_K$ and then $n = n_B$. The next tables contain the simulation results for x_K, rounded to two digits after the decimal point, and for n_B.

Table 1: case $r = 1$ and $t = 2$

a	30	50	100	200
x_K	17.71	19.36	21.50	23.59
n_B	18	20	22	24

Table 2: case $r = 1$ and $t = 3$

a	30	50	100	200
x_K	19.19	20.93	23.18	25.35
n_B	20	21	24	26

If we multiply x_K with 4 and round off to the next integer we get the corresponding results for $r = 0.5$ instead of $r = 1$ (compare results in [5]). If we consider the difference $x_K(a) - 2\ln a$ for the given a, then it increases only slowly. This seems to manifest the above derived asymptotic estimate. But for a more accurate analysis we would need values $x_K(a)$ for still greater a.

References

[1] **Bechhofer, R. E. :** *A Single Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances.* Ann. Math. Statist. 25, 16 – 39 (1954)

[2] **Rasch, D.**, and **Schott, D. :** *Mathematische Statistik für Mathematiker, Natur- und Ingenieurwissenschaftler.* Wiley-VCH 2016

[3] **Rasch, D.**, and **Schott, D. :** *Mathematical Statistics.* Wiley 2018

[4] **Rasch, D., Takuya, Y., Schott, D.**, and **Pilz, J. :** *Statistical Selection Procedures for Expectations – a Review and Recent Simulation Results.* Paper for the 10th International Workshop on Simulation and Statistics, Salzburg 2019

[5] **Schott, D. :** *How to get in the Top Ten? An Analysis of the Bechhofer Selection Problem in Statistics.* In: Proceedings of the 1st Northern-Light Symposium, Hafencity University Hamburg, April 2018. Wismarer Frege-Reihe, Heft 02/2018, 7 – 22

[6] **Schott, D. :** *Monotone Functions Generated by Improper Integrals and Applications.* To be submitted

received: May 22, 2018

Author:

Dieter Schott
Hochschule Wismar
Fakultät für Ingenieurwissenschaften
Bereich Elektrotechnik und Informatik
Philipp-Müller-Str. 14
D-23966 Wismar

e-mail: dieter.schott@hs-wismar.de

Rostock. Math. Kolloq. **71**, 68 – 99 (2017/18)

Subject Classification (AMS)
11A63, 39A10, 11B68, 05A10

Manfred Krüppel

Two-Scale Difference Equations with a Parameter and Power Sums related to Digital Sequences

ABSTRACT. This paper is a direct continuation of [19] concerning the representation of power sums related to digital sequences. Foundation is beside article [19] an existence theorem for differentiable solutions of certain two-scale difference equations with a parameter. By means of such solutions and a method developed in [19] we are able to give an explicit representation for general sums related to digital sequences. In particular, we give a summation formula for power sums of the sum of digits and incidentally, we find a new property of the Bernoulli polynomials.

KEY WORDS. Two-scale difference equations with a parameter, power sums of digital sums, Bernoulli polynomials, generating functions

1 Introduction

We consider a two-scale difference equation with a parameter $x \in X \subseteq \mathbb{R}$ of the form

$$\varphi\left(\frac{t}{p}, x\right) = \sum_{r=0}^{p-1} c_r(x)\varphi(t-r, x) \qquad (t \in \mathbb{R}) \tag{1.1}$$

with an integer $p > 1$ and real or complex coefficients $c_r(x)$ where $c_0(x)c_{p-1}(x) \neq 0$ and

$$\sum_{r=0}^{p-1} c_r(x) = 1 \qquad (x \in X). \tag{1.2}$$

It is known that for such $x \in X$ where $|c_r(x)| < 1$ for $r = 0, 1, \dots, p-1$ the equation (1.1) has a solution $\varphi(t, x)$ satisfying

$$\varphi(t, x) = 0 \quad \text{for} \quad t < 0, \qquad \varphi(t, x) = 1 \quad \text{for} \quad t > 1 \tag{1.3}$$

which is continuous with respect to t, cf. [18], see also [7], [11]. We show that if coefficients $c_r(x)$ are k-times differentiable then the solution $\varphi(t, x)$ is k-times differentiable with respect to x (Theorem 2.1, Theorem 2.2). This result is the base for the derivation of a formula for the sum of certain digital sequences where we use similar methods as in [19].

In particular, we investigate digital exponential sums (Section 3) and the digital power sums

$$S_k(N) := \sum_{n=0}^{N-1} s(n)^k \tag{1.4}$$

where $s(n)$ denotes the sum of digits of the integer n in the p-adic representation, i.e. if $n = \sum n_k p^k$ then $s(n) = \sum n_k$. In binary case $p = 2$ first Trollope [26] in 1968 has given an explicit expression for the sum $S_1(N)$. Delange [5] gave a simple proof and generalized the result to digits in arbitrary basis $p > 1$. The well-known Trollope-Delange formula reads

$$S_1(N) = \frac{1}{2} N \log_2 N + NG(\log_2 N) \tag{1.5}$$

where $G(u)$ is an 1-periodic continuous, nowhere differentiable function which can be expressed by means of the Takagi function. In 1994 the Trollope-Delange formula (1.5) was also proved in [6] by use of classical tools from analytic number theory, namely the Mellin-Perron formulae, see [6, Theorem 3.1, Remark 4.5].

For the basis $p = 2$ Coquet [3] in 1986 proved that

$$\frac{1}{N} S_2(N) = \left(\frac{\log_2 N}{2}\right)^2 + \log_2 N \; G_{2,1}(\log_2 N) + G_{2,0}(\log_2 N)$$

where $G_{2,1}(u)$, $G_{2,0}(u)$ are 1-periodic continuous functions and that for arbitrary integer $k > 1$ the power sum $S_k(N)$ can be represented in the form

$$\frac{1}{N} S_k(N) = \sum_{\ell=0}^{k} (\log_2 N)^\ell G_{k,\ell}(\log_2 N) \tag{1.6}$$

where $G_{k,\ell}(u)$ are 1-periodic functions, in particular $G_{k,k}(u) = \frac{1}{2^k}$. He also found certain recurrence relations between the functions $G_{k,\ell}$. In [22], by means of binomial measures a more explicit representation of $G_{k,\ell}$ was given and their continuity was proved, cf. also [23] and [15]. In [17] it was proved that the functions $G_{k,\ell}$ $(\ell = 0, 1, \ldots, k-1)$ are nowhere differentiable. In 2012 Girgensohn [8] gives a new representation for $S_k(N)$ by use of functional equations and generating function techniques. If $q_k(t)$ is a sequence of polynomials given by $q_0(t) = 1$ and the recursion

$$q_{k+1}(t) = t(2q_k(t) - q_k(t-1)) \qquad (k \geq 0) \tag{1.7}$$

then it holds

$$S_k(N) = \sum_{\ell=0}^{k} \binom{k}{\ell} N 2^{-\ell} q_\ell(\log_2 N) f_{k-\ell}(x) \tag{1.8}$$

with certain 1-periodic continuous functions $f_k(x)$ and $x = \frac{N-p(N)}{p(N)}$ where $p(N) = 2^{[\log_2 N]}$ is the largest power of 2 less than or equal to N, cf. [8, Section 5].

For an arbitrary integer basis $p > 1$ in 2000 Muramoto et al. [20] have proved by means of multinomial measures that

$$\frac{1}{N} S_k(N) = \sum_{\ell=0}^{k} (\log_p N)^\ell H_{k,\ell}(\log_p N) \tag{1.9}$$

where $H_{k,\ell}(u)$ are 1-periodic continuous functions. In case $N = p^n$ it follows that $\frac{1}{p^n} S_k(p^n)$ can be represented as polynomial

$$\frac{1}{p^n} S_k(p^n) = \sum_{\ell=0}^{k} a_{k,\ell}\; n^\ell \tag{1.10}$$

with the coefficients $a_{k,\ell} = H_{k,\ell}(n) = H_{k,\ell}(0)$ since the functions $H_{k,\ell}(u)$ are 1-periodic. Certainly, the coefficients $a_{k,\ell}$ and also the polynomial

$$P_k(t) = \sum_{\ell=0}^{k} a_{k,\ell}\; t^\ell \tag{1.11}$$

depend on p. So equation (1.10) can be written in the form

$$\frac{1}{p^n} S_k(p^n) = P_k(n). \tag{1.12}$$

For the basis $p = 10$ the polynomials $P_k(n)$ were computed in [12] for $n = 1, 2, \ldots, 8$:

$$\begin{aligned}
\frac{1}{10^n} S_1(10^n) &= \frac{9}{2} n \\
\frac{1}{10^n} S_2(10^n) &= \frac{81}{4} n^2 + \frac{33}{4} n \\
\frac{1}{10^n} S_3(10^n) &= \frac{729}{8} n^3 + \frac{891}{8} n^2 \\
\frac{1}{10^n} S_4(10^n) &= \frac{6561}{16} n^4 + \frac{8019}{8} n^3 + \frac{3267}{16} n^2 - \frac{3333}{40} n \\
\frac{1}{10^n} S_5(10^n) &= \frac{59049}{32} n^5 + \frac{120285}{16} n^4 + \frac{147015}{32} n^3 - \frac{29997}{16} n^2 \\
\frac{1}{10^n} S_6(10^n) &= \frac{531441}{64} n^6 + \frac{3247695}{64} n^5 + \frac{3969405}{64} n^4 - \frac{1080783}{64} n^3 - \frac{329967}{32} n^2 + \frac{15873}{4} n \\
\frac{1}{10^n} S_7(10^n) &= \frac{4782969}{128} n^7 + \frac{40920957}{128} n^6 + \frac{83357505}{128} n^5 - \frac{56133}{128} n^4 - \frac{20787921}{64} n^3 + \frac{99999}{8} n^2 \\
\frac{1}{10^n} S_8(10^n) &= \frac{43046721}{256} n^8 + \frac{122762871}{64} n^7 + \frac{750217545}{128} n^6 + \frac{76284747}{32} n^5 - \frac{1372208607}{256} n^4 \\
&\quad + \frac{677777479}{64} n^3 + \frac{371092563}{320} n^2 - \frac{33333333}{80} n
\end{aligned}$$

Figure 1. The first polynomials $P_k(n)$ for the basis $p = 10$

These results were obtained as follows. If $f_k(x)$ is a sequence of functions defined by

$$f_0(x) = (1 + x + x^2 + \cdots + x^9)^n$$

and for $k \geq 1$ by

$$f_k(x) = x f'_{k-1}(x)$$

then it holds

$$S_k(10^n) = f_k(n)$$

cf. [4, Section 3]. So starting with $f_0(x)$ the polynomials in Figure 1 were computed by repeated differentiation, multiplication by x, and finally substitution $x = 1$, cf. [4, p. 342]. We see in Figure 1 that for odd $k > 1$ the linear term of $P_k(n)$ vanishes.

In this paper we give a new derivation of (1.9) as application of two-scale difference equations with a parameter (Theorem 4.3, Corollary 4.4). The main result can be written as

$$\sum_{k=0}^{\infty} \frac{1}{k!} \frac{1}{N} S_k(N) z^k = \left(\sum_{k=0}^{\infty} \frac{1}{k!} P_k(L) z^k \right) \left(\sum_{k=0}^{\infty} \frac{1}{k!} F_k(L) z^k \right) \qquad (z \in \mathbb{C}) \tag{1.13}$$

where $L = \log_p N$, where $P_k(t)$ are polynomials (1.11) satisfying (1.12), and where $F_0(u) = 1$ and $F_k(u)$ are 1-periodic continuous functions with $F_k(0) = 0$ for $k \geq 1$ (Theorem 6.11). In view of the Cauchy product relation (1.13) means that for $k \geq 0$ we have

$$\frac{1}{N} S_k(N) = \sum_{\ell=0}^{k} \binom{k}{\ell} P_\ell(L) F_{k-\ell}(L).$$

The polynomials $P_k(t)$ are given by their generating function

$$\sum_{k=0}^{\infty} \frac{1}{k!} P_k(t) z^k = \left(\frac{e^{pz} - 1}{p(e^z - 1)} \right)^t \qquad (z \in \mathbb{C}) \tag{1.14}$$

(Proposition 6.2), and the functions $F_k(u)$ are determined by the equation (1.1) with the coefficients

$$c_r(x) = \frac{e^{rx}}{1 + e^x + \cdots + e^{(p-1)x}} \qquad (r = 0, 1, \ldots, p-1) \tag{1.15}$$

in the following way: if equation (1.1) with (1.15) has the solution $\varphi(t, x)$ satisfying (1.3) and if $F(u, x)$ denotes the 1-periodic function with respect to u given by

$$F(u, x) := \frac{\varphi(p^u, x)}{(1 + e^x + \cdots + e^{(p-1)x})^u} \qquad (u \leq 0)$$

then

$$F_k(u) = \frac{\partial^k}{\partial x^k} F(u, x) \bigg|_{x=0}.$$

Moreover, the coefficients $a_{k,\ell} = a_{k,\ell}(p)$ of $P_k(t)$ are polynomials of degree at most k in p which are given by

$$a_{k,\ell}(p) = \frac{(-1)^k k!}{\ell!} \sum_{k_1+\cdots+k_\ell=k} \frac{k!}{k_1!\cdots k_\ell!} \left\{ \prod_{n=1}^{\ell} \frac{B_{k_n}}{k_n \cdot k_n!}(p^{k_n}-1) \right\} \tag{1.16}$$

where $k_1, \ldots, k_\ell$ are positive integers and where B_k denotes the Bernoulli numbers. In particular $a_{k,k}(p) = (\frac{p-1}{2})^k$ for $k \geq 0$, $a_{k,0}(p) = 0$ for $k \geq 1$ and

$$a_{k,1}(p) = \frac{(-1)^k B_k}{k}(p^k - 1) \qquad (k \geq 1)$$

(Proposition 5.4). Hence, for odd $k > 1$ we have $a_{k,1}(p) = 0$ which is the reason that for these k the linear term of $P_k(n)$ vanishes, cf. Figure 1.

In this paper several times we need the Bernoulli polynomials $B_n(t)$ which are given by their generating function

$$\frac{ze^{tz}}{e^z - 1} = \sum_{n=0}^{\infty} \frac{B_n(t)}{n!} z^n \qquad (|z| < 2\pi) \tag{1.17}$$

and which have the form

$$B_n(t) = \sum_{\nu=0}^{n} \binom{n}{\nu} B_\nu t^{n-\nu} \tag{1.18}$$

where $B_n = B_n(0)$ are the Bernoulli numbers. They satisfy the difference equations

$$B_n(t+1) - B_n(t) = nt^{n-1} \tag{1.19}$$

which imply the sum formula

$$\sum_{i=0}^{N-1} i^n = \tilde{B}_n(N) \tag{1.20}$$

with the modified Bernoulli polynomials

$$\tilde{B}_n(t) = \frac{1}{n+1}\{B_{n+1}(t) - B_{n+1}\} \tag{1.21}$$

of degree $n+1$ which have the generating function

$$\sum_{n=0}^{\infty} \frac{\tilde{B}_n(t)}{n!} z^n = \frac{e^{tz} - 1}{e^z - 1} \qquad (|z| < 2\pi). \tag{1.22}$$

Finally, let us mention a new property of the Bernoulli polynomials. We show that the polynomial $\frac{1}{p}\tilde{B}_k(p) - (\frac{p-1}{2})^k$ is divisible by $p+1$. For more details cf. Remark 5.8.

2 Functional equations with a parameter

As in the Introduction mentioned we consider the two-scale difference equation (1.1) with a parameter $x \in X \subseteq \mathbb{R}$ and coefficients $c_r(x)$ satisfying $c_0(x)c_{p-1}(x) \neq 0$ and

$$c_0(x) + \cdots + c_{p-1}(x) = 1 \qquad (x \in X). \tag{2.1}$$

We investigate solutions $\varphi(t,x) : \mathbb{R} \times X \mapsto \mathbb{R}$ satisfying the conditions

$$\varphi(t,x) = 0 \quad \text{for} \quad t < 0, \qquad \varphi(t,x) = 1 \quad \text{for} \quad t > 1 \tag{2.2}$$

and all $x \in X$. It is easy to see that if $\varphi(t,x)$ is a solution of the following system of equations

$$\varphi\left(\frac{r+t}{p}, x\right) = c_r(x)\varphi(t,x) + g_r(x) \qquad (t \in [0,1], \quad x \in X) \tag{2.3}$$

with

$$g_r(x) = \sum_{k=0}^{r-1} c_k(x) \tag{2.4}$$

so that $g_0(x) = 0$ and $g_p(x) = 1$ for all $x \in X$ then the function

$$\varphi_0(t,x) := \begin{cases} 0 & \text{for} \quad t < 0 \\ \varphi(t,x) & \text{for} \quad 0 \leq t \leq 1 \\ 1 & \text{for} \quad t > 1 \end{cases} \tag{2.5}$$

is a solution of (1.1). We are interested to solutions of (1.1) which are continuous with respect to t and differentiable with respect to x. The set of all such functions we denote by **D**. If $\varphi(t,x)$ belongs to **D** then it follows by differentiation of (1.1) with respect to x that the partial derivative $\varphi_x(t,x) = \frac{\partial}{\partial x}\varphi(t,x)$ satisfies

$$\varphi_x\left(\frac{t}{p}, x\right) = \sum_{r=0}^{p-1} c_r(x)\varphi_x(t-r,x) + \Psi_1(t,x) \qquad (t \in \mathbb{R}, x \in X) \tag{2.6}$$

where

$$\Psi_1(t,x) = \sum_{r=0}^{p-1} c_r'(x)\varphi(t-r,x) \qquad (t \in \mathbb{R}, x \in X) \tag{2.7}$$

and by differentiation of (2.3) we get for $r \in \{0, 1, \ldots, p-1\}$ the equations:

$$\varphi_x\left(\frac{r+t}{p}, x\right) = c_r(x)\varphi_x(t,x) + \psi_r(t,x) \qquad (t \in [0,1], \, x \in X) \tag{2.8}$$

where

$$\psi_r(t,x) = c_r'(x)\varphi(t,x) + g_r'(x). \tag{2.9}$$

Theorem 2.1 *Assume that $c_0(x), c_1(x), \dots, c_{p-1}(x)$ $(x \in X)$ are differentiable functions with bounded derivatives and that*

$$L := \sup\{|c_0(x)|, \dots, |c_{p-1}(x)| : x \in X\} < 1. \tag{2.10}$$

Then there exists exactly one solution $\varphi(t,x) \in \mathbf{D}$ of (1.1) satisfying (1.3). Moreover, there exists the partial derivative $\varphi_x(x,t)$ which satisfies (2.8) and which is continuous with respect to t.

Proof: First we determine the solution $\varphi(t,x)$ for $(t,x) \in [0,1] \times X$. For this we put $L' := \sup\{|c_0'(x)|, \dots, |c_{p-1}'(x)| : x \in X\}$ and choose $\varepsilon > 0$ so small that $K := L + \varepsilon L' < 1$. Note that $\mathbf{D}$ is a Banach space with the norm

$$\|u\|_{\mathbf{D}} := \|u\|_\infty + \varepsilon \left\| \frac{\partial u}{\partial x} \right\|_\infty$$

where $\|u\|_\infty = \sup\{|u(x,t)| : (x,t) \in [0,1] \times X\}$, and that

$$\Omega := \{u \in \mathbf{D} : \; u(0,x) = 0, \; u(1,x) = 1, \; \forall x \in X\}$$

is a closed subset of $\mathbf{D}$. For $u \in \Omega$ we define an operator T for all $x \in X$ by $(Tu)(0,x) := 0$ and

$$(Tu)(t,x) := c_r(x)u(pt-r,x) + g_r(x) \qquad \text{for} \quad t \in \left(\frac{r}{p}, \frac{r+1}{p}\right]$$

where $r = 0, 1, \dots, p-1$ so that in view of $u(1,x) = 1$ for all $x \in X$ and (2.4)

$$(Tu)\left(\frac{r+1}{p}, x\right) = c_r(x)u(1,x) + g_r(x) = c_r(x) + g_r(x) = g_{r+1}(x).$$

We show that T maps Ω into itself. At first we have $(Tu)(0,x) = 0$ and $(Tu)(1,x) = g_p(x) = 1$ by (2.4) and (2.1). Next, $(Tu)(t,x)$ is continuous with respect to t. This is clear at each point (t,x) with $t \in (\frac{r}{p}, \frac{r+1}{p})$ $(r = 0, \dots, p-1)$ and left-hand continuous at $t = \frac{r+1}{p}$. But it is also right-hand continuous at $t = \frac{r+1}{p}$ with $r = 0, 1, \dots, p-2$ since for $0 < h < 1$ we have

$$(Tu)\left(\frac{r+1+h}{p}, x\right) = c_{r+1}(x)u(1+h,x) + g_{r+1}(x)$$

which converges to $c_{r+1}(x)u(1,x) + g_{r+1}(x) = g_{r+2}(x) = (Tu)(\frac{r+1}{p}, x)$ as $h \to 0$.

Moreover, for $x \in X$ the partial derivative $\frac{\partial}{\partial x}Tu$ exists and is given for $(0,x)$ by $\frac{\partial}{\partial x}(Tu)(0,x) = 0$ and for (t,x) with $t \in (\frac{r}{p}, \frac{r+1}{p}]$ by

$$\frac{\partial}{\partial x}(Tu)(x,t) = c_r'(x)u(pt-r,x) + c_r(x)\frac{\partial}{\partial x}u(pt-r,x) + g_r'(x).$$

Hence, indeed T maps Ω into Ω. For $u_1, u_2 \in \Omega$ it holds

$$\|Tu_1 - Tu_2\|_\infty \leq L\|u_1 - u_2\|_\infty$$

and

$$\left\|\frac{\partial}{\partial x}Tu_1 - \frac{\partial}{\partial x}Tu_2\right\|_\infty \leq L'\|u_1 - u_2\|_\infty + L\left\|\frac{\partial}{\partial x}u_1 - \frac{\partial}{\partial x}u_2\right\|_\infty .$$

Therefore, and in view of $K = L + \varepsilon L'$ we have

$$\begin{aligned} \|Tu_1 - Tu_2\|_{\mathbf{D}} &= \|Tu_1 - Tu_2\|_\infty + \varepsilon\left\|\frac{\partial}{\partial x}Tu_1 - \frac{\partial}{\partial x}Tu_2\right\|_\infty \\ &\leq L\|u_1 - u_2\|_\infty + \varepsilon L'\|u_1 - u_2\|_\infty + \varepsilon L\left\|\frac{\partial}{\partial x}u_1 - \frac{\partial}{\partial x}u_2\right\|_\infty \\ &\leq K\|u_1 - u_2\|_{\mathbf{D}} \end{aligned}$$

By Banach's fixed point theorem there exists exactly one fixed point, i.e. (1.1) has exactly one solution $\varphi(t, x) \in \mathbf{D}$ which is defined for $(t, x) \in [0, 1] \times X$ with $\varphi(0, x) = 0$ and $\varphi(1, x) = 1$ for all $x \in X$. Hence, if we continue $\varphi(t, x)$ by $\varphi(t, x) = 0$ for $t < 0$ and $\varphi(t, x) = 1$ for $t > 1$ then we get a solution $\varphi(t, x) : \mathbb{R} \mapsto \mathbb{R} \times X$ of (1.1) which is continuous with respect to t.

Next we show that the partial derivative $\varphi_x(t, x)$ is continuous with respect to $t \in [0, 1]$. Differentiation of (2.3) with respect to x yields that $\varphi_x(t, x)$ satisfies the equations (2.8) where the functions (2.9) are continuous with respect to t. It follows by the result of GIRGENSOHN [7, Theorem 1] that for each $x \in X$ the function $\varphi_x(t, x)$ is continuous with respect to $t \in [0, 1]$.

It remains to show that $\varphi_x(0, x) = \varphi_x(1, x) = 0$ for all $x \in X$. According to (2.8) and (2.9), both with $t = 0$, we have

$$\varphi_x(0, x) = c_0(x)\varphi_x(0, x) + \psi_0(0, x)$$

where $\psi_0(0, x) = g_0'(x) = 0$, see (2.4). In view of $c_0(x) \neq 0$ and $c_0(x) \neq 1$, see (2.10), it follows $\varphi_x(0, x) = 0$. Moreover, (2.8) and (2.9) imply for $r = p - 1$ and $t = 1$

$$\varphi_x(1, x) = c_{p-1}(x)\varphi_x(1, x) + \psi_{p-1}(x)$$

where

$$\begin{aligned} \psi_{p-1}(x) &= c_{p-1}'(x)\varphi(1, x) + g_{p-1}'(x) \\ &= c_{p-1}'(x) + g_{p-1}'(x) \\ &= g_p'(x) \end{aligned}$$

owing to (2.4) with $r = p$. But $g_p(x) = c_0(x) + \cdots + c_{p-1}(x) = 1$ for all $x \in X$ according to (2.1) so that $\psi_{p-1}(x) = g_p'(x) = 0$. It follows $\varphi_x(1, x) = c_{p-1}(x)\varphi_x(1, x)$ and hence $\varphi_x(1, x) = 0$ since $c_{p-1}(x) \neq 1$, see (2.10). □

Theorem 2.2 *Under the suppositions of Theorem 2.1 it holds that if the functions $c_0(x), \ldots, c_{p-1}(x)$ are k-times differentiable then the solution $\varphi(t,x)$ of (1.1) is also k-times differentiable with respect to x and the k-th partial derivative*

$$\varphi_x^{(k)}(t,x) := \frac{\partial^k}{\partial x^k}\varphi(t,x) \tag{2.11}$$

is continuous with respect to t. For $k \geq 1$ we have $\varphi^{(k)}(t,x) = 0$ for $t \notin (0,1)$ and all $x \in X$, and

$$\varphi_x^{(k)}\left(\frac{t}{p}, x\right) = \sum_{r=0}^{p-1} c_r(x)\varphi_x^{(k)}(t-r,x) + \Psi_k(t,x) \qquad (t \in \mathbb{R}, x \in X) \tag{2.12}$$

where $\Psi_k(t,x)$ is recursively given by (2.7) and

$$\Psi_k(t,x) = \sum_{r=0}^{p-1} c_r'(x)\varphi_x^{(k-1)}(t-r,x) + \frac{\partial}{\partial x}\Psi_{k-1}(t,x). \tag{2.13}$$

Proof: The first part is a consequence of Theorem 2.1. Starting with (2.6) equation (2.12) with (2.13) can be proved by induction. If (2.12) with (2.13) is true for $k-1$ then by the product rule for the differentiation we get

$$\varphi_x^{(k)}\left(\frac{t}{p}, x\right) = \sum_{r=0}^{p-1} c_r(x)\varphi_x^{(k)}(t-r,x) + \sum_{r=0}^{p-1} c_r'(x)\varphi_x^{(k-1)}(t-r,x) + \frac{\partial}{\partial x}\Psi_{k-1}(t,x).$$

So the proof is complete. □

Next we use so-called Knopp function of the form

$$H(t) = \sum_{j=0}^{\infty} \frac{h(p^j t)}{p^j} \qquad (t \in \mathbb{R}) \tag{2.14}$$

with the generating, 1-periodic function $h(t)$ with $h(0) = 0$, cf. [11] or [16]. Obviously, (2.14) implies

$$H\left(\frac{t}{p}\right) = h\left(\frac{t}{p}\right) + \frac{1}{p}H(t) \qquad (t \in \mathbb{R}) \tag{2.15}$$

Conversely, if $H(.)$ satisfies (2.15) then $H(.)$ has the form (2.14).

Proposition 2.3 *Under the suppositions of Theorem 2.2 it holds that in case $c_r(0) = \frac{1}{p}$ for $r = 0, 1, \ldots, p-1$ we have for $\varphi_x^{(k)}(t,0)$ with $k \geq 1$ the representation*

$$\varphi_x^{(k)}(t,0) = \sum_{j=0}^{\infty} \frac{1}{p^j} h_k(p^j t) \qquad (0 \leq t \leq 1)$$

where h_k is 1-periodic continuous function given by $h_k(t) = \Psi_k(t,0)$ for $0 \leq t < 1$.

Proof: Let $f(t)$ be a function of period 1 given by $f(t) = \varphi^{(k)}(t,0)$ for $0 \leq t \leq 1$. The relation $\varphi^{(k)}(t,x) = 0$ for $k \geq 1$, all $t \notin (0,1)$ and $x \in X$ implies $f(0) = f(1) = 1$. In view of $c_r(0) = \frac{1}{p}$ equation (2.12) for $x = 0$ and with $r + t$ in place of t yields

$$f\left(\frac{r+t}{p}\right) = \frac{1}{p} f(t) + \Psi_k\left(\frac{r+t}{p}, 0\right) \qquad (0 \leq r \leq p-1, 0 \leq t \leq 1)$$

and $f(0) = f(1) = 0$ implies $\Psi_k(0,0) = \Psi(1,0) = 0$, i.e. $h_k(0) = h_k(1) = 0$. It follows that

$$f(t) = \sum_{j=0}^{\infty} \frac{1}{p^j} h_k(p^j t) \qquad (t \in \mathbb{R}).$$

So the theorem is proved. □

3 Digital exponential sums

For integer $N \geq 1$ we investigate the digital exponential sum

$$S(N,x) := \sum_{n=0}^{N-1} e^{xs(n)} \qquad (x \in \mathbb{R}) \tag{3.1}$$

where $s(n)$ denotes the sum of digits of the integer n in the p-adic representation of n. For this we begin with a results of [19] concerning a formula for the sum

$$S(N) := \sum_{n=0}^{N-1} C_n \tag{3.2}$$

where C_n is an arbitrary sequence which is given by the p initial values $C_0 = 1, C_1, \ldots, C_{p-1}$ such that

$$C := C_0 + \cdots + C_{p-1} > 0 \tag{3.3}$$

and which satisfies the recurrence formula

$$C_{kp+r} = C_k C_r \qquad (k \in \mathbb{N}, r = 0, 1, \ldots, p-1). \tag{3.4}$$

If the conditions (3.3) and (3.4) are fulfilled then the two-scale difference equation

$$\varphi\left(\frac{t}{p}\right) = \frac{1}{C} \sum_{r=0}^{p-1} C_r \varphi(t-r)$$

with C from (3.3) has in case $|C_r| < C$ a continuous solution $\varphi = \varphi_0$ satisfying $\varphi_0(t) = 0$ for $t < 0$ and $\varphi_0(t) = 1$ for $t > 1$, cf. [19], and it holds

Proposition 3.1 ([19]) *For $n \in \mathbb{N}$ the sum (3.2) can be represented as*

$$S(N) = N^\alpha F(\log_p N)$$

with $\alpha = \log_p C$ and an 1- periodic continuous function F which is given by

$$F(u) = \frac{\varphi_0(p^u)}{p^{\alpha u}} \qquad (u \leq 0).$$

Certainly, Proposition 3.1 is also valid if we consider a sequence $C_n = C_n(x)$ depending on a parameter x provided that the above conditions are fulfilled. In particular, for the sequence

$$C_n(x) := e^{xs(n)} \qquad (x \in \mathbb{R}) \tag{3.5}$$

we have

$$C(x) := C_0(x) + C_1(x) + \cdots + C_{p-1}(x) = 1 + e^x + \cdots + e^{(p-1)x} \tag{3.6}$$

since $s(n) = n$ for $n = 0, 1, \ldots, p-1$ and in view of $s(kp+r) = s(k) + s(r)$ for $k = 0, 1, 2, \ldots$ and $r = 0, 1, \ldots, p-1$. Hence, we have that

$$C_{kp+r}(x) = e^{xs(kp+r)} = e^{x(s(k)+s(r))} = C_k(x)C_r(x),$$

cf. (3.4). For $c_r(x) = \frac{C_r(x)}{C(x)}$ we have

$$c_0(x) + c_1(x) + \ldots + c_{p-1}(x) = 1$$

and $c_r(0) = \frac{1}{p}$ for $r \in \{0, 1, \ldots, p-1\}$. By Theorem 2.2 equation (1.1) with the actual coefficients $c_r(x)$, i.e.

$$\varphi\left(\frac{t}{p}, x\right) = \sum_{r=0}^{p-1} \frac{e^{rx}}{C(x)} \varphi(t-r, x) \qquad (t \in \mathbb{R}) \tag{3.7}$$

with $C(x)$ from (3.6) has a solution $\varphi_0(t, x)$ which is continuous with respect to t and arbitrary often differentiable with respect to $x \in X$.

According to Proposition 3.1 we have

Proposition 3.2 *For $N \in \mathbb{N}$ the sum $S(N, x)$ from (3.1) can be represented as*

$$S(N, x) = N^{\alpha(x)} F_0(\log_p N, x) \tag{3.8}$$

where

$$\alpha(x) = \log_p C(x) \tag{3.9}$$

with $C(x)$ from(3.6) and an 1-periodic function $F_0(u, x)$ with respect to u which is given by

$$F_0(u, x) = \frac{\varphi_0(p^u, x)}{p^{\alpha(x)u}} \qquad (u \leq 0, x \in \mathbb{R}) \tag{3.10}$$

which is continuous and 1-periodic with respect to u.

By Theorem 2.2 the function $F_0(u,x)$ is arbitrary often differentiable with respect to x and the k-th partial derivative

$$F_k(u,x) := \frac{\partial^k}{\partial x^k} F_0(u,x) \tag{3.11}$$

is 1-periodic with respect to u. So the functions

$$F_k(u) := F_k(u,0) \tag{3.12}$$

are continuous and 1-periodic.

Proposition 3.3 *We have*

$$F_0(u) = 1. \tag{3.13}$$

Proof: For $x = 0$ equation (3.7) takes the form

$$\varphi\left(\frac{t}{p}, 0\right) = \sum_{r=0}^{p-1} \frac{1}{p} \varphi(t-r, 0) \qquad (t \in \mathbb{R})$$

with the unique solution $\varphi(t,0) = t$ for $0 \le t \le 1$, cf. [18]. Further, owing to (3.9) we get

$$\alpha(0) = \log_p C(0) = \log_p p = 1$$

and hence for $u \le 0$

$$F_0(u,0) = \frac{\varphi_0(p^u, 0)}{p^{\alpha(0)u}} = \frac{p^u}{p^u} = 1 \qquad (u \le 0).$$

The periodicity of $F_0(u)$ implies $F_0(u) = 1$ for all real u. □

4 Power sums of the sum of digits

Now, for integer $N \ge 1$ we investigate the power sums of the sum of digits

$$S_k(N) = \sum_{n=0}^{N-1} s(n)^k \tag{4.1}$$

where $k \ge 0$ is an integer and use that

$$\frac{\partial^k}{\partial x^k} S(N,x)\bigg|_{x=0} = \sum_{n=0}^{N-1} s(n)^k e^{xs(n)}\bigg|_{x=0} = S_k(N).$$

So according to Proposition 3.2 we have

Proposition 4.1 *For $N \in \mathbb{N}$ the power sum (4.1) can be represented as*

$$S_k(N) = \frac{\partial^k}{\partial x^k} N^{\alpha(x)} F_0(\log_p N, x)\bigg|_{x=0} \tag{4.2}$$

As abbreviation we put

$$c(x) = \frac{C'(x)}{C(x)} = \frac{e^x + 2e^{2x} + \cdots + (p-1)e^{(p-1)x}}{1 + e^x + \cdots + e^{(p-1)x}} \tag{4.3}$$

with $C(x)$ from (3.6). In particular

$$c(0) = \frac{1}{p} S_1(p) = \frac{p-1}{2}. \tag{4.4}$$

For $k \in \mathbb{N}$ and $\ell = 0, \ldots, k$ we introduce functions $c_{k,\ell}(x)$ as follows:

$$\left.\begin{array}{lcll} c_{k,k}(x) & = & c(x)^k & \text{for} \quad k \geq 0 \\ c_{k,0}(x) & = & 0 & \text{for} \quad k \geq 1 \\ c_{k+1,\ell}(x) & = & c_{k,\ell-1}(x)c(x) + c'_{k,\ell}(x) & \text{for} \quad k \geq 1, \ \ 1 \leq \ell \leq k \end{array}\right\} . \tag{4.5}$$

So $c_{0,0}(x) = 1$ in view of $c(x) \neq 0$. It is easy to see that for $k \geq 1$ it holds

$$c_{k,1}(x) = c^{(k-1)}(x), \qquad c_{k,k-1}(x) = \frac{k(k-1)}{2} c(x)^{k-2} c'(x). \tag{4.6}$$

Lemma 4.2 *For integer $k \geq 0$ we have*

$$\frac{d^k}{dx^k} N^{\alpha(x)} = N^{\alpha(x)} \sum_{\ell=0}^{k} c_{k,\ell}(x)(\log_p N)^\ell. \tag{4.7}$$

Proof: Formula (4.7) is true for $k = 0$ since $c_{0,0}(x) = 1$. In view of $\alpha(x) = \log_p C(x)$ we get

$$\frac{d}{dx} N^{\alpha(x)} = N^{\alpha(x)} \log N \frac{1}{\log p} \frac{C'(x)}{C(x)} = N^{\alpha(x)} c(x) \log_p N$$

i.e. for $k = 1$ formula (4.7) is true, too. Assume (4.7) is true for a fixed k. Then we get

$$\frac{d^{k+1}}{dx^{k+1}} N^{\alpha(x)} = N^{\alpha(x)} \sum_{\ell=0}^{k} c_{k,\ell}(x)c(x)(\log_p N)^{\ell+1} + N^{\alpha(x)} \sum_{\ell=0}^{k} c'_{k,\ell}(x)(\log_p N)^\ell$$

which in view of (4.5) and $c_{k,k}(x)c(x) = c(x)^{k+1} = c_{k+1,k+1}(x)$ yields the assertion. □

Theorem 4.3 *For integer $N \geq 0$ we have*

$$\sum_{n<N} s(n)^k e^{xs(n)} = N^{\alpha(x)} \sum_{\ell=0}^{k} (\log_p N)^\ell H_\ell(\log_p N, x) \tag{4.8}$$

with $\alpha(x) = \log_p C(x)$ and

$$H_\ell(u,x) = \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} c_{k-\kappa,\ell}(x) F_\kappa(u,x) \tag{4.9}$$

with $F_k(u,x)$ from (3.11).

Proof: We use (3.1) and (3.8) with $\alpha = \alpha(x) = \log_p C(x)$. By Leibniz's formula and Lemma 4.2 we have with $u = \log_p N$

$$\begin{aligned}
\frac{\partial^k}{\partial x^k} N^\alpha F_0(u,x) &= \sum_{\kappa=0}^{k} \binom{k}{\kappa} \frac{\partial^\kappa}{\partial x^\kappa} N^\alpha \frac{\partial^{k-\kappa}}{\partial x^{k-\kappa}} F_0(u,x) \\
&= N^\alpha \sum_{\kappa=0}^{k} \binom{k}{\kappa} \sum_{\ell=0}^{\kappa} u^\ell c_{\kappa,\ell}(x) F_{k-\kappa}(u,x) \\
&= N^\alpha \sum_{\ell=0}^{k} \sum_{\kappa=\ell}^{k} \binom{k}{\kappa} u^\ell c_{\kappa,\ell}(x) F_{k-\kappa}(u,x).
\end{aligned}$$

Replacing κ by $k-\kappa$ we get

$$\frac{\partial^k}{\partial x^k} N^\alpha F_0(u,x) = N^\alpha \sum_{\ell=0}^{k} \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} u^\ell c_{k-\kappa,\ell}(x) F_\kappa(u,x)$$

which in view of (3.1) and (3.8) yields (4.8) with (4.9). □

Corollary 4.4 *For the power sum (4.1) we have*

$$\frac{1}{N} S_k(N) = \sum_{\ell=0}^{k} (\log_p N)^\ell H_\ell(\log_p N) \tag{4.10}$$

where $H_\ell(u) = H_\ell(u,0)$ from (4.9), i.e.

$$H_\ell(u) = \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} c_{k-\kappa,\ell} F_\kappa(u) \tag{4.11}$$

with

$$c_{k,\ell} := c_{k,\ell}(0) \tag{4.12}$$

and $F_k(u)$ from (3.12).

Lemma 4.5 *For integer $k \geq 0$ we have*

$$c^{(k)}(0) = \frac{(-1)^{k+1}B_{k+1}}{k+1}(p^{k+1}-1) \tag{4.13}$$

with the Bernoulli numbers B_{k+1}.

Proof: From (4.3) we have by Leibniz's formula

$$c^{(k)}(x) = \sum_{n=0}^{k} \binom{k}{n} C^{(n+1)}(x) \left(\frac{1}{C(x)}\right)^{(k-n)} . \tag{4.14}$$

In order to compute $c^{(k)}(0)$ first note that

$$C^{(n+1)}(0) = 1 + 2^{n+1} + \ldots + (p-1)^{n+1} = \tilde{B}_{n+1}(p), \tag{4.15}$$

cf. (1.20). Moreover, in view of

$$C(x) = 1 + e^x + \cdots + e^{(p-1)x} = \frac{e^{px}-1}{e^x-1}$$

and

$$\frac{1}{C(x)} = \frac{e^x-1}{e^{px}-1} = \frac{e^{\frac{1}{p}(px)}-1}{e^{px}-1} = \sum_{n=0}^{\infty} \frac{\tilde{B}_n(\frac{1}{p})}{n!}(px)^n \qquad \left(|x| < \frac{2\pi}{p}\right),$$

cf. (1.22) with $z = px$ and $t = \frac{1}{p}$, we find

$$\left(\frac{1}{C(x)}\right)^{(n)}\Bigg|_{x=0} = p^n \tilde{B}_n\left(\frac{1}{p}\right) . \tag{4.16}$$

From (4.14), (4.15) and (4.16) we get

$$c^{(k)}(0) = \sum_{n=0}^{k} \binom{k}{n} \tilde{B}_{n+1}(p) p^{k-n} \tilde{B}_{k-n}\left(\frac{1}{p}\right) .$$

In view of the Cauchy product of two power series we see that

$$\sum_{k=0}^{\infty} \frac{c^{(k)}(0)}{k!} z^k = \sum_{n=0}^{\infty} \frac{\tilde{B}_{n+1}(p)}{n!} z^n \sum_{n=0}^{\infty} \frac{p^n \tilde{B}_n(\frac{1}{p})}{n!} z^n \tag{4.17}$$

which in view of (1.22) is convergent for $|z| < \frac{2\pi}{p}$. Moreover, from (1.22) we find

$$\sum_{n=0}^{\infty} \frac{\tilde{B}_{n+1}(p)}{n!} z^n = \frac{d}{dz}\left(\frac{e^{pz}-1}{e^z-1}\right) = \frac{pe^{pz}(e^z-1)-(e^{pz}-1)e^z}{(e^z-1)^2},$$

$$\sum_{n=0}^{\infty} \frac{\tilde{B}_n(\frac{1}{p})}{n!}(pz)^n = \frac{e^z-1}{e^{pz}-1},$$

and by (4.17) we get

$$\sum_{k=0}^{\infty} \frac{c^{(k)}(0)}{k!} z^k = \frac{pe^{pz}}{e^{pz}-1} - \frac{e^z}{e^z-1} \qquad (z \neq 0, |z| < 2\pi).$$

According to (1.17) we have for $z \neq 0$ and $|z| < 2\pi$

$$\frac{e^z}{e^z-1} = \frac{1}{1-e^{-z}} = -\sum_{n=0}^{\infty} \frac{B_n}{n!}(-z)^{n-1}$$

and

$$\frac{pe^{pz}}{e^{pz}-1} = -\sum_{n=0}^{\infty} \frac{pB_n}{n!}(-pz)^{n-1}.$$

Hence,

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{c^{(k)}(0)}{k!} z^k &= \sum_{n=0}^{\infty} \frac{(-1)^n B_n}{n!}(p^n-1)z^{n-1} \\ &= \sum_{k=0}^{\infty} \frac{(-1)^{k+1} B_{k+1}}{(k+1)!}(p^{k+1}-1)z^k \end{aligned}$$

with $k = n-1$. This implies (4.13). □

Remark 4.6 Note that in case $p = 2$ we have $c(x) = \frac{e^x}{1+e^x} = 1 - \frac{1}{1+e^x}$ so that in view of

$$\frac{1}{1+e^x} = \frac{1}{2}\left(1 - \tanh\left(\frac{x}{2}\right)\right)$$

we get $c^{(k)}(0) = \frac{(-1)^{k+1} B_{k+1}}{k+1}(2^{k+1}-1)$.

5 Specific power sums

We begin with a formula for the digital power sum

$$S_k(pN) = \sum_{n<pN} s(n)^k. \tag{5.1}$$

Proposition 5.1 *For the sums (5.1) we have*

$$S_k(pN) = \sum_{\ell=0}^{k} \binom{k}{\ell} S_{k-\ell}(p) S_\ell(N) \tag{5.2}$$

where $S_0(N) = N$.

Proof: Write $n = pm + r$ with $0 \leq r \leq p - 1$ and $0 \leq m \leq N - 1$, we get in view of $s(pm + r) = s(m) + s(r)$ that

$$\begin{aligned}\sum_{n<pN} s(n)^k &= \sum_{r=0}^{p-1} \sum_{m<N} \{s(m) + s(r)\}^k \\ &= \sum_{r=0}^{p-1} \sum_{m<N} \sum_{\kappa=0}^{k} \binom{k}{\kappa} s(m)^\kappa s(r)^{k-\kappa} \\ &= \sum_{\kappa=0}^{m} \binom{k}{\kappa} \sum_{r=0}^{p-1} s(r)^{k-\kappa} \sum_{m<N} s(m)^\kappa .\end{aligned}$$

This yields the assertion. □

For $N = p^n$ we get from (4.10)

$$\frac{1}{p^n} S_k(p^n) = \sum_{\ell=0}^{k} a_{k,\ell}\, n^\ell \tag{5.3}$$

with coefficients $a_{k,\ell}$ depending on p which owing to (4.11) are given by

$$a_{k,\ell} = \sum_{\kappa=0}^{k-\kappa} \binom{k}{\kappa} c_{k-\kappa,\ell} F_\kappa(0), \tag{5.4}$$

cf. [4]. In particular for $n = 1$ we have $S_k(p) = \tilde{B}_k(p)$ with the modified Bernoulli polynomials $\tilde{B}_k(\cdot)$, cf. (1.21), and (5.3) implies

$$\sum_{\ell=0}^{k} a_{k,\ell}(p) = \frac{1}{p} \tilde{B}_k(p). \tag{5.5}$$

Lemma 5.2 *For the coefficients $a_{k,\ell} = a_{k,\ell}(p)$ we have the relation*

$$\sum_{\nu=\kappa+1}^{k} \binom{\nu}{\kappa} a_{k,\nu} = \sum_{\ell=\kappa}^{k-1} \binom{k}{\ell} \frac{1}{p} S_{k-\ell}(p) a_{\ell,\kappa} \tag{5.6}$$

Proof: We use (5.2) with $N = p^n$. By (5.3) we have

$$\begin{aligned}\frac{1}{p^{n+1}} S_k(p^{n+1}) &= \sum_{\ell=0}^{k} a_{k,\ell}(n+1)^\ell \\ &= \sum_{\ell=0}^{k} a_{k,\ell} \sum_{\kappa=0}^{\ell} \binom{\ell}{\kappa} n^\kappa \\ &= \sum_{\kappa=0}^{k} \sum_{\ell=\kappa}^{m} a_{k,\kappa} \binom{\ell}{\kappa} n^\kappa\end{aligned}$$

and

$$\begin{aligned}\sum_{\kappa=0}^{k}\binom{k}{\kappa}\frac{1}{p}S_{k-\kappa}(p)\frac{1}{p^n}S_\kappa(p^n) &= \sum_{\kappa=0}^{k}\binom{k}{\kappa}\frac{1}{p}S_{k-\kappa}(p)\sum_{k=0}^{\kappa}a_{\ell,\kappa}n^\kappa \\ &= \sum_{\kappa=0}^{k}\sum_{\ell=\kappa}^{k}\binom{k}{\ell}\frac{1}{p}S_{k-\ell}(p)a_{\ell,\kappa}n^\kappa.\end{aligned}$$

According to (5.2) we get

$$\sum_{\nu=\kappa}^{k}\binom{\nu}{\kappa}a_{k,\nu}=\sum_{\ell=\kappa}^{k}\binom{k}{\ell}\frac{1}{p}S_{k-\ell}(p)a_{\ell,\kappa}$$

which yields (5.6). □

We already know from Proposition 3.3 that $F_0(u)=1$ for $u\in\mathbb{R}$.

Proposition 5.3 *For the values $F_k(0)$ with $k\geq 1$ we have*

$$F_k(0)=0 \qquad (k\geq 1). \tag{5.7}$$

For all $k\in\mathbb{N}_0$ and $\ell=0,1,\dots,k$ we have $a_{k,\ell}=c_{k,\ell}$. In particular $a_{k,k}=(\frac{p-1}{2})^k$ for $k\geq 0$ and $a_{k,0}=0$ for $k\geq 1$. The further numbers $a_{k,\ell}$ are uniquely determined by

$$\ell a_{k,\ell}(p)=\sum_{\mu=\ell-1}^{k-1}\binom{k}{\mu}\frac{1}{p}S_{k-\mu}(p)a_{\mu,\ell-1}(p)-\sum_{\nu=\ell+1}^{k}\binom{\nu}{\ell-1}a_{k,\nu}(p). \tag{5.8}$$

Proof: From (5.4) we get $a_{k,k}=c_{k,k}F_0(0)=(\frac{p-1}{2})^k$ for $k\geq 0$ according to (4.5), (4.4) and $F_0(.)=1$. Formula (5.8) follows from (5.6). If $a_{k',\ell'}$ are given for $0\leq k'<k$, $0\leq\ell'\leq k'$ and for $k'=k$, $\ell<\ell'\leq k$ then $a_{k,\ell}$ is determined by (5.8).

Next we show that $a_{k,0}=0$ for $k\geq 1$. At first we get from (5.3) with (5.4) in case $k=1$ that

$$\frac{1}{p^n}S_1(p^n)=a_{1,0}+a_{1,1}n=a_{1,0}+\frac{1}{p}S_1(p)n$$

since $a_{1,1}=\frac{1}{p}S_1(p)$ and $n=1$ implies $a_{1,0}=0$. Now, equation (5.4) for $\kappa=0$ yields

$$\sum_{\nu=1}^{m}a_{k,\nu}=\sum_{\ell=0}^{k-1}\binom{k}{\kappa}\frac{1}{p}S_{k-\ell}(p)a_{\ell,0}.$$

Using (5.5) it follows

$$\frac{1}{p}S_k(p)-a_{k,0}=\frac{1}{p}S_k(p)+\sum_{\ell=1}^{k-1}\binom{k}{\ell}\frac{1}{p}S_{k-\ell}(p)a_{\ell,0}$$

and

$$a_{k,0}=-\sum_{\ell=1}^{k-1}\binom{k}{\ell}\frac{1}{p}S_{k-\ell}(p)a_{\ell,0}.$$

Now, $a_{1,0}=0$ implies $a_{k,0}=0$ for $k\geq 1$.

Finally we show that $a_{k,\ell} = c_{k,\ell}$. Equation (5.4) for $\ell = 0$ yields

$$a_{k,0} = \sum_{\kappa=0}^{k} \binom{k}{k} c_{k-\kappa,0} F_\kappa(0) = c_{0,0} F_k(0) = F_k(0),$$

i.e. $F_0(0) = 1$ and $F_k(0) = 0$ for $k \geq 1$. Now, equation (5.4) yields

$$a_{k,\ell} = \sum_{j=0}^{k} \binom{k}{j} c_{k-j,\ell} F_j(0) = c_{k,\ell} F_0(0) = c_{k,\ell}$$

and the proposition is proved completely. □

We already know that $a_{k,0} = 0$ for $k \geq 1$ and

$$a_{k,k} = \left(\frac{p-1}{2}\right)^k \qquad (k \geq 0). \tag{5.9}$$

The first $a_{k,\ell} = a_{k,\ell}(p)$ are computed by means of (5.8)

$$\begin{array}{lllll}
a_{1,1} = \frac{p-1}{2} & & & & \\
a_{2,1} = \frac{p^2-1}{12} & a_{2,2} = (\frac{p-1}{2})^2 & & & \\
a_{3,1} = 0 & a_{3,2} = \frac{(p-1)^2(p+1)}{8} & a_{3,3} = (\frac{p-1}{2})^3 & & \\
a_{4,1} = -\frac{p^4-1}{120} & a_{4,2} = \frac{(p-1)^2(p+1)^2}{48} & a_{4,3} = \frac{(p-1)^3(p+1)}{8} & a_{4,4} = (\frac{p-1}{2})^4 & \\
a_{5,1} = 0 & a_{5,2} = -\frac{(p-1)^2(p+1)(p^2+1)}{48} & a_{5,3} = \frac{5(p-1)^3(p+1)^2}{96} & a_{5,4} = \frac{5(p-1)^4(p+1)}{48} & a_{5,5} = (\frac{p-1}{2})^5
\end{array}$$

Figure 2. The first numbers $a_{k,\ell}$

In the following we need the

Proposition 5.4 *For $k \geq 1$ we have*

$$a_{k,1}(p) = \frac{(-1)^k B_k}{k}(p^k - 1) \tag{5.10}$$

with the Bernoulli numbers B_k, and

$$a_{k,k-1}(p) = \binom{k}{2} \left(\frac{p-1}{2}\right)^{k-1} \frac{p+1}{6}. \tag{5.11}$$

Proof: We use $a_{k,\ell}(p) = c_{k,\ell}(0)$, cf. Proposition 5.3, and both relations in (4.6), i.e. $c_{k,1}(x) = c^{(k-1)}(x)$ and $c_{k,k-1}(x) = \binom{k}{2} c(x)^{k-2} c'(x)$ with $c(x)$ from (4.3). First we compute $a_{k,1} = a_{k,1}(p)$. Now, $a_{k,1}(p) = c^{(k-1)}(0)$ so that (5.10) follows from Lemma 4.5.

We know that $a_{k,k-1} = \binom{k}{2} c^{k-2}(0)c'(0)$ where $c(0) = \frac{1}{p}(1+2+\cdots+(p-1)) = \frac{p-1}{2}$, i.e, $a_{k,k-1} = \binom{k}{2}(\frac{p-1}{2})^{k-2}c'(0)$ where $c'(0)$ is independent of k. In particular, $a_{2,1} = c'(0)$. From (5.10) we know that $a_{2,1} = \frac{p^2-1}{12}$ and it follows (5.11). □

Proposition 5.5 *For $k \geq 2$ and $1 \leq \ell < k$ we have that $a_{k,\ell}(p)$ are polynomials in p of degree at most k with $a_{k,\ell}(-1) = 0$. Moreover,*

$$a_{k,\ell}(p) = \left(\frac{p-1}{2}\right)^{\ell} \tilde{a}_{k,\ell}(p) \tag{5.12}$$

where $\tilde{a}_{k,\ell}(p)$ are polynomials in p of degree at most $k-\ell$ which are given by $\tilde{a}_{k,k}(p) = 1$ and

$$\ell \tilde{a}_{k,\ell}(p) = \sum_{\mu=\ell-1}^{k-1} \binom{k}{\mu} \frac{2\tilde{B}_{k-\mu}(p)}{p(p-1)} \tilde{a}_{\mu,\ell-1}(p) - \sum_{\nu=\ell+1}^{k} \binom{\nu}{\ell-1} \left(\frac{p-1}{2}\right)^{\nu-\ell} \tilde{a}_{k,\nu}(p). \tag{5.13}$$

Proof: From (5.4) we get $a_{k,k} = c_{k,k}F_0(0) = (\frac{p-1}{2})^k$ for $k \geq 0$ and $a_{k,0} = c_{0,0}F_k(0) = F_k(0)$ for $k \geq 1$. Assume that $a_{k',\ell'}(p)$ are given polynomials with *deg* $a_{k',\ell'}(p) \leq k'$ if $0 \leq k' < k$, $0 \leq \ell' \leq k'$ and if $k' = k$, $\ell < \ell' \leq k$. Then $a_{k,\ell}(p)$ is determined by (5.8) and *deg* $a_{k,\ell}(p) \leq k$.

We show that $a_{k,\ell}(-1) = 0$ for all $k \geq 2$ and $\ell = 0, \ldots, k-1$. This is true for $a_{k,0} = 0$ and $a_{k,k-1}$ according to (5.11) with $k \geq 2$. Assume that for $k \geq 2$ we have $a_{k',\ell'}(-1) = 0$ if $0 \leq k' < k$, $0 \leq \ell' \leq k'-1$ and if $k' = k$, $\ell < \ell' \leq k-1$. Then from (5.8) we get in view of $a_{\ell-1,\ell-1}(-1) = (-1)^{\ell-1}$ and $a_{k,k}(-1) = (-1)^k$, cf. (5.9), that

$$\begin{aligned} \ell a_{k,\ell}(-1) &= \binom{k}{\ell-1}(-1)\tilde{B}_{k-\ell+1}(-1)(-1)^{\ell-1} - \binom{k}{\ell-1}(-1)^k \\ &= \binom{k}{\ell-1}(-1)^{\ell}\left\{\tilde{B}_{k-\ell+1}(-1) - (-1)^{k-\ell}\right\} \\ &= 0 \end{aligned}$$

since $\tilde{B}_n(-1) = (-1)^{n-1}$ which follows from (1.19) with $t = -1$ and (1.21) where $B_n = B_n(0)$.

Next we show (5.12) which is true for all $a_{k,k}$ and $a_{k,0}$, $k = 0, 1, 2, \ldots$ since $a_{k,k} = (\frac{p-1}{2})^k$ and $a_{k,0} = 0$ for $k \geq 1$. Assume (5.12) is true for all $a_{k',\ell'}$ with $0 \leq k' < k$ and $0 \leq \ell' \leq k'$ as well as for $a_{k,\ell'}$ with $\ell < \ell' \leq k$. Then by division of (5.8) with $(\frac{p-1}{2})^{\ell}$ we get (5.13) which implies that indeed $\tilde{a}_{k,\ell}(p)$ is a polynomial in p and the supposition is proved by induction. □

Proposition 5.6 *For the power sum* (4.1) *with $N \in \mathbb{N}$ and $L = \log_p N$ we have*

$$\frac{1}{N} S_k(N) = \left(\frac{p-1}{2} L\right)^k + \sum_{\ell=0}^{k-1} \left(\frac{p-1}{2} L\right)^\ell \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} \tilde{a}_{k-\kappa,\ell}(p) F_\kappa(L) \tag{5.14}$$

where $\tilde{a}_{k,\ell}(p)$ are polynomials in p of degree at most $k-\ell$, given by $\tilde{a}_{k,k}(p) = 1$ and the recursion (5.13).

Proof: We use Corollary 4.4, $F_0(\cdot) = 1$ and $c_{k,\ell} = a_{k,\ell}(p) = (\frac{p-1}{2})^\ell \, \tilde{a}_{k,\ell}(p)$. So we get

$$\begin{aligned} \frac{1}{N} S_k(N) &= \sum_{\ell=0}^{k} L^\ell \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} \left(\frac{p-1}{2}\right)^\ell \tilde{a}_{k-\kappa,\ell}(p) F_\kappa(L) \\ &= \left(\frac{p-1}{2} L\right)^k + \sum_{\ell=0}^{k-1} \left(\frac{p-1}{2} L\right)^\ell \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} \tilde{a}_{k-\kappa,\ell}(p) F_\kappa(L) \end{aligned}$$

and (5.14) is proved. □

Remark 5.7 In view of (5.11) and (5.12) formula (5.14) yields for arbitrary integer k the asymptotic relation

$$\frac{1}{N} S_k(N) = \left(\frac{p-1}{2} L\right)^k + \left(\frac{p-1}{2} L\right)^{k-1} \left\{ \binom{k}{2} \frac{p+1}{6} + k F_1(L) \right\} + o(L^{k-1}) \tag{5.15}$$

as $N \to \infty$. In case $k = 1$ we get from (5.14) the formula of Trollope-Delange

$$\frac{1}{N} S_1(N) = \frac{p-1}{2} L + F_1(L), \tag{5.16}$$

in case $k = 2$

$$\frac{1}{N} S_2(N) = \left(\frac{p-1}{2} L\right)^2 + \frac{p-1}{2} L \left\{ \frac{p+1}{6} + 2F_1(L) \right\} + F_2(L) \tag{5.17}$$

which for $p = 2$ is known by Coquet, cf. [3], and in case $k = 3$

$$\begin{aligned} \frac{1}{N} S_3(N) &= \left(\frac{p-1}{2} L\right)^3 + \left(\frac{p-1}{2} L\right)^2 \left\{ \frac{p+1}{2} + 3F_1(L) \right\} \\ &\quad + \frac{p-1}{2} L \left\{ \frac{p+1}{3} F_1(L) + 3F_2(L) \right\} + F_3(L), \end{aligned}$$

cf. also [22] for $p = 2$ or [17, Theorem 6.3]. In case $k = 4$ we get

$$\begin{aligned} \frac{1}{N} S_4(N) &= \left(\frac{p-1}{2} L\right)^4 + \left(\frac{p-1}{2} L\right)^3 \{p + 1 + 4F_1(L)\} \\ &\quad + \left(\frac{p-1}{2} L\right)^2 \left\{ \frac{p+1}{12} + 2(p+1)F_1(L) + 6F_2(L) \right\} \\ &\quad + \frac{p-1}{2} L \left\{ -\frac{p+1}{60} + (p+1)F_2(L) + 4F_3(L) \right\} + F_4(L). \end{aligned}$$

Remark 5.8 In case $N = p$ we have $L = 1$, and in view of $F_0(1) = 1$ and $F_k(1) = 0$ for $k > 0$ as well as $S_k(p) = 1^k + 2^k + \cdots + (p-1)^k = \tilde{B}_k(p)$, cf. (1.20), we get from (5.14) that

$$\frac{1}{p}\tilde{B}_k(p) = \left(\frac{p-1}{2}\right)^k + \sum_{\ell=0}^{k-1}\left(\frac{p-1}{2}\right)^\ell \tilde{a}_{k,\ell}(p) \tag{5.18}$$

cf. also (5.5) and (5.12). Because $\tilde{a}_{k,\ell}(-1) = 0$ for $\ell < k$ we have that $\frac{1}{p}\tilde{B}_k(p) - (\frac{p-1}{2})^k$ is divisible by $p+1$. Hence $\frac{1}{n+1}\tilde{B}_k(n+1) - (\frac{n}{2})^k = \frac{1}{n+1}(1^k + 2^k + \cdots + n^k) - (\frac{n}{2})^k$ is divisible by $n+2$. So, in particular, for $k = 1, 2, \ldots$ we have

$$\begin{aligned}
\tfrac{1}{n+1}\sum_{i=1}^{n} i - \tfrac{n}{2} &= 0\\
\tfrac{1}{n+1}\sum_{i=1}^{n} i^2 - (\tfrac{n}{2})^2 &= \tfrac{1}{12}n(n+2)\\
\tfrac{1}{n+1}\sum_{i=1}^{n} i^3 - (\tfrac{n}{2})^3 &= \tfrac{1}{8}n^2(n+2)\\
\tfrac{1}{n+1}\sum_{i=1}^{n} i^4 - (\tfrac{n}{2})^4 &= \tfrac{1}{240}n(n+2)(33n^2+6n-4)\\
\tfrac{1}{n+1}\sum_{i=1}^{n} i^5 - (\tfrac{n}{2})^5 &= \tfrac{1}{96}n^2(n+2)(13n^2+6n-4)
\end{aligned}$$

and so on.

6 Power series and generating functions

We start with the power sums $S_k(p^n)$, cf. (4.1) with $N = p^n$.

Proposition 6.1 *For $n \in \mathbb{N}$ we have*

$$\sum_{k=0}^{\infty}\frac{1}{k!}S_k(p^n)z^k = \left(\frac{e^{pz}-1}{e^z-1}\right)^n \qquad (z \in \mathbb{C}). \tag{6.1}$$

Proof: We prove (6.1) by induction on n. In case $n = 1$ we use (1.20) and (1.22) so that

$$\sum_{k=0}^{\infty}\frac{1}{k!}S_k(p)z^k = \sum_{k=0}^{\infty}\frac{1}{k!}\tilde{B}_k(p)z^m = \frac{e^{pz}-1}{e^z-1}$$

where we have convergence for all $z \in \mathbb{C}$ in view of

$$\frac{e^{pz}-1}{e^z-1} = 1 + e^z + \cdots + e^{(p-1)z}.$$

Assume (6.1) is true for a certain $n \geq 0$. Then we have in view of Proposition 5.1 with $N = p^n$ and the Cauchy product of two power series

$$\begin{aligned}
\sum_{k=0}^{\infty}\frac{1}{k!}S_k(p^{n+1})z^k &= \left(\sum_{k=0}^{\infty}\frac{1}{k!}S_m(p^n)z^k\right)\left(\sum_{k=0}^{\infty}\frac{1}{k!}S_k(p)z^k\right) = \left(\frac{e^{pz}-1}{e^z-1}\right)^n\frac{e^{pz}-1}{e^z-1}\\
&= \left(\frac{e^{pz}-1}{e^z-1}\right)^{n+1}.
\end{aligned}$$

□

Now we consider the polynomials

$$P_k(t) := \sum_{\ell=0}^{k} a_{k,\ell}(p)t^\ell \tag{6.2}$$

with the coefficients $a_{k,\ell}(p)$ given by (5.8). We remember that in particular, $a_{k,k} = (\frac{p-1}{2})^k$ and that $\frac{1}{p^n}S_k(p^n) = P_k(n)$, cf. (5.3). According to Figure 2 the first polynomials $P_k(t)$ read:

$$\begin{aligned}
P_0(t) &= 1\\
P_1(t) &= \tfrac{p-1}{2}t\\
P_2(t) &= \tfrac{p^2-1}{12}t + (\tfrac{p-1}{2})^2t^2\\
P_3(t) &= \tfrac{(p-1)^2(p+1)}{8}t^2 + (\tfrac{p-1}{2})^3t^3\\
P_4(t) &= -\tfrac{p^4-1}{120}t + \tfrac{(p-1)^2(p+1)^2}{48}t^2 + \tfrac{(p-1)^3(p+1)}{8}t^3 + (\tfrac{p-1}{2})^4t^4\\
P_5(t) &= -\tfrac{(p-1)^2(p+1)(p^2+1)}{48}t^2 + \tfrac{5(p-1)^3(p+1)^2}{96}t^3 + \tfrac{5(p-1)^4(p+1)}{48}t^4 + (\tfrac{p-1}{2})^5t^5
\end{aligned}$$

Figure 3. The first polynomials $P_k(t)$

Proposition 6.2 *The polynomials (6.2) have the generating function*

$$\sum_{k=0}^{\infty} \frac{1}{k!}P_k(t)z^k = \left(\frac{e^{pz}-1}{p(e^z-1)}\right)^t \qquad (z \in \mathbb{C}) \tag{6.3}$$

and starting with $P_0(t) = 1$ they satisfy the recursions

$$P_k(t) = t\sum_{\ell=1}^{k} \binom{k-1}{\ell-1} P_{k-\ell}(t)a_{\ell,1}(p) \tag{6.4}$$

where

$$a_{\ell,1}(p) = \frac{(-1)^\ell B_\ell}{\ell}(p^\ell - 1)$$

with the Bernoulli numbers B_ℓ, cf. (5.10).

Proof: According to $\frac{1}{p^n}S_k(p^n) = P_k(n)$, cf. (5.3), and (6.1) we have for $n \in \mathbb{N}$

$$\sum_{k=0}^{\infty} \frac{1}{k!}P_k(n)z^k = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{1}{p^n}S_k(p^n)z^k = \left(\frac{e^{pz}-1}{p(e^z-1)}\right)^n \qquad (z \in \mathbb{C}) \tag{6.5}$$

so that (6.3) is true for $t = n \in \mathbb{N}$. We show that for all $t \in \mathbb{C}$

$$\left(\frac{e^{pz}-1}{p(e^z-1)}\right)^t = \sum_{k=0}^{\infty} \frac{1}{k!}Q_k(t)z^k \qquad (z \in \mathbb{C}) \tag{6.6}$$

where $Q_k(t)$ are polynomials with respect to t of degree k. For fix p and t we put

$$f(z) := \left(\frac{e^{pz}-1}{p(e^z-1)}\right)^t = (g(z))^t \tag{6.7}$$

with

$$g(z) = \frac{e^{pz}-1}{p(e^z-1)} = \frac{1}{p}(1+e^z+\cdots+e^{(p-1)z}) \qquad (z \in \mathbb{C})$$

and we have $Q_k(t) = f^{(k)}(z)|_{z=0}$, in particular $Q_0(t) = f(0) = 1$. Formula (6.7) yields $\log f(z) = t \log g(z)$. Hence $f'(z) = tf(z)\frac{g'(z)}{g(z)}$ and by the product rule of Leibniz we get

$$f^{(k+1)}(z) = t\sum_{\ell=0}^{k}\binom{k}{\ell}f^{(k-\ell)}(z)\left(\frac{g'(z)}{g(z)}\right)^{(\ell)}.$$

Note that $\frac{g'(z)}{g(z)} = c(z)$ with $c(z)$ from (4.3) and by Lemma 4.5 we have

$$\left.\left(\frac{g'(z)}{g(z)}\right)^{(\ell)}\right|_{z=0} = \frac{(-1)^{\ell+1}B_{\ell+1}}{\ell+1}(p^{\ell+1}-1)$$

so that

$$Q_{k+1}(t) = t\sum_{\ell=0}^{k}\binom{k}{\ell}Q_{k-\ell}(t)\frac{(-1)^{\ell+1}B_{\ell+1}}{\ell+1}(p^{\ell+1}-1). \tag{6.8}$$

It follows by induction on k that $Q_k(t)$ are polynomials with respect to t of degree k. We know that $Q_0(t) = 1$ and owing to (6.8) we get $Q_1(t) = tQ_0(t)B_1(p-1) = t(-\frac{1}{2})(p-1)$. Assume that $Q_k(t)$ with fixed $k \geq 1$ is a polynomial of degree k then (6.8) implies that Q_{k+1} is a polynomial of degree $k+1$. Finally, $Q_k(t) = P_k(t)$ for all t since $Q_k(n) = P_k(n)$ for all integer $n \geq 1$ according to (6.5) and (6.6). We get (6.4) from (6.8) if we replace Q by P as well as $k+1$ by k and $\ell+1$ by ℓ. □

Remark 6.3 **1.** A consequence of (6.3) is the following additions theorem

$$P_k(s+t) = \sum_{\ell=0}^{k}\binom{k}{\ell}P_\ell(s)P_{k-\ell}(t). \tag{6.9}$$

2. Formula (6.3) with $n=1$ yields in view of (1.21) the values for $P_k(1)$, namely

$$P_k(1) = \frac{1}{p}\tilde{B}_k(p) \qquad (k = 0,1,2,\ldots). \tag{6.10}$$

For the values $P_k(-1)$ we have

$$P_k(-1) = p^{k+1}\tilde{B}_k\left(\frac{1}{p}\right) \qquad (k = 0,1,2,\ldots) \tag{6.11}$$

which follows from

$$\left(\frac{e^{pz}-1}{p(e^z-1)}\right)^{-1} = p\frac{e^{\frac{1}{p}(pz)}-1}{(e^{pz}-1)} = p\sum_{k=0}^{\infty}\frac{\tilde{B}_k(\frac{1}{p})}{k!}(pz)^k = \sum_{k=0}^{\infty}\frac{p^{k+1}\tilde{B}_k(\frac{1}{p})}{k!}z^k$$

and (6.3) with $n=-1$. We remember that we have by (5.3)

$$P_k(n) = \frac{1}{p^n}S_k(p^n) \qquad (n=0,1,2,\ldots), \tag{6.12}$$

in particular, $P_0(n)=0$ and $P_k(1)=\frac{1}{p}S_k(p)=\frac{1}{p}\tilde{B}_k(p)$, cf. (5.5).

Proposition 6.4 *In case $p=2$ the polynomials $P_k(t)$ satisfy the recursion*

$$P_0(t)=1 \qquad \textit{and} \qquad P_{k+1}(t) = t\left(P_k(t)-\frac{1}{2}P_k(t-1)\right) \quad \textit{for} \quad k\geq 0.$$

Proof: From (6.3) with $p=2$ we get

$$\sum_{k=0}^{\infty}\frac{P_k(t)}{k!}z^k = \left(\frac{e^{2z}-1}{2(e^z-1)}\right)^t = \left(\frac{e^z+1}{2}\right)^t.$$

Note that

$$\sum_{k=0}^{\infty}\frac{P_{k+1}(t)}{k!}z^k = \sum_{k=1}^{\infty}\frac{P_k(t)}{(k-1)!}z^{k-1} = \frac{d}{dz}\left(\sum_{k=0}^{\infty}\frac{P_k(t)}{k!}z^k\right)$$

and

$$\frac{d}{dz}\left(\frac{e^z+1}{2}\right)^t = t\left(\frac{e^z+1}{2}\right)^{t-1}\frac{1}{2}e^z.$$

Further,

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{P_k(t)-\frac{1}{2}P_k(t-1)}{k!}z^k &= \left(\frac{e^z+1}{2}\right)^t - \frac{1}{2}\left(\frac{e^z+1}{2}\right)^{t-1} \\ &= \left(\frac{e^z+1}{2}\right)^{t-1}\frac{1}{2}e^z\end{aligned}$$

Compare of coefficients implies assertion. □

Remark 6.5 By equating coefficients of t^ℓ in relation (6.4) we find a new recursion for the polynomials $a_{k,\ell}(p)$, namely

$$a_{k,\ell}(p) = \sum_{j=1}^{k-\ell+1}\binom{k-1}{j-1}a_{k-j,\ell-1}(p)a_{j,1}(p), \tag{6.13}$$

cf. (5.8).

Now for integer $\ell \geq 0$ we introduce the generating functions of $a_{k,\ell} = a_{k,\ell}(p)$ by

$$G_\ell(z) := \sum_{k=0}^{\infty} \frac{a_{k,\ell}}{k!} z^k. \tag{6.14}$$

Note that $G_0(z) = 1$ since $a_{0,0} = 1$ and $a_{k,0} = 0$ for $k \geq 1$ and that for $\ell \geq 1$

$$G_\ell(z) = \sum_{k=0}^{\infty} \frac{a_{k+1,\ell}}{(k+1)!} z^{k+1} \tag{6.15}$$

since $a_{0,\ell} = 0$, i.e. $G_\ell(0) = 0$. In particular

$$G_1(z) = \sum_{k=1}^{\infty} \frac{(-1)^k B_k}{k \cdot k!} (p^k - 1) z^k \tag{6.16}$$

with the Bernoulli numbers B_k, cf. (5.10).

Proposition 6.6 *For $\ell \geq 1$ we have*

$$G_\ell(z) = \frac{1}{\ell!} G_1(z)^\ell \tag{6.17}$$

with $G_1(z)$ from (6.16).

Proof: Let be $\ell \geq 1$. Relation (6.13) with $k+1$ instead of k and $\ell+1$ instead of ℓ can be written as

$$\begin{aligned} a_{k+1,\ell+1} &= \sum_{j=1}^{k-\ell+1} \binom{k}{j-1} a_{k+1-j,\ell} a_{j,1} \\ &= \sum_{i=0}^{k} \binom{k}{i} a_{k-i,\ell} a_{i+1,1} \end{aligned}$$

with $i = j-1$ where we have used that $a_{k-i,\ell} = 0$ for $i \geq k - \ell + 1$ in view of $a_{m,n} = 0$ for $m < n$. So we have after multiplication with z^k

$$\frac{a_{k+1,\ell+1}}{k!} z^k = \sum_{i=0}^{k} \frac{a_{k-i,\ell}}{(k-i)!} z^{k-i} \frac{a_{i+1,1}}{i!} z^i$$

and summation over k yields in view of the Cauchy product and the relations

$$\sum_{k=0}^{\infty} \frac{a_{k+1,\ell+1}}{k!} z^k = \left(\sum_{k=0}^{\infty} \frac{a_{k+1,\ell+1}}{(k+1)!} z^{k+1} \right)' = G'_{\ell+1}(z)$$

and

$$\sum_{i=0}^{\infty} \frac{a_{i+1,1}}{i!} z^i = \left(\sum_{i=0}^{\infty} \frac{a_{i+1,1}}{(i+1)!} z^{i+1} \right)' = G'_1(z)$$

that

$$G'_{\ell+1}(z) = G_\ell(z)G'_1(z)$$

which is also valid for $\ell = 0$ since $G_0(z) = 1$.

Now we can prove (6.17) by induction on ℓ. Obviously, it is true for $\ell = 1$. Assume that it is true for an integer $\ell \geq 1$. Then we get

$$\begin{aligned} G'_{\ell+1}(z) &= G_\ell(z)G'_1(z) \\ &= \frac{1}{\ell!}G_1(z)^\ell G'_1(z) \\ &= \frac{1}{(\ell+1)!}\frac{d}{dz}G_1(z)^{\ell+1}. \end{aligned}$$

So $G_{\ell+1}(z) = \frac{1}{(\ell+1)!}G_1(z) + c$ where $c = 0$ in view of $G_1(0) = 0$ and $G_{\ell+1}(0) = 0$. □

Proposition 6.7 *The polynomials $a_{k,\ell}(p)$ have the explicit representation*

$$a_{k,\ell}(p) = \frac{(-1)^k k!}{\ell!} \sum_{k_1+\cdots+k_\ell=k} \frac{k!}{k_1!\cdots k_\ell!}\left(\prod_{n=1}^{\ell} \frac{B_{k_n}}{k_n \cdot k_n!}(p^{k_n}-1)\right) \tag{6.18}$$

where $k_1, \ldots, k_\ell$ are positive integers and where B_k are the Bernoulli numbers. Moreover,

$$a_{k,\ell}\left(\frac{1}{p}\right) = \frac{(-1)^\ell}{p^k}a_{k,\ell}(p). \tag{6.19}$$

Proof: By Proposition 6.6 we have in view of (6.14) and (6.16)

$$\sum_{k=0}^{\infty}\frac{a_{k,\ell}(p)}{k!}z^k = \frac{1}{\ell!}\left(\sum_{k=1}^{\infty}\frac{(-1)^k B_k}{k\cdot k!}(p^k-1)z^k\right)^\ell. \tag{6.20}$$

Applying the multimonial theorem (cf. H. Hall [10], Combinatorical theory Wiley (1986)) we get for the right-hand side of (6.20) with positive integers k_i

$$\frac{1}{\ell!}\sum_{k_1+\ldots+k_\ell=k}\frac{k!}{k_1!\cdots k_\ell!}\left(\prod_{n=1}^{\ell}\frac{(-1)^{k_n}B_{k_n}}{k_n\cdot k_n!}(p^{k_n}-1)z^{k_1+\cdots+k_\ell}\right)$$

which in view of $(-1)^{k_1}\cdots(-1)^{k_\ell} = (-1)^k$ if $k_1+\cdots+k_\ell = k$ is equal to

$$\frac{1}{\ell!}\sum_{k=\ell}^{\infty}(-1)^k\sum_{k_1+\cdots+k_\ell=k}\frac{k!}{k_1!\cdots k_\ell!}\left(\prod_{n=1}^{\ell}\frac{B_{k_n}}{k_n\cdot k_n!}(p^{k_n}-1)\right)z^k$$

Now comparing coefficients of z^k in (6.20) yields (6.18).

From (6.18) with $\frac{1}{p}$ instead of p we get in view of

$$\begin{aligned}\sum_{k_1+\cdots+k_\ell=k}\left(\prod_{n=1}^{\ell}\frac{B_{k_n}}{k_n\cdot k_n!}\left(\frac{1}{p^{k_n}}-1\right)\right) &= \frac{1}{p^k}\sum_{k_1+\cdots+k_\ell=k}\left(\prod_{n=1}^{\ell}\frac{B_{k_n}}{k_n\cdot k_n!}(1-p^{k_n})\right)\\ &= \frac{(-1)^\ell}{p^k}\sum_{k_1+\cdots+k_\ell=k}\left(\prod_{n=1}^{\ell}\frac{B_{k_n}}{k_n\cdot k_n!}(p^{k_n}-1)\right)\end{aligned}$$

the relation (6.19). □

Remark 6.8 Let $A_{k,\ell}$ be the main coefficient of the polynomial $a_{k,\ell}(p)$, that means

$$a_{k,\ell}(p) = A_{k,\ell}p^k + o(p^k) \qquad (k\to\infty). \tag{6.21}$$

Then from (6.18) we see that

$$A_{k,\ell} = (-1)^\ell a_{k,\ell}(0)$$

and (5.5) implies in view of (1.21) and (1.18) that

$$\sum_{\ell=0}^{k} A_{k,\ell} = \frac{1}{k+1} \tag{6.22}$$

and that

$$\sum_{\ell=0}^{k} a_{k,\ell}(0) = \frac{B_k}{k+1} \tag{6.23}$$

with the Bernoulli numbers B_k, see Figure 2.

We know already from Proposition 5.5 that for $k\geq 2$ and $1\leq \ell < k$ the polynomials $a_{k,\ell}(p)$ are divisible by $p+1$.

Proposition 6.9 *For integer $\ell \geq 1$ and $r \geq 1$ the polynomial $a_{\ell+2r,\ell}(p)$ is divisible by $(p+1)^2$, see Figure 2.*

Proof: We write short $a_{k,\ell}$ for $a_{k,\ell}(p)$ and use induction on ℓ. The assertion is true for $\ell = 1$ since $a_{2r+1,1}(p) = 0$ according to (5.10) and $B_{2r+1} = 0$. Assume that for a fixed $\ell \geq 1$ the polynomials $a_{\ell+2r,\ell}(p)$ for all $r \geq 1$ are divisible by $(p+1)^2$. By (6.13) with $\ell+1+2r$ instead of k and $\ell+1$ instead of ℓ we get for arbitrary integer $r\geq 1$

$$\begin{aligned}a_{\ell+1+2r,\ell+1}(p) &= \sum_{j=1}^{2r+1}\binom{\ell+2r}{j-1}a_{\ell+1+2r-j,\ell}(p)a_{j,1}(p)\\ &= a_{\ell+2r,\ell}a_{1,1} + \binom{\ell+2r}{1}a_{\ell+2r-1,\ell}a_{2,1} + \cdots + \binom{\ell+2r}{2r}a_{\ell,\ell}a_{2r+1,1}.\end{aligned}$$

By induction assumption the first product $a_{\ell+2r,\ell}a_{1,1}$ is divisible by $(p+1)^2$. The last product $a_{\ell,\ell}a_{2r+1,1} = 0$ since $a_{2r+1,1} = 0$ for $r \geq 1$. Moreover, all another products $a_{\ell+2r-1,\ell}a_{2,1}, \ldots, a_{\ell+1,\ell}a_{2r,1}$ are also divisible by $(p+1)^2$ since each of the both factors (polynomials) is divisible by $p+1$ according to Proposition 5.5. Consequently, $a_{\ell+1+2r,\ell+1}(p)$ is divisible by $(p+1)^2$ and the assertion is proved by induction. □

Remark 6.10 Comparison with recursion (1.7) and formula (1.8) yields $q_k(t) = 2^k P_k(t)$ for the polynomials $q_k(t)$ introduced in [8].

Theorem 6.11 *For $N \in \mathbb{N}$ and $L = \log_p N$ we have*

$$\sum_{k=0}^{\infty} \frac{1}{k!}\frac{1}{N} S_k(N) z^k = \left(\sum_{k=0}^{\infty} \frac{1}{k!} P_k(L) z^k\right)\left(\sum_{k=0}^{\infty} \frac{1}{k!} F_k(L) z^k\right) \qquad (z \in \mathbb{C}) \qquad (6.24)$$

with the polynomials (6.2) and

$$\sum_{k=0}^{\infty} \frac{1}{k!} F_k(L) z^k = \left(\sum_{k=0}^{\infty} \frac{1}{k!}\frac{1}{N} S_k(N) z^k\right)\left(\sum_{k=0}^{\infty} \frac{1}{k!} P_k(-L) z^k\right) \qquad (z \in \mathbb{C}). \qquad (6.25)$$

Proof: For the power sum (4.1) we have by Corollary 4.4 in view of $c_{k,\ell} = a_{k,\ell}(p)$, cf. Proposition 5.3, that

$$\begin{aligned} \frac{1}{N} S_k(N) &= \sum_{\ell=0}^{k} L^\ell \sum_{\kappa=0}^{k-\ell} \binom{k}{\kappa} a_{k-\kappa,\ell} F_\kappa(L) \\ &= \sum_{\kappa=0}^{k} \sum_{\ell=0}^{k-\kappa} \binom{k}{\kappa} a_{k-\kappa,\ell} L^\ell F_\kappa(L) \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{k!}\frac{1}{N} S_k(N) &= \sum_{\kappa=0}^{k} \frac{1}{(k-\kappa)!} \sum_{\ell=0}^{k-\kappa} a_{k-\kappa,\ell}(p) L^\ell \frac{1}{\kappa!} F_\kappa(L) \\ &= \sum_{\kappa=0}^{k} \frac{1}{(k-\kappa)!} P_{k-\kappa}(L) \frac{1}{\kappa!} F_\kappa(L). \end{aligned}$$

In view of the Cauchy product of two power series we get (6.24) with the polynomials (6.2). If we replace t by $-t$ in (6.24) and in (6.2) we see that (6.24) implies (6.25). □

Remark 6.12 **1.** For $N \in \mathbb{N}$ and $L = \log_p N$ we have by Theorem 6.11

$$\frac{1}{N} S_k(N) = \sum_{\ell=0}^{k} \binom{k}{\ell} P_k(L) F_{k-\ell}(L) \qquad (6.26)$$

and

$$F_k(L) = \sum_{\ell=0}^{k} \binom{k}{\ell} P_\ell(-L) \frac{1}{N} S_{k-\ell}(N). \tag{6.27}$$

For the case $p = 2$ one can find in [8] a similarly representation of $S_k(N)$ by means of generating functions.

2. Up to now we only know about the 1-periodic functions $F_k(u)$ with $k \geq 1$ that $F_k(0) = 0$. By means of (6.27) we are able to compute the values $F_k(u)$ for $u = \log_p N$ if $N \leq p$ since for these N the sums $S_{k-\ell}(N)$ are the usual power sums

$$\sum_{n=0}^{N-1} n^{k-\ell} = \tilde{B}_{k-\ell}(N)$$

cf. (1.20). According to (6.27) we get for $u = u_N := \log_p N$ with $N \leq p$ that

$$F_k(u_N) = \sum_{\ell=0}^{k} \binom{k}{\ell} P_\ell(-u_N) \frac{1}{N} \tilde{B}_{k-\ell}(N). \tag{6.28}$$

where $\tilde{B}_n(\cdot)$ are the generalized Bernoulli polynomials, cf. (1.21).

References

[1] **Berg, L.**, and **Krüppel, M.** : *De Rham's singular function, two-scale difference equations and Appell polynomials.* Result. Math. **38**, 18 – 47 (2000)

[2] **Berg, L.**, and **Krüppel, M.** : *Eigenfunctions of two-scale difference equations and Appell polynomials.* Z. Anal. Anw. **20**, 457 – 488 (2001)

[3] **Coquet, J.** : *Power sums of digital sums.* J. Number Theory **22**, 161 – 176 (1986)

[4] **Cooper, C.**, **Kennedy, R. E.**, and **Renberg, M.** : *On certain sums of functions of base b expansions.* The Fibonacci Quaterly **36** No 5, 407 – 415 (1998)

[5] **Delange, H.** : *Sur la fonction sommatoire de la fonction "somme de chiffres".* Enseig. Math. (2) **21**, 31 – 47 (1975)

[6] **Flajolet, F.**, **Grabner, P.**, **Kirschenhofer,P.**, **Prodinger, H.**, and **Tichy, R.F.** : *Mellin transforms and asymptotics: digital sums.* Theoret. Comput. Sci. **123**, 291 – 314 (1994)

[7] **Girgensohn, R.** : *Functional equations and nowhere differentiable functions.* Aequationes Math. **46**, 243 – 256 (1993)

[8] **Girgensohn, R. :** *Digital Sums and Functional Equations.* Integers **12**, No. 1, 141 – 160 (2012)

[9] **Grabner, P. J., Kirschenhofer, P., Prodinger, H.**, and **Tichy, R. F. :** *On the moments of the sum-of-digit function.* Bergum G. E. (ed.) et al., Fibonacci numbers and Applications. Vol. 5: Proceedings of the fifth international conference on Fibonacci numbers and their applications, University of St. Andrews, Scotland 20 – 24, 1992. Dordrecht: Kluwer Akademic Publishers. 263 – 271 (1993)

[10] **Hall, H. :** *Combinatorical theory.* Wiley (1986)

[11] **Kairies, H.-H. :** *Functional equations for peculiar functions.* Aequationes Math. **53**, 207 – 241 (1997)

[12] **Kennedy, R. E.**, and **Cooper, C. :** *Sums of powers of digital sums.* Fibonacci Quaterly **31** No. 4, 341 – 345 (1993)

[13] **Knuth, D. E. :** *Johann Faulhaber and Sums of Powers.* Math. Comp. **61**, No. 203, 277 – 294 (1993)

[14] **Kobayashi, Z., Muramoto, K., Okada, T., Sekiguchi, T.**, and **Shiota, Y. :** *An Explicit Formula of the Newman-Coquet Exponential Sum.* Interdisciplinary Information Sciences, Vol. **13**, No. 1, 1 – 6 (2007)

[15] **Krüppel, M. :** *De Rham's singular function, its partial derivatives with respect to the parameter and binary digital sums.* Rostock. Math. Kolloq. **64**, 69 – 86 (2009)

[16] **Krüppel, M. :** *Functional equations for Knopp functions and digital sums.* Rostock. Math. Kolloq. **65**, 85 – 101 (2010)

[17] **Krüppel, M. :** *The partial derivatives of de Rham's singular function and power sums of binary digital sums.* Rostock. Math. Kolloq. **66**, 45 – 67 (2011)

[18] **Krüppel, M. :** *On the solution of two-scale difference equations.* Rostock. Math. Kolloq. **67**, 59 – 88 (2012)

[19] **Krüppel, M. :** *Two-scale difference equations and power sums related to digital sequences.* Rostock. Math. Kolloq. **68**, 45 – 64 (2013)

[20] **Muramoto, K., Okada, T., Sekiguchi, T.**, and **Shiota, Y. :** *Digital sum problems for the p-adic expansion of natural numbers.* Interdiscip. Inf. Sci., Vol. 6, No. 2, 105 – 109 (2000)

[21] **Muramoto, K., Okada, T., Sekiguchi, T.**, and **Shiota, Y. :** *An explicit formula of subblock occurrences for the p-adic expansion.* Interdisciplinary Information Sciences, Vol. 8, No. 1, 115 – 121 (2002)

[22] **Okada, T., Sekiguchi, T.**, and **Shiota, Y. :** *Applications of binomial measures to power sums of digital sums.* Journal of Number Theory **52**, 256 – 266 (1995)

[23] **Okada, T., Sekiguchi, T. :** *Shiota, Y..* An explicit formula of the exponential sums of digital sumsJapan J. Industr. Appl. Math. **12**, 425 – 438 (1995)

[24] **Stein, A. H. :** *Exponential sums of sum-of-digit functions.* Illinois J. Math. **30**, 660 – 675 (1986)

[25] **Tenenbaum, G. :** *Sur la non-différentiabilité de fonctions périodiques accosiées à certaines formules sommatoires.* Algorithms and combinatorics **13**, 117 – 128, Springer Verlag 1997

[26] **Trollope, J. R. :** *An explicit expression for binary digital sums.* Math. Mag. **41**, 21 – 25 (1968)

[27] **Zeidlin, D. :** *On a general exact formula for sums of powers of digital sums.* Abstracts of the American Mathematical Society **95**, 431 (1994)

received: May 28, 2018

Author:

Manfred Krüppel
Institut für Mathematik,
Universität Rostock,
Ulmenstr. 69,
D-18057 Rostock,
Germany

e-mail: manfred.krueppel@uni-rostock.de

Hints for Authors

Rostock. Math. Kolloq. appears once or twice per year.

Submission

Papers should be sent by e-mail to

romako@uni-rostock.de

or by mail to

Universität Rostock
Institut für Mathematik
Universitätsplatz 1
D-18051 Rostock

We will only consider original contributions which are related to research areas at the University of Rostock. All papers will be reviewed.

AMS-Subject-Classification

Please add one or two AMS-classification numbers which describe the content of your paper.

Manuscript Format

Papers should be written in German or English. Please send your article as a Latex-file and a pdf-file or a ps-file. The Latex-file should not contain self defined commands. Textwidth should be 165 mm and font size 12pt.

Authors Adresses

Please add the current complete adresses of all authors including first name / surname / institution / department / street / house number / postal code / place / country / e-mail-address.

Bibliography

Current numbers within the text ([3], [4]; [7, 8, 9]) refer to the corresponding item in the bibliography at the end of the article, which has the headline References. Please follow the structure of the examples:

[3] **Zariski, O.**, and **Samuel, P.:** *Commutative Algebra.* Princeton 1958

[4] **Steinitz, E.:** *Algebraische Theorie der Körper.* J. Reine Angew. Math. **137**, 167-309 (1920)

[8] **Gnedenko, B.W.:** *Über die Arbeiten von C.F. Gauß zur Wahrscheinlichkeitsrechnung.* In: Reichardt, H. (Ed.): C.F. Gauß, Gedenkband anläßlich des 100. Todestages. S. 193-204, Leipzig 1957

Each citation should be written in the original language. Cyrillic letters must be transliterated as it is usual in libraries.

Hinweise für Autoren

Das Rostock. Math. Kolloq. erscheint ein- bis zweimal pro Jahr.

Einreichen

Senden Sie bitte Ihre Arbeiten per e-mail an

romako@uni-rostock.de

oder per Post an

Universität Rostock
Institut für Mathematik
Universitätsplatz 1
D-18051 Rostock

Wir berücksichtigen nur Originalarbeiten, die in Bezug zu einem der Rostocker Arbeitsgebiete stehen. Alle Arbeiten werden begutachtet.

AMS-Subject-Klassifikation

Bitte geben Sie zur inhaltlichen Einordnung der Arbeit ein bis zwei Klassifizierungsnummern (AMS-Subject-Classification) an.

Manuskript

Manuskripte sollen in Deutsch oder Englisch abgefasst sein. Bitte senden Sie uns Ihre Arbeit als als LaTeX- und als PDF–Datei bzw. PS-Datei. In der LaTeX -Datei sollten selbst definierte Befehle vermieden werden. Die Textbreite sollte 165 mm betragen und die Schriftgröße 12pt.

Adressen der Autoren

Die aktuelle, vollständige Adresse des Autors sollte enthalten: Vornamen Name / Institution / Struktureinheit / Straße Hausnummer / Postleitzahl Ort / Land / e-mail-Adresse.

Literaturzitate

Literaturzitate sind im Text durch laufende Nummern (vgl. [3], [4]; [7, 8, 10]) zu kennzeichnen und am Schluss der Arbeit unter der Zwischenüberschrift **Literatur** zusammenzustellen. Hierbei ist die durch die nachfolgenden Beispiele veranschaulichte Form einzuhalten.

[3] **Zariski, O.**, and **Samuel, P.:** *Commutative Algebra.* Princeton 1958

[4] **Steinitz, E.:** *Algebraische Theorie der Körper.* J. Reine Angew. Math. **137**, 167-309 (1920)

[8] **Gnedenko, B.W.:** *Über die Arbeiten von C.F. Gauß zur Wahrscheinlichkeitsrechnung.* In: Reichardt, H. (Ed.): C.F. Gauß, Gedenkband anläßlich des 100. Todestages. S. 193-204, Leipzig 1957

Die Angaben erfolgen in Originalsprache; bei kyrillischen Buchstaben sollte die (bibliothekarische) Transliteration verwendet werden.