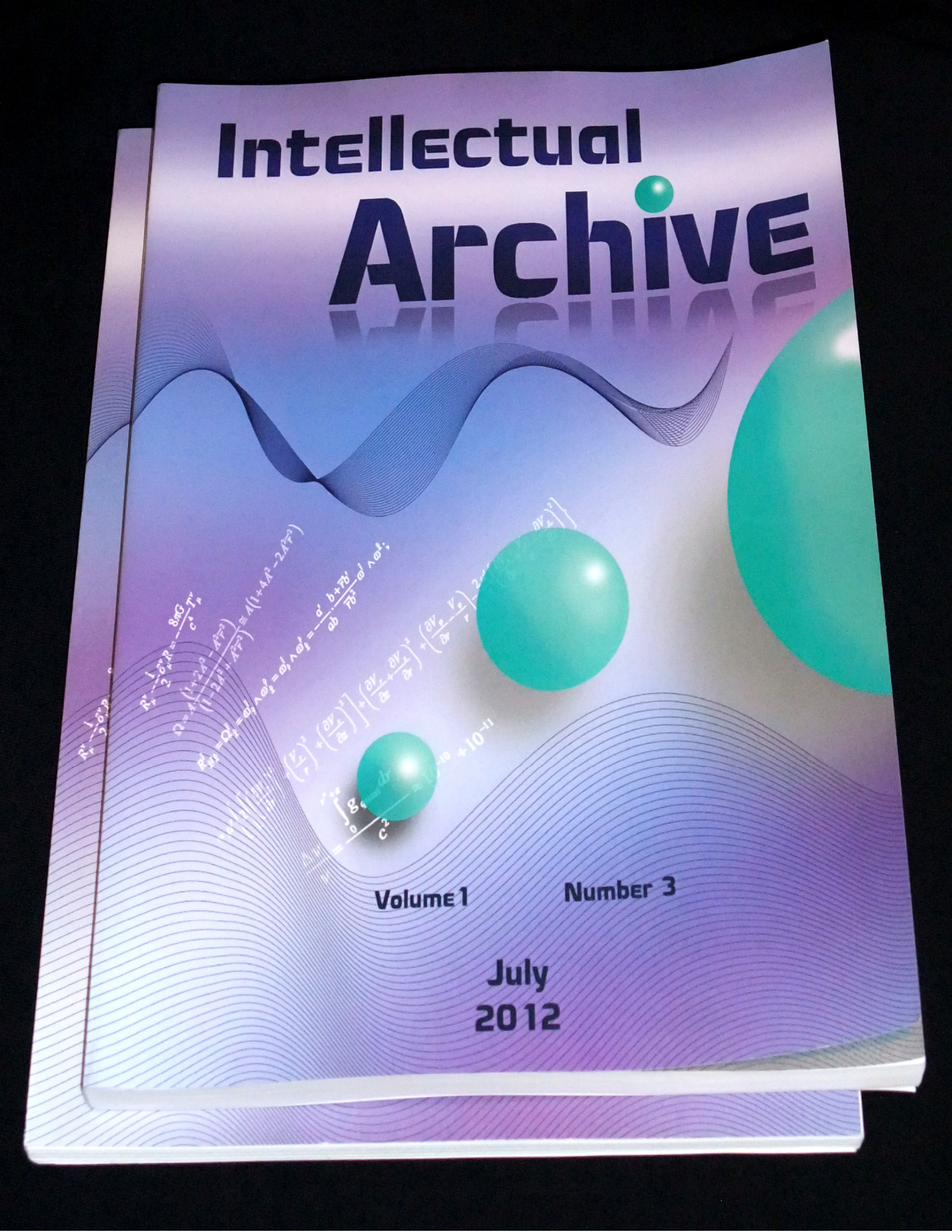
Intellectual Archive



Volume 1 Number 3

July 2012

Intellectual Archive

Volume 1 Number 3

July
2012

IntellectualArchive

Volume 1, Number 3

Publisher : Shiny World Corp.
Address : 9350 Yonge Street
P.O.Box 61533,
Richmond Hill, Ontario
L4C 3N0
Canada

E-mail : support@IntellectualArchive.com
Web Site : www.IntellectualArchive.com
Series : Journal
Frequency : Monthly
Month : July of 2012
ISSN : 1929-4700
Trademark : **IntellectualArchive™**

© 2012 Shiny World Corp. All Rights Reserved. No reproduction allowed without permission.
Copyright and moral rights of all articles belong to the individual authors.

Intellectual Archive

Volume 1 Number 3 July 2012

Table of Contents

Toronto, July 2012

The density maximum of He^4 at the lambda point modeled by the stochastic quantum hydrodynamic analogy

Piero Chiarelli

National Council of Research of Italy, Area of Pisa, 56124 Pisa, Moruzzi 1, Italy

Interdepartmental Center "E.Piaggio" University of Pisa
Phone: +39-050-315-2359
Fax: +39-050-315-2166

Email: pchiare@ifc.cnr.it.

Abstract. The lambda point in liquid He^4 is a well established phenomenon acknowledged as an example of Bose-Einstain condensation. This is generally accepted, but there are serious discrepancies between the theory and experimental results, namely the lower value of the transition temperature T_λ and the negative value of dT_λ/dP. These discrepancies can be explained in term of the quantum stochastic hydrodynamic analogy (QSHA). The QSHA shows that at the $He^4{}_I \rightarrow He^4{}_{II}$ superfluid transition the quantum coherence length λ_c becomes of order of the distance up to which the wave function of a couple of He^4 atoms extends itself. In this case, the $He^4{}_2$ state is quantum and the quantum pseudo-potential brings a repulsive interaction that leads to the negative dT_λ/dP behavior. This fact overcomes the difficulty to explain the phenomenon by introducing a Hamiltonian inter-atomic repulsive potential that would obstacle the gas-liquid transition.

PACS numbers: 67.40.-w; 03.75.Hh; 03.75.Nt; 05.30.5p

Keywords: Lambda point; Liquid He^4; Maximum density, Low temperature critical dynamics; Ballistic to diffusive transition; Anomalous transport

1. Introduction

To explain the $He^4{}_I \rightarrow He^4{}_{II}$ superfluid transition London [1] in 1938 made the hypothesis that He^4 lambda point might be an example of Bose-Einstain (BE) condensation (BEC). This hypothesis was based upon the similarity between the shape of the heat capacity of an ideal boson gas at the BEC transition and the data for the $He^4{}_I \rightarrow He^4{}_{II}$ transition. This convincement was reinforced by the observation that there is no similar phase transition in the Fermi liquid He^3. However, even if the basic BEC hypothesis is acknowledged, looking in details some discrepancies exist [2]. Among those, two are the majors: (1) the calculated BE transition temperature T_B for an ideal gas is 3.14 K while the measured one for the He^4 is of 2.17K. (2) The variation of the transition temperature T_λ with pressure is negative and is opposite in sign to that expected from the BEC. The standard way out is to address the differences to the fact that the BEC theory is applied to an ideal gas while the He^4 is clearly not, since it shows a van der waals-like liquid gas phase transition. Therefore, the inter-molecular potential must be taken into account when we calculate the transition temperature T_B and its variation with temperature.
The BEC theory [3] affirms that below the BE temperature T_B the number of particles , N_e, in the excited state reads

$$N_e = h^{-3} \int \frac{1}{\exp[\varepsilon / kT] - 1} d^3 q d^3 p \qquad (1)$$

and hence

$$N_e(\mathrm{T}) = 2{,}612\ V\frac{2\pi(2m)^{3/2}}{h^3}T^{3/2}. \quad (2)$$

At the BE temperature T_B, it is assumed that all the particles go in the excited state so that

$$T_B = \frac{h^2}{2\pi mk}\left(\frac{N}{2{,}612V}\right)^{3/2}. \quad (3)$$

As equation (3) shows, the increase of pressure, leading to the volume V decrease, will bring to the increase of T_B. This contradicts what is experimentally observed at lambda point where $d\,T_\lambda/dP$ is negative.

By considering the van der Waals state equation

$$\mathrm{P} = \{\mathrm{n\ k\ T}/(\mathrm{V} - \mathrm{n\ b})\} - \mathrm{a\ n}^2/\mathrm{V}^2, \quad (5)$$

(where n is the number of molecules, b is the fourfold atomic volume and a is the mean inter-molecular potential energy derived by using the rigid sphere approximation given by (21-22)) that for punctual particles (i.e., b = 0) with an attractive potential (i.e., a>0) reads

$$\mathrm{P} = \mathrm{n\ k\ T}/\mathrm{V} - \mathrm{a\ n}^2/\mathrm{V}^2 \quad (6)$$

we can see that the pressure decrease $-\mathrm{a\ n}^2/\mathrm{V}^2$ is a consequence of the attractive intermolecular potential. This is equivalent to a compression of the ideal gas and, since the integration in (1) is carried out on the system volume, we can say that a cohesive intermolecular potential reduces the system volume and by (3) that T_B increases.

Therefore, given the ideal gas pressure $\mathrm{P}_{IG} \cong \mathrm{n\ k\ T}/\mathrm{V}$, the variation of BEC temperature ΔT_B has the same sign of the pressure variation ΔP (with respect to the real gas) according to the expression

$$\Delta T_B \propto \Delta\mathrm{P} = \mathrm{P}_{IG} - \mathrm{P} \cong \mathrm{a\ n}^2/\mathrm{V}^2 > 0 \quad (7)$$

Moreover, given $dT_B/dP \sim d\Delta T_B/dP$ it follows that

$$dT_B/dP \propto \mathrm{a\ n}^2\, d(\mathrm{V}^{-2})/dP > 0 \quad (8)$$

since V decreases with the pressure.

Feyman [4] in 1953 and later Butler and Friedman [5,6] calculated in detail the contribution of the inter-molecular potential for a bosonic system showing that it would need a repulsive potential, causing an expansion of the gas, in order to lower T_B as one might expect from (7).

Shortly afterwards, ter Haar [7], pointed out that the repulsive potential was unphysical and would hinder the gas-liquid transition from taking place.

Recently, Deeney et al. [8] showed that a quantum source of energy leading to the expansion of the condensate may explain the negative dT_λ/dP behavior. The QSHA model supports this hypothesis showing that the quantum pseudo potential (QPP) (that acts only in the quantum condensed state) generate a repulsive force leading to the anomalous behavior at lambda point.

The QPP is a well-defined potential energy in the Madelung's quantum hydrodynamic analogy (QHA). It is responsible for the realization of the eigenstates and the consequent quantum dynamics. As shown by Weiner [9], this energy is a real energy of the system and consists in the difference between the quantum energy and the classical one.

If fluctuations are present, the stochastic quantum hydrodynamic analogy (QSHA) shows that the quantum potential may have a finite range of interaction [10] so that dynamics owing a larger scale acquire the classical behavior. On the contrary on a scale shorter than the quantum coherence length λ_c the quantum behavior is restored [10].

Following this approach, when the couples of He^4 molecules lie at a distance smaller or equal to the quantum coherence length λ_c, the atomic dynamics becomes quantum (the related quantum pseudo potential interaction appears) and the systems makes the $He^4{}_I \rightarrow He^4{}_{II}$ transition (see section A.3. in appendix [A]).

In the following the effects of the quantum pseudo potential energy onto the BEC temperature as well as on the sign of dT_λ/dP are derived.

2. The QSHA equation of motion

The QHA-equations are based on the fact that the Schrödinger equation, applied to a wave function $\psi_{(q,t)} = A_{(q,t)} \exp[i\, S_{(q,t)}/\hbar]$, is equivalent to the motion of a fluid with particle density $n_{(q,t)} = A^2_{(q,t)}$ and a velocity $\dot{q}_\alpha(q,t) = m^{-1}\, \partial S_{(q,t)}/\partial q_\alpha$, governed by the equations [11]

$$\partial_t n_{(q,t)} + \nabla_q \bullet (n_{(q,t)} \nabla_q \dot{q}) = 0\,, \tag{9.a}$$

$$\dot{q} = \nabla_p H\,, \tag{9.b}$$

$$\dot{p} = -\nabla_q (H + V_{qu})\,, \tag{9.c}$$

$$H = \frac{p \bullet p}{2m} + V_{(q)}\,. \tag{9.d}$$

By defining $\upsilon^H_j \equiv (\partial H/\partial p_\alpha, -\partial H/\partial q_\beta)$, $\upsilon^{qu}_j \equiv (0, -\partial V^{qu}/\partial q_\beta)$ we can ideally subdivide the phase-space velocity into the Hamiltonian and quantum part to read $\upsilon^Q_j = \upsilon^H_j + \upsilon^{qu}_j$. Moreover, n is the number of structureless particles of the system whose mass is m and V^{qu} is the quantum pseudo-potential that originates the quantum non-local dynamics and reads

$$V_{qu} = -(\frac{\hbar^2}{2m}) n^{-1/2} \nabla_q \bullet \nabla_q n^{1/2}\,. \tag{10}$$

When fluctuations are considered into the hydrodynamic quantum equation of motion, the resulting stochastic QHA dynamics preserve the quantum behavior on a scale shorter than the theory defined quantum coherence length λ_c [10]. Moreover, in the case of non-linear systems, on very large scale the local classical behavior can be achieved when the quantum pseudo potential has a finite range of interaction given by the non-locality length λ_L [10] (with $\lambda_L > \lambda_c$).

Following the procedure given in reference [10], with $n_{(q,t)} = \int_{-\infty}^{+\infty} \rho_{(q,,p,,t)}\, dp_1 ... dp_{3n}$, where $\rho_{(q,,p,,t)}$ is the probability density function (PDF) of the system (whose spatial density $n_{(q,t)}$ represents the squared wave function modulus) , the QSHA equation of motion can be established to read

$$\partial_t n_{(q,t)} = -\nabla_q \bullet (n_{(q,t)} \nabla_q \dot{q}) + \eta_{(q_\alpha, t, \Theta)}\,, \tag{11}$$

$$< \eta_{(q_\alpha, t)}, \eta_{(q_\alpha + \lambda, t+\tau)} >= \underline{\mu} \frac{4m(k\Theta)^2}{\pi^3 \hbar^2} exp[-(\frac{\lambda}{\lambda_c})^2]\, \delta(\tau) \delta_{\alpha\beta} \tag{12}$$

$$\lambda_c = (\frac{\pi}{2})^{3/2} \frac{\hbar}{(2mk\Theta)^{1/2}} \tag{13}$$

where Θ is a measure of the noise amplitude.

Moreover, given that (for the mono-dimensional case) the quantum potential range of interaction λ_L (for $\lambda_L > \lambda_c$) reads [10]

$$\lambda_L = 2\lambda_c \frac{\int_0^{\infty} |\, q^{-1} \frac{dV_{qu}}{dq} \,| dq}{|\frac{dV_{qu}}{dq}|_{(q=\lambda_c)}}, \tag{14}$$

Where the origin (0,0) is the point of minimum Hamiltonian potential energy that is the rest mean position of the particle, for $\lambda_c \cup \lambda_L << \Delta\Omega$ equations (11-13) acquire the classical stochastic form

$$\partial_t \mathrm{n}_{(q,t)} = -\nabla_q \bullet (\mathrm{n}_{(q,t)} \nabla_q \dot{q}) + \eta_{(q_\alpha, t, \Theta)} \tag{15}$$

$$< \eta_{(q_\alpha, t)}, \eta_{(q_\alpha + \lambda, t+\tau)} >= \underline{\mu}\, \delta_{\alpha\beta} \frac{k\Theta}{\lambda_c} \delta(\lambda)\delta(\tau) \tag{16}$$

$$\begin{aligned} \dot{q} = \frac{p}{m} = \nabla_q \lim_{\Delta\Omega/\lambda_L \to \infty} \frac{S}{m} = -\nabla_q \{ \lim_{\Delta\Omega/\lambda_L \to \infty} \frac{1}{m} \int_{t_0}^{t} dt (V_{(q)} + V_{qu(\mathrm{n}_0)} + I^*) \} \\ = -\frac{1}{m} \nabla_q \{ \int_{t_0}^{t} dt (V_{(q)} + I^*) \} = \frac{p_{cl}}{m} + \frac{\Delta p_{st}}{m} \end{aligned} \tag{17}$$

The reader who is interested in more detail about the QSHA picture of a gas phase can refer to the Appendix A.

3. Determination of the quantum potential at the $He^4_I \rightarrow He^4_{II}$ transition

In order to calculate the experimental outputs of the $He^4_I \rightarrow He^4_{II}$ superfluid transition we make use of the well-established statistical method of the Virial expansion that fits very fine for van der Waals fluids. This is possible in the stochastic hydrodynamic analogy since the presence of the quantum pseudo potential brings in the Virial expansion the quantum contribution to the system energy. A central point to derive the thermodynamic quantity by means of the Virial approach is the knowledge of the interaction in the pair of molecules (quantum potential included). Therefore, we firstly calculate the features of the He^4 - He^4 couple interaction.

As shown in ref. [19], the He^4 -He^4 interaction can be satisfying approximated by means of a square well potential of depth $\mathcal{U}^*$ and width 2Δ such as

$V_{LJ(q)} = \infty \quad x < \sigma$ (18)

$V_{LJ(q)} = -\mathcal{U}^* \qquad \sigma < x < \sigma + 2\Delta$ (19)

$V_{LJ(q)} = 0 \qquad x > \sigma + 2\Delta$ (20)

(where $\sigma + \Delta$ is about the mean molecular (half) distance) and by introducing the self states wave functions

$\psi = B \sin[K_n (x - \sigma)] \qquad \sigma < x < \sigma + 2\Delta \qquad E_n > -\mathcal{U}$ (21)

$\psi = B \sin[K_n (2\Delta)]] \exp[-\Gamma_n (x - (\sigma + 2\Delta))] \qquad x > \sigma + 2\Delta \qquad E_n < 0$ (22)

where $\Gamma_n = (-2mE_n/\hbar^2)^{1/2}$, $K_n = (2m(\mathcal{U} + E_n)/\hbar^2)^{1/2}$, into relation (10), the quantum potential reads

$$V^{qu}(n) = -(\hbar^2 / 2m)\,\Gamma_n^2 = E_n \qquad x > \sigma + 2\Delta \tag{23}$$

$$V^{qu}(n) = (\hbar^2 / 2m)\,K_n^2 = (\mathcal{U} + E_n) \qquad \sigma < x < \sigma + 2\Delta \tag{24}$$

Where the values E_n are given by the trigonometric equation

$$\tan [K_n (2\Delta)] = -K_n / \Gamma_n = -(-(\mathcal{U} + E_n) / E_n)^{1/2} \qquad E_n < 0 \tag{25}$$

and hence

$$\Delta = (\hbar^2 / 8km)^{1/2} \arctan [-(-(\mathcal{U} + E_0) / E_0)^{1/2}] / (\mathcal{U}/k + E_0/k)^{1/2} =$$
$$= 1.231 \times 10^{-10} \{\arctan [-(-(\mathcal{U} + E_0) / E_0)^{1/2}]\} / (\mathcal{U}/k + E_0/k)^{1/2} \tag{26}$$

Moreover, by assuming that the mean square well deepness $\mathcal{U}^*$ is slightly smaller than the L-J potential one $\mathcal{U}$ ($\mathcal{U}/k \cong$ 10.9 °k) and by evaluating that the value of the energy E_0 of the fundamental state at the transition is about

$$-E_0/k \sim T_{cr} = 5.19\ ^\circ K \tag{27}$$

we obtain

$$\Delta \gtrsim 1.20 \times 10^{-10}\ m = 2.3\ \text{Bohr}. \tag{28}$$

if we choose $\mathcal{U}^*$ to obtain the value for "a" given by (22) it follows that

$$\mathcal{U}^*/k \cong [V_{cr} / N_A((\sigma + 2\Delta)^3 - (\sigma)^3)]\ \mathcal{U}/k = 0.82\ \mathcal{U}/k = 8.9\ ^\circ K,$$

from where, it follows that

$$\Delta \sim 1.54 \times 10^{-10}\ m = 2.9\ \text{Bohr} \tag{29}$$

and that the mean He^4_2 atomic distance

$$\sigma + \Delta \cong 3.82 \times 10^{-10}\ m = 7.2\ \text{Bohr} \tag{30}$$

that well agrees with the values 7.1 Bohr given in ref. [16]. Moreover, as shown in Appendix [B] the above results well agree with the Virial expansion applied to the He^4.
On this data, we can check that the coherence length of the deterministic quantum state λ_c is coherently of order of the intermolecular distance at the $He^4_I \rightarrow He^4_{II}$ superfluid transition.
Reaching the lambda point (let's suppose by He^4 - He^4 cooling), the mean half atomic distance decreases to the value $\sigma + \Delta$ of the fundamental state and the wave function variance decreases to 2Δ (the He^4 atoms lie almost inside the potential well). Therefore, assuming that the quantum coherence length λ_c becomes much bigger that to the dimension of space domain where the wave function is relevant (i.e., the well width of 2Δ) to read

$$\lambda_c = \left(\frac{\pi}{2}\right)^{3/2} \frac{\hbar}{(2mk\Theta)^{1/2}} > 2\Delta, \tag{31}$$

for the couple of He^4 - He^4 atoms, at lambda point it follows that

$$\Theta_\lambda < 0.59 \times 10^{-19}\ \Delta^{-2}\ ^\circ K \tag{32}$$

and hence, by (28) that

$\Theta_\lambda < 4{,}0$ °K
or, more precisely, by using (29) that

$$\Theta_\lambda < 2{,}49 \text{ °K} \qquad (34)$$

Even if Θ is not exactly the thermodynamic temperature T, the result (47) is very satisfying since it correctly gives the order of magnitude of the transition temperature of the lambda point. The fact that Θ is close to T can be intuitively understood with the fact that going toward the absolute null temperature, correspondingly, Θ must decrease since the systems fluctuations must vanish in both cases.
As shown in [10]a relation between Θ and T can be established for an ideal gas at equilibrium. In this case, the thermodynamic temperature T converges to the vacuum fluctuation amplitude Θ in going toward the to absolute zero. In the case of a real gas and its fluid phase, a bit of difference between Θ and T may exists for $\Theta \neq 0$.
The result (34) definitely says that below a temperature of about 2,5°K degrees Kelvin the quantum potential enters more and more in the $He^4{}_I$–$He^4{}_I$ pair interaction. As it is shown in the following section, this well agrees with the features of the He lambda point that clearly shows how the increase of He^4 density (the sign of the quantum potential interaction) starts before the transition $He^4{}_I \rightarrow He^4{}_{II}$ takes place.

3.2. The sign of $\Delta T_\lambda = T_\lambda - T_B$ and of dT_λ/dP at He^4 lambda point

The above equation (22) holds for normal fluid phases at a temperature above the superfluid transition one. Below the superfluid transition temperature, as shown by (31 and 47) the quantum coherence length λ_c becomes larger than the inter-atomic He^4 - He^4 distance and hence the quantum potential contributes to the molecular energy and it must be taken into account in the calculation of the mean inter-molecular potential energy "a" that reads

$$a \equiv -2\pi \int_{r_0}^{\infty} (V_{LJ(r)} + V^{qu})\, r^2\, dr = -2\pi \{ \int_{r_0}^{\infty} V_{LJ(r)}\, r^2\, dr + \int_{r_0}^{\infty} V^{qu}\, r^2\, dr \} = a^{cl} + a^{qu} \qquad (35)$$

where

$$a^{qu} = -2\pi \int_{r_0}^{\infty} V^{qu}\, r^2\, dr = -2\pi\, (\mathcal{U} + E_0) \int_{r_0}^{\sigma+\lambda_{\mathcal{N}}} r^2\, dr = -\tfrac{2}{3}\pi\, (\mathcal{U} + E_0)\, [(\sigma + \lambda_{\mathcal{N}})^3 - 2^{1/2}(\sigma)^3] < 0. \qquad (36)$$

From (36) we can observe that a^{qu} is negative since from (24) $V^{qu} = (\mathcal{U} + E_n)$ is positive. Therefore, below the superfluid transition temperature, the state equation (20) reads:

$$\{P + a^{cl}\, n^2/V^2 + a^{qu}\, n^2/V^2\}\, (V - n\, b) = \{P + a^{cl}\, n^2/V^2 + \Delta P^{qu}\}\, (V - n\, b) = n\, k\, T \qquad (37)$$

where $\Delta P^{qu} = a^{qu}\, n^2 / V^2$, so that the pressure for $He^4{}_I$ and $He^4{}_{II}$ respectively reads

$$P_{I\,(He^4{}_I)} \cong \{n\, k\, T/(V - n\, b)\} - a^{cl}\, n^2/V^2 \qquad (38)$$

$$P_{II(\,He^4{}_{II})} = \{n\, k\, T/(V - n\, b)\} - a^{cl}\, n^2/V^2 - \Delta P^{qu} \qquad (39)$$

Where it is posed $V \cong V_I \cong V_{II}$ since the fluid phase is poorly compressible. By using the same criterion of (7) the variation $\Delta T_\lambda = T_\lambda - T_{B\,(He^4_I)}$ has the same sign of the pressure difference to read

$$\Delta P = (P_I - P_{II}) = \Delta P^{qu} = a^{qu}\, n^2 / V^2 < 0 \qquad (40)$$

and the sign of dT_λ/dP is the same of the derivative

$$d\Delta P^{qu}/dP = a^{qu}\, n^2\, d\,(V^{-2})/dP < 0. \qquad (41)$$

given that $d(V^{-2})/dP$ is positive.
Therefore, the quantum potential of the QSHA leads to both ΔT_λ and dT_λ/dP negative.
Finally, in order to show that the result obtained above is a direct consequence of the convex harmonic quantum potential, we use its more precise expression given in appendix C, where the interaction of a couple of He^4 - He^4 atoms is approximated by a harmonic well (as given by the atomic Lennard-Jones potential) and coherently found to be

$$V^{qu}_{(q,t)} = -(\hbar^2/2m)|\psi|^{-1}\partial^2|\psi|/\partial q\partial q = -(2\hbar^2/m)\, K_0^4\,(q - \underline{q})^2 + (\hbar^2/m)\, K_0^2, \qquad (42)$$

where

where $K_0 = (2m(\mathcal{U} + E_0)/\hbar^2)^{1/2}$ and $\underline{q}$ is the mean He^4 - He^4 inter-atomic distance.

4. Discussion

The negative sign of both $\Delta T_\lambda = T_\lambda - T_B$ and of dT_λ/dP are the direct consequence of the convex harmonic quantum potential that leads to a repulsive inter-atomic force so that the pressure of the superfluid He^4_{II} is higher of that one it would assume the He^4_I at the same temperature. Due to the repellent quantum potential energy, the passage from the He^4_{II} state to the equivalent He^4_I one (submitted to a lower pressure) needs less kinetic energy to happen and hence T_λ is smaller than the condensation temperature $T_{B\,(He^4_I)}$. Moreover, since in He^4_{II} a higher pressure than in He^4_I is needed to maintain the same atomic distance, when the temperature is lowered at constant pressure near the lambda point (crossing T_λ) a decrease in density is produced as we get closer to the transition $He^4_I \rightarrow He^4_{II}$. Therefore, during the cooling process the He^4 shows a maximum in its density just above the $He^4_I \rightarrow He^4_{II}$ transition as confirmed by the experimental outputs.
It must be noted that for the realization of the maximum density, the crossover between the rate of change of the He^4_I thermal shrinking and that one of the He^4_{II} quantum dilatation is needed.
Moreover, since the density maximum is at 2.2 °K while T_λ =2.17 °K, we can infer that the quantum interaction starts little bit before the transition temperature (i.e., Θ_λ < 2,49 °K) as (37) well signals (a larger and large fraction of He^4 atoms fall in the quantum interaction closer and closer we get to T_λ).
Moreover, it is worth mentioning that the QSHA model does not exclude the possibility of similar maximum density phenomena close to liquid-solid transitions (such as that of water) since, in this case, the quantum interaction between the atoms in a crystal is also set by the quantum potential whose interaction range λ_L becomes larger than the typical inter-atomic distances (see appendix [C]). This fact well agrees with the similarity between the $He^4_I \rightarrow He^4_{II}$ and the water-ice transitions widely accepted by the scientific community in the field. Also in this case, in order to have the maximum density at the liquid solid transition, the quantum dilation must overcomes the thermal shrinking velocity.

5. Conclusion

The finite range the quantum interaction in the QSHA is able to explain the controversial aspect of negative dT_λ/dP at the He^4 lambda point without the introduction of a non-physical repulsive atomic potential that would hinder the gas-liquid phase transition [20,21]. The quantum pseudo-potential of the QSHA model is exactly the required potential: it is repulsive as widely requested by the scientific community to explain the maximum density of He^4 lambda point, but it also has the property to disappear in the classical phase and to not hinder the liquid-gas phase transition as any Hamiltonian potential would do.

The pseudo-potential of the QSHA approach also explains both why the lambda transition temperature T_λ is smaller than the BE one T_B and why the liquid He^4 has a maximum in its density just above the lambda point in agreement with the experimental measurements. The model puts in evidence that the perfect BE condensation is a phenomenon that happens between an ideal gas and its condensed quantum phase. As far as it concerns the liquid He^4, the phenomenon is slightly different being, by the fact, a transition between a real gas (in the fluid phase) and its quantum condensed phase so that the transition temperature is smaller.

Finally, it must be noted that even if the $He^4_I \rightarrow He^4_{II}$ is very well described by Montecarlo numerical simulation of standard quantum equations, the SQHA gives a modeling explanation that leads to a figurative comprehension of such a phenomenon that is complemental to that one coming by the numerical methods.

The SQHA kinetic equation is not a semi-empirical kinetic equation as those used to study the system behavior and its universality class near the phase transitions [23] but is a microscopic theoretical model from which exact Langevin kinetic equation can be obtained by the standard procedure of coarse-graining [24].

Appendix A

The QSHA model for gas and condensed phases

A.1. Analysis of the quantum potential of localized free particles

In order to elucidate the particle PDF evolution, in a classical phase (i.e., the mean inter-particle distance bigger than λ_L) we inspect the interplay between the Hamiltonian potential and the quantum potential that define the quantum non-locality length.

Fixed the PDF at the initial time, then the Hamiltonian potential and the quantum one determine the evolution of the PDF that on its turn modifies the quantum potential.

A Gaussian PDF has a parabolic repulsive quantum potential, if the Hamiltonian potential is parabolic too (the free case is included), when the PDF wideness adjusts itself to produce a quantum potential that exactly compensate the force of the Hamiltonian one, the Gaussian states becomes stationary (eigenstates). In the free case, the stationary state is the flat Gaussian (with an infinite variance) so that any Gaussian PDF expands itself following the ballistic dynamics of quantum mechanics [see Unpublished note 1] [12].

From the general point of view, we can say that if the Hamiltonian potential grows faster than a harmonic one, the wave equation of a self-state is more localized than a Gaussian one (its PDF decreases faster), and by (10) this leads to a stronger-than -linear quantum potential (also at large distance).

On the contrary, a Hamiltonian potential that grows slower than a harmonic one will produce a less localized (stationary) PDF that decreases slower than the Gaussian one [see Appendix D], so that the quantum potential grows less-than-linearly and may lead to a finite quantum non-locality length by (15).

As shown in ref. [10], the large distances exponential-decay of the PDF such $\lim_{|q| \rightarrow \infty} |n^{1/2}| \approx \exp[-P^h(q)]$ with $h < 3/2$ is a sufficient condition to have a finite quantum non-locality length.

Thence, we can enucleate three typologies of quantum potential interactions:

(1) $h > 2$ strong quantum potential that leads to quantum force $\partial V^{qu}(q)/\partial q$ that grows faster than linearly and λ_L is infinite (*super-ballistic* free particle PDF expansion)

$$\lim_{q\rightarrow\infty} |\partial V^{qu}(q)/\partial q| > q^{1+\varepsilon}. \qquad (\varepsilon > 0) \qquad (A.1)$$

(2) $h = 2$ strong quantum potential that leads to quantum force $\partial V^{qu}(q)/\partial q$ that grows linearly and λ_L is infinite (*ballistic* free PDF enlargement)

$\lim_{q\to\infty} |\partial V^{qu}(q)/\partial q| \propto q^{1}$. (A.2)

(3) $2 > h \geq 3/2$ middle quantum potential; the integrand of (15) ax well as the quantum force may be vanishing at large distance to read

$\text{Const} \geq \lim_{q\to\infty} |q^{-1}\, \partial V^{qu}(q)/\partial q| > q^{-1}$. (A.3)

but λ_L may be still infinite (*under-ballistic* free PDF expansion).

(4) $h < 3/2$ week quantum potential leading to quantum force that becomes vanishing at large distance following the asymptotic behavior

$\lim_{q\to\infty} |q^{-1}\, \partial V^{qu}(q)/\partial q| \approx q^{-(1+\varepsilon)}$, $(\varepsilon > 0)$ (A.4)

with a finite λ_L for $\Theta \neq 0$ (*asymptotically vanishing* free PDF expansion).

A.2. Free pseudo-Gaussian particles of a gas phase in presence of noise

Gaussian particles generate a quantum potential that has an infinite range of interaction and hence they do not admit macroscopic local dynamics.

Nevertheless, imperceptible deviation by the perfect Gaussian PDF may possibly lead to finite quantum non-locality length [see Appendix D]. Particles that are inappreciably less localized than the Gaussian ones (let's name them as pseudo-Gaussian) own $\partial V^{qu}(q)/\partial q$ that can sensibly deviate by the linearity so that the quantum non-locality length may be finite.

In the case of a free pseudo-Gaussian particle we can say that λ_L extends itself at least up to the Gaussian core (where the quantum force is linear). At a distance much bigger than λ_L for $h < 3/2$, the expansive quantum force becomes vanishing.

On short distance, for $q << \lambda_c$, the noise is progressively suppresses (i.e., the deterministic quantum dynamics is established). Therefore, it follows that:

(1) For $q << \lambda_c < \lambda_L$, the evolution is quantum ballistic.

(2) For $q >> \lambda_L > \lambda_c$ the evolution is classically stochastic.

(3) For $\langle \Delta q^2 \rangle^{1/2} << \lambda_c < \lambda_L$, the quantum deterministic state with $h = 2$ is approached by the free pseudo-Gaussian particle.

(4) For $\langle \Delta q^2 \rangle^{1/2} >> \lambda_L > \lambda_c$ and for $h < 3/2$ the expansion dynamics of the free pseudo-Gaussian PDF are almost diffusive.

If at the initial time, we have the pseudo Gaussian PDF confined on a micro-scale (i.e., $\langle \Delta q^2 \rangle^{1/2} << \lambda_c < \lambda_L$), the expansion of the particle PDF is always very fast (ballistic).

For $\lambda_c << \langle \Delta q^2 \rangle^{1/2} < \lambda_L$ the noise will add diffusion to the PDF ballistic enlargement.

This until $\langle \Delta q^2 \rangle^{1/2}$ reaches the length of λ_L. Then, when $\langle \Delta q^2 \rangle^{1/2} >> \lambda_L$, the expansion dynamics slow down toward the diffusive one.

When the (pseudo-Gaussian) PDF has reached the mesoscopic scale ($\langle \Delta q^2 \rangle^{1/2} \sim \lambda_L$), we can infer that its core expands ballistically while its tail diffusively.

Since the outermost expansion is slower than the innermost, there is an accumulation of PDF (ρ is a conserved quantity) in the middle region ($q \sim \lambda_L$) generating, as time passes, a slower and slower (than the Gaussian one) PDF decrease so that (for a free particle) the quantum potential and λ_L decrease (and cannot increase) in time.

In force of these arguments (i.e., the core quantum ballistic enlargement is faster than the diffusive outermost classical one), the free pseudo-Gaussian states (with $h < 3/2$) are self-sustained and remain pseudo-Gaussian in time.

As far as it concerns the particle de-localization at very large times, the asymptotically vanishing quantum potential does not completely avoid such a problem since the (Θ-noise driven) diffusion spreading of the molecular PDF remains (even if it is much slower than the quantum ballistic one).

If the particle PDF confinement cannot be achieved in the case of one or few molecules, on the contrary, in the case of a system of a huge number of structureless particles (with a repulsion core as in the case of the Lennard-Jones (L-J) potentials) the PDF localization can come from the interaction (collisions) between the molecules.

More analytically, we can say that in a rarefied gas phase at the collision, when two particles get at the distance of order of the L-J potential minimum r_0, the quantum non-locality length becomes sensibly large and bigger than the inter-particle distance of interaction r_0 (since for a sufficiently deep L-J well the potential is approximately quadratic and the associated state largely Gaussian) [see Appendix C].
After the collision, when the molecules are practically free, the PDF starts to expand again and λ_L decreases again. It will never reach the flat Gaussian configuration since, in a finite time, the molecule undergoes another collision taking a bit of PDF squeezing leading to a new increase of the $\lambda_L / <\Delta q^2>^{1/2}$ ratio. The overall effect of this process is that the random collisions among the huge number of free particles in a gas phase with L-J type intermolecular potential, maintain their localization. Finally we observe that more rarefied is the gas phase closer is λ_L to its minimum value λ_c . Thence, except for solids where λ_L is sensibly higher than λ_c [see Appendix C] we assume for classic liquid and gas phases $\lambda_L \approx \lambda_c$.

A.3. Condensed phase and $He^4_I \rightarrow He^4_{II}$ transition

When both the lengths λ_c and λ_L are much smaller than the smallest physical length of the system (so that the resolution of the descriptive scale can be of order or bigger than λ_L) the macroscopic classical description arises. This for instance happens in a rarefied gas phase of L-J interacting particles where λ_L as well as λ_c are very small compared to the intermolecular mean distance (except for few colliding molecules).
On the contrary, when the mean inter-particle distance becomes comparable with the quantum non-locality length, the classical description may break down because the quantum potential enters in the particle interaction. Furthermore, if the wave function of the interacting particles is localized on a length of order or smaller than the quantum coherence length λ_c , the quantum deterministic description takes place for the bounded states of the couples of molecules.
In the classical regime, the Virial expansion furnishes an elegant conceptual understanding for passing from a gas to a condensed phase for molecules having finite range of interaction *even in non-equilibrium condition* [21].
In the classical treatment of the Virial expansion, the energy function does not include the quantum potential and hence converges to the classical value failing, for instance, to predict the law of the specific heat for solids where the quantum dynamics enters in the atoms interaction.
In the frame of the QHA description, the quantum potential energy (that changes at each stationary state) added to the classical value of the energy, leads to the variety of the quantum energy eigenvalues. This is very clearly shown in Ref. [9], the energy of the quantum eigenstates is composed by the sum of the two terms: one steams from the classical Hamiltonian while the other one by the quantum potential, leading to the correct eigenvalue E_n. Therefore, in principle the Virial approach can be applied (in the QSHA model) both for quantum as well classical molecular interactions
Since in a crystal the atoms fall in the linear range of interaction, the quantum non-locality λ_L is larger than the inter-molecular distance [see Appendix C] and the system shows quantum characteristics (in those properties depending by the molecular state).
Usually, for crystalline solids the inter-atomic distance lies in the harmonic range of the L-J interaction even at temperature higher than the room one due to the great deepness of the potential well [see Appendix C].
When, at higher thermal oscillations, the mean molecular distance starts to increases by the equilibrium position r_0 toward the non-linear range of the L-J inter-molecular potential, we have a transition from the solid phase to the liquid one [22]. During this process, the inter-particle wave function extends itself more and more in the non-linear L-J zone so that the quantum potential weakens and λ_L decreases [see Appendix C].
For deep L-J intermolecular potential well, this happens at high temperature and we have a direct transition from the solid to the classical fluid phase.
For small potential well, the liquid phase can persist down to a very low temperature. In this case, even if λ_L may result smaller than the inter-particle distance (so that the liquid phase is maintained), decreasing the temperature, and hence the amplitude Θ of fluctuations, when λ_c grows and becomes of order of the mean molecular distance, the liquid phase may acquire quantum properties (about those depending by the molecular interaction such as the viscosity). The fluid-superfluid transition can happen if the temperature of the fluid can be lowered up to the transition point before the solid phase takes place (i.e., $\lambda_L < r_0$).
Therefore, it worth noting that the mechanism that brings to the quantum inter-atomic interaction in a solid is different from that one in a superfluid: in the former the linearity of the interaction leads to a quantum non-

locality length λ_L larger than the typical atomic distance while in the latter is the decrease of Θ, by lowering the temperature, that increases λ_c up to the mean inter-atomic length.
Even if the relation between the PDF noise fluctuations amplitude Θ and the temperature T of an ensemble of particles is not straight [10], it can be easily acknowledged that when we cool a system toward the absolute zero (with steps of equilibrium) also the noise amplitude Θ reduces to zero since the energy fluctuations of the system must vanish. Thence, even there is not a fix linear relation between the fluctuation amplitude Θ and the temperature we expect lower values of Θ for lower values of the temperature [10].

Appendix B

The Virial expansion applied to the He fluid

As far as it concerns the first point, we have that from the standard Virial expansion [13] the state equation of (classical) real gas accounting only for double collisions, reads:

$$P\,V = n\,k\,T\,\{1 - (n\,(a/kT - b)/\,V)\}, \tag{18}$$

that under the standard substitution [14]

$$\{1 + n\,b\,/\,V\} \cong \{1 - n\,b\,/\,V\}^{-1} \tag{19}$$

leads to the van der Waals equation

$$\{P + a\,n^2\,/V^2\}\,(V - n\,b) = n\,k\,T \tag{20}$$

where P is the pressure, V the volume, n the number of molecules,

$$b = \tfrac{2}{3}\,\pi\,r_0^3 = V_{cr}\,/\,3N_A \tag{21}$$

is the fourfold atomic volume [15] and

$$a = -2\,\pi \int_{r_0}^{\infty} V_{(r)}\,r^2\,dr \cong 4\pi\,r_0^3\,\mathcal{U}\,/3 = 2\mathcal{U}\,V_{cr}\,/\,3N_A \tag{22}$$

is the mean inter-molecular potential energy derived by using the rigid sphere approximation [13] that reads

$$V_{(r)} = \infty, \qquad x < r_0 \tag{23}$$

$$V_{(r)} = V_{LJ(q)} = 4\,\mathcal{U}\,[(\sigma/q)^{12} - (\sigma/q)^6]\,, \qquad x > r_0 \tag{24}$$

where $\mathcal{U} = -V_{LJ(r_0)}$ is the well depth of the L-J intermolecular potential. Moreover, by using the relation[15]

$$a = 9\,k\,T_{cr}V_{cr}\,/\,8\,N_A, \tag{25}$$

from (22), for He^4 $\mathcal{U}$/k reads

$$\mathcal{U}\,/\,k = 27\,T_{cr}\,/\,16 = 8.77°K, \tag{26}$$

satisfactory close to the value $\mathcal{U}\,/\,k = 10.9$ given by quantum Monte Carlo models [16] and to the value $\mathcal{U}\,/\,k = 11.07$ given by He^4 -He^4 scattering [17].
Moreover, by using for helium [18] the value of

$$V_{cr} = 5.7 \times 10^{-5}\ m^3/moles, \tag{27}$$

it follows that

$r_0 \cong 2.56 \times 10^{-10}\,m = 4.8$ bohr (28)

where

$r_0 = 2^{1/6}\sigma$ (29)

is the point of minimum for the L-J intermolecular potential with

$\sigma = 2^{-1/6} r_0 = 2.32 \times 10^{-10}\,m \cong 4.35$ Bohr.

Appendix C

Quantum non-locality length of L-J bounded states

In order to calculate the quantum potential and its non-locality length for a L-J potential well, we can assume the harmonic approximation

$V_{LJ(q)} = \frac{1}{2}\, k\, (q - r_0)^2 + C,$ (C.1)

where

$k = 4\, K_0^4\, \hbar^2/m$ (C.2)

where $K_0 = (2m(\mathcal{U} + E_0)/\hbar^2)^{1/2}$, and where the constant C can be calculated by the energy eigenvalue of the fundamental state

$C = E_0 - V^{qu}{}_{0\,(q-\underline{q}=0)}$, (C.3)

leading to a Gaussian wave function whose series expansion at second order coincides with that one of eqs. (21-22) having the same eigenvalue E_0 and mean position $\underline{q} = \sigma + \Delta$ that reads

$\psi_0 = B \exp[-K_0^2(q-\underline{q})^2] \cong B\,[1 - K_0^2(q-\underline{q})^2] \cong B \sin[K_0(q-\sigma)] \qquad |q - r_0| << 2\Delta/\pi.$ (C.4)

The convex quadratic quantum potential associated to the wave function ψ_0 reads

$V^{qu}{}_{(q,t)} = -(\hbar^2/2m)|\psi|^{-1}\partial^2|\psi|/\partial q \partial q = -(2\hbar^2/m)\, K_0^4\, (q - \underline{q})^2 + (\hbar^2/m)\, K_0^2$ (C.5)

that leads to the quantum force

$-\partial V^{qu}/\partial q = 2\, K_0^4\, (\hbar^2/m)\, (q - \underline{q})$ (C.6)

and to

$C = E_0 - V^{qu}{}_{0\,(q=0)} = \frac{1}{2}\hbar\,(k/m)^{1/2} - (\hbar^2/m)\, K_0^2 = 0.$

Given the simple exponential PDF decrease of (21-22) for $x > \sigma + 2\Delta$ (that leads to a vanishing quantum potential as well as to vanishing small quantum force), we can disregard the contribution to the quantum non-locality length for $x > \sigma + 2\Delta$.
Thence, by (13) it follows that,

$$\lambda_L \cong 2\lambda_c \frac{\int_{\sigma+\Delta}^{\sigma+2\Delta} |\, q^{-1} \frac{dV_{qu}}{dq}\,| dq}{|\, \frac{dV_{qu}}{dq}\,|_{(q=\lambda_c)}} \qquad \text{(C.7)}$$

$|d\, V^{qu}/dq| = 2\, K_0^4\, (\hbar^2/m)\, \lambda_c$

$|-q^{-1}\partial\, V^{qu}/\partial q| = |2\, K_0^4\, (\hbar^2/m)|$

$$\lambda_L \cong 2 \int_{\sigma+\Delta}^{\sigma+2\Delta} dq = 2\Delta = \lambda_c$$

Appendix D

Pseudo-Gaussian PDF

If a system admits the large-scale classical dynamics, the PDF cannot acquire an exact Gaussian shape because it would bring to an infinite quantum non-locality length.
In section (III.B.1) we have shown that for $h < 3/2$ (when the PDF decreases slower than a Gaussian) a finite quantum length is possible.
The Gaussian shape is a physically good description of particle localization but irrelevant deviations from it, at large distance, are decisive to determine the quantum non-locality length.
For instance, let's consider the pseudo-Gaussian function type

$$n_{(q,t)} = \exp[-(q-\underline{q})^2/\langle\Delta q^2\rangle[1 + [(q-\underline{q})^2/\Lambda^2 f(q-\underline{q})]]], \qquad \text{(D.1)}$$

where $f(q-\underline{q})$ is an opportune regular function obeying to the conditions

$\Lambda^2 f(0) >> \langle\Delta q^2\rangle$ and $\lim_{|q-\underline{q}|\to\infty} f(q-\underline{q}) << (q-\underline{q})^2/\Lambda^2$.

For small distance $(q-\underline{q})^2 << \Lambda^2 f(0)$ the above PDF is physically indistinguishable from a Gaussian, while for large distance we obtain the behavior

$$\lim_{(q-\underline{q})\to\infty} n_{(q,t)} = \exp[-\Lambda^2 f(q-\underline{q})/\langle\Delta q^2\rangle]. \qquad \text{(B.3)}$$

For instance, we may consider the following examples

i. $f(q-\underline{q}) = 1$ $\quad \lim_{|q-q_0|\to\infty} n_{(q,t)} = \exp[-\Lambda^2/\langle\Delta q^2\rangle]$; (D.4)

ii. $f(q-\underline{q}) = 1+|q-\underline{q}|$ $\quad \lim_{|q-q_0|\to\infty} n_{(q,t)} \propto \exp[-\Lambda^2|q-\underline{q}|/\langle\Delta q^2\rangle]$; (D.5)

iii. $f(q-\underline{q}) = 1 + \ln[1+|q-\underline{q}|^h]$ $\quad (0< h <2)$ $\quad \lim_{|q-q_0|\to\infty} n_{(q,t)} \propto |q-\underline{q}|^{-h\Lambda^2/\langle\Delta q^2\rangle}$; (D.6)

iv. $f(q-\underline{q}) = 1 + |q-\underline{q}|^h$ $\quad (0< h <2)$ $\quad \lim_{|q-q_0|\to\infty} n_{(q,t)} \propto \exp[-\Lambda^2|q-\underline{q}|^h/\langle\Delta q^2\rangle]$. (D.7)

All cases (i-iv) lead to a finite quantum non-locality length λ_L.
In the case (iv)(D.1) reads

$$n_{pg(q,t)} = \exp[-((q-\underline{q})^2/\langle\Delta q^2\rangle\{1 + \Lambda^{-2}(q-\underline{q})^2/|q-\underline{q}|^h\})] \quad (h<2). \qquad \text{(D.8)}$$

Given that for the PDF(D.8)

$\lim_{|q-\underline{q}|\to\infty} |\psi| = \lim_{|q-\underline{q}|\to\infty} n_{pg}^{1/2} = \exp[-\Lambda^2(q-\underline{q})^h/2<\Delta q^2>]$,

the quantum potential for $|q| >> |\underline{q}|$ reads:

$$\lim_{(q-\underline{q})\to\infty} V^{qu} = -(\hbar^2/2m)|\psi|^{-1}\partial^2|\psi|/\partial q\partial q = -(\hbar^2/2m)[(\Lambda^4 h^2(q-\underline{q})^{2(h-1)}/4<\Delta q^2>^2) - h(h-1)(q-\underline{q})^{(h-2)}], \qquad (D.9)$$

leading, for $h \neq 2$, to the quantum force

$$\lim_{(q-q_0)\to\infty} \partial V^{qu}/\partial q = -(\hbar^2/2m)[\Lambda^4(2h-1)h^2(q-\underline{q})^{2h-3}/4<\Delta q^2>^2) - \Lambda^2 h(h-1)(h-2)(q-\underline{q})^{(h-3)}/2<\Delta q^2>], \qquad (D.10)$$

that for $h < 3/2$ gives $\lim_{(q-\underline{q})\to\infty} \partial V^{qu}/\partial q = 0$.

References

1. F. London, Nature 141 (1938) 643.
2. P. Papon, J. Leblon, P.H.E. Meijer, The Physics of Phase Transition, Springer-Verlagh, Berlin, 2002.
3. A. M. Guenault, Statistical Physics, Kluwer Academic, Dordrecht, 1995.
4. R.P. Feynman, Phys. Rev, 91 (1953) 1291.
5. S.T. Butler, M.H. Friedman, Phys. Rev. 98 (1955) 287.
6. ibid [5] p. 294.
7. D. ter Haar, Phys. Rev. 95 (1954) 895.
8. F.A: Deeney, J.P.O'Leary, P. O'Sullivan, Phys. Lett. A 358 (2006) 53.
9. Weiner, J.H., *Statistical Mechanics of Elasticity* (John Wiley & Sons, New York, 1983), p. 317.
10. P.Chiarelli, "*Large-Scale classical behavior of fluctuating quantum states in non-linear systems*" arXiv quantum-ph 1107.4198
11. Ibid [9] p. 315.
12. Ibid [9] p. 406.
13. Y. B. Rumer, M. S. Ryvkin, *Thermodynamics, Statistical Physics, and Kinetics* (Mir Publishers, Moscow, 1980), p. 333.
14. ibid [13] p. 334.
15. ibid [13] p. 56.
16. J. B. Anderson, C. A. Traynor and B. M. Boghosian, J. Chem. Phys. **99** (1), 345 (1993).
17. R.A. Aziz and M.A. Slaman, Metrologia **27**, 211 (1990).
18. Teragon Research 2518 26th Avenue San Francisco, CA 94116, http://www.trgn.com/database/cryogen.html;
19. S. Noegi and G.D. Mahan, arXiv:0909.3078v1 (2009).
20. R. J. Donnelly and C. F. Barenghi, "The observed properties of liquid Helium at the saturated vapor pressure"; http://darkwing.uoregon.edu/~rjd/vapor1.htm.
21. ibid [13] p. 325.
22. ibid [13] p. 260.
23. Täuber, U.C., *"Field Theory Approaches to Nonequilibrium Dynamics"* arXiv: 0511743v2 **[**cond-mat**.**stat-mech] 13 Jun 2006
24. M., A., Muñoz, "Some puzzling problems in nonequilibrium field theories" arXiv:0210645 [cond-mat] 2002.

Photon diffraction and interference

J.C. Hodge[1]*
[1]Blue Ridge Community College, 100 College Dr., Flat Rock, NC, 28731-1690

Abstract

Some observations of light are inconsistent with a wave–like model. Other observations of light are inconsistent with a traditional particle–like model. A single model of light has remained a mystery. Newton's speculations, Democritus's speculations, the Bohm interpretation of quantum mechanics, and the fractal philosophy are combined. The resulting model of photon structure and dynamics is tested by toy computer experiments. The simulations include photons from a distance, in Young's experiment, and from a laser. The patterns on the screens show diffraction patterns fit by the Fresnel equation. The model is consistent with the Afshar experiment.

Interference, Young's experiment, Afshar's experiment PACS 42.50Ct, 42.25Hz, 42.25.Fx

1 INTRODUCTION

A single model of light has remained a mystery. Black body radiation, the photoelectric effect, and the Compton effect observations reject the wave-in-space model of light. The reflection, diffraction, interference, polarization, and spectrographic observations reject the traditional particle model of light. The challenge of uniting the Newtonian and quantum worlds is to develop laws of motion of photons that obtain the diffraction experimental observations. The goal is to model a set of photon characteristics that can produce interference patterns.

The popular Copenhagen interpretation of quantum mechanics suggests the simultaneous existence of both apparently mutually exclusive concepts resulting in the principle of complementarity (wave–particle duality).

The Afshar experiment (Afshar 2005) may challenge the principle of complementarity in quantum mechanics. Coherent light was passed through dual pinholes, past a series of wires placed at interference minima, and through a condensing lens. The resulting images showed the dual pinholes that suggested the which-way information had been recovered.

*E-mail:jc_hodge@blueridge.edu

1 INTRODUCTION

The Bohm interpretation of quantum mechanics is an alternative to the Copenhagen interpretation (Dürr,et al. 2009; Goldstein 2009). It is a causal, "hidden variable", and, perhaps, a deterministic model. Physics in the Newtonian and cosmological scales revolve around the concept that the motion of a single particle can be modeled. The Bohm interpretation posits particles have a definite position and momentum at all times. Particles are guided by a "pilot wave" in a Ψ–field that satisfies the Schrödinger equation, that acts on the particles to guide their path, that is ubiquitous, and that is non-local. The probabilistic nature of the Schrödinger equation results because measurements detect a statistical distribution. The origin of the Ψ–field and the dynamics of a single photon are unmodeled. The Ψ–field of Bohmian mechanics acts on particles and produces interference patterns with photons through slits.

Democritus of Abdera speculated that because different animals ate similar food, matter consists of an assembly of identical, smallest particles that recombine to form other types of matter. His term "atom" has been applied to an assembly of nucleons and electrons. Today, the Democritus concept applies as well to nucleons and smaller particles. Therefore, the Democritus "atom" is a still smaller, single type of particle.

That light moves slower through denser materials is well known. The index of refraction is the ratio of the speeds of light in differing media. Frizeau's experiment (Sartori 1996) measured a different speed of light between a medium moving toward a source and a medium moving away from a source. Both the "ether drag" model of Fresnel and the Special Theory of relativity are consistent with Frizeau's result. Because the index of refraction varies with the wavelength of light, refraction causes the analysis of light into its component wavelengths. This experimental observation is considered to decisively rule out the corpuscular model of light.

The gravitational lens phenomena may have two causes (Will 2001). The model for the gravitational lens phenomenon is that the convergent gravitational field (1) attracts light, (2) causes the curvature of coherent light, or (3) both. The gravitational attraction of photons was suggested in Newton's *Opticks* speculations. This produces a velocity $\vec{v}$ perpendicular to $\vec{c}$. This gravitational affect on light alone is insufficient to explain the gravitational lens phenomena yielding a calculated angular deviation approximately half the observed value.

The curvature-of-coherent-light model is that the light from stars is coherent and that coherence implies mutual attraction. An analogy is that of a rod with its axis maintained along the streamlines of the gravitational field. The rod travels perpendicular to its axis. The inner part of the rod slows or has a time dilation relative to the outer part. As light emerges from the convergent gravitational field, the rate of change in the direction of $\vec{c}$ and $\vec{v}$ decreases or is dampened. This is similar to the refraction of light by differing "media" densities. This model can account for the observed value of gravitational lensing (Eddington 1920, p. 209).

Models of photons have assumed the photon structure to be similar to other matter particles. That is, photons are assumed to be three-dimensional. The conclusion drawn from the well-known Michelson-Morley experiment's null re-

sult was that light is a wave, that there is no aether for light to "wave" in, and that a Fitzgerald or Lorentz type of contraction existed.

Fractal cosmology has been shown to fit astronomical data on the scale of galaxy clusters and larger (Baryshev and Teerikorpi 2002). At the opposite extreme, models such as Quantum Einstein Gravity have been presented that suggest a fractal structure on the near Planck scale (Lauscher and Reuter 2005). However, the Democritus' concept of the smallest particle suggests a lower limit of the fractal self-similarity.

The distinction between incoherent and coherent light is whether the light forms a diffraction/interference pattern when passed through one or more slits in a mask. Because interference has been thought to be a wave phenomenon, coherence is considered a property of waves. However, the definition of coherence allows a distribution of particles to be coherent.

Coherence is obtained (1) by light traveling a long distance such as from a star as seen in the Airy disk pattern in telescope images, (2) by the pinhole in the first mask in Young's experiment, and (3) from a laser.

Young's double-slit experiment has become a classic experiment because it demonstrates the central mysteries of the physics and of the philosophy of the very small. Incoherent light from a source such as a flame or incandescent light impinges on a mask and through a small pinhole or slit. The light through the first slit shows no diffraction effects. The light that passes through the slit is allowed to impinge on a second mask with two narrow, close slits. The light that passes through the two slits produces an interference pattern on a distant screen. The first slit makes the light coherent. The intensity pattern on the screen is described by the Huygens-Fresnel equation.

The assumptions of Fresnel model of diffraction include: (1) The Huygens' Principle that each point in a wave front emits a secondary wavelet. (2) The wavelets destructive and constructive interference produces the diffraction pattern. (3) The secondary waves are emitted in only the forward direction, which is the so called "obliquity factor" (a cosine function). (4) The wavelet phase advances by one-quarter period ahead of the wave that produced them. (5) The wave has a uniform amplitude and phase over the wave front in the slit and zero amplitude and no effect from behind the mask. (6) Note the Fresnel model has a slight arc of the wave front across the slit. That is, the distribution of energy in the plane of the slit varies. The Fresnel model with larger distance between the mask and the screen or with condensing lenses before and after the mask degenerates into the Fraunhofer diffraction model.

The intensity patterns produced by multiple slits can be compared to the intensities of the single slit pattern of equal total width. Thus, the resulting pattern may be regarded as due to the joint action of interference between the waves coming from corresponding points in the multiple slits and of diffraction from each slit. Diffraction in the Fresnel model is the result of interference of all the secondary wavelets radiating from the different elements of the wave front. The term diffraction is reserved for the consideration of the radiating elements that is usually stated as integration over the infinitesimal elements of the wave front. Interference is reserved for the superposition of existing waves, which is

usually stated as the superposition of a finite number of beams.

Sommerfield (1954) gave a more refined, vector treatment for phase application. However, these other more complex models make many simplifying assumptions. When the apertures are many wavelengths wide and the observations are made at appreciable distance from the screen, intensity results are identical to the Fresnel model.

Newton in his book Opticks (1730) speculated light was a stream (ray) of particles. The aether in query 17 overtakes (travels faster) the rays of light and directs the rays' path. Newton's analogy was of water waves. That is, Newton was using a self-similarity (fractal) postulate. The rays of light recede from denser parts of the aether in query 19. The aether grows denser from bodies in query 20 and this causes gravity in query 21. Newton seems to have suggested light is particles that are directed by the aether to produce the wave phenomena. The prevailing models of the 19th century considered light to be a wave. The prevailing interpretation of Newton's model is that Newton was suggesting light is both a wave and a particle rather than two entities having differing effects like a rock (photon) creating transverse waves in water (aether).

The Maxwell equations followed by the Special Theory of Relativity posited the velocity $\vec{c}$ of photons was constant in the absence of matter. This says nothing about the speed of gravity (aether) waves.

The cosmological, scalar potential model (SPM) was derived from considerations of galaxies and galaxy clusters (Hodge 2004, 2006a,b,c,d,e) [1]. The SPM posits a plenum exists whose density distribution creates a scalar potential ρ (erg) field. The term "plenum" was chosen to distinguish the concept from "space" in the relativity sense and from "aether". The term "space" is reserved for a passive backdrop to measure distance, which is a mathematical construct. The plenum follows Descartes (1644) description of the plenum. The plenum is infinitely divisible, fills all volume between matter particles, is ubiquitous, flows to volumes according to the heat equation, is influenced by matter, is compressible in the sense that the amount of plenum in a volume may change, and influences matter.

Consideration of the electromagnetic signal of the Pioneer Anomaly led to the postulate that matter caused a depression in the plenum (Hodge 2006e). The redshift of the Pioneer Anomaly on the solar system scale ignored several terms because they had significant values only on the galactic scale (Hodge 2006e). In addition to the propositions of the galactic SPM, matter is posited to cause a static[2] warp in the ρ field in accordance with the Newtonian spherical property. "Static" because matter is neither a Source nor a Sink of energy. Because the ρ field near matter must attract other matter, the matter decreases the ρ field. The ρ field then causes matter attraction according to established gravitational physics and causes the frequency change of the EM signal[3]. Matter merely modifies the ρ energy flowing from Sources to Sinks. The SPM of the Pioneer

[1] A more complete discussion of this model is in Hodge (Theory of Everything)

[2] "Static" such as caused by a stationary electron in a stationary EM field.

[3] This concept is from General Relativity where matter shapes the geometrical "space" and "space" directs matter.

Anomaly is an effect on only the EM signal and is a blueshift superimposed on the Doppler redshift of the receding spacecraft.

Hodge (2004) showed that if the surface of the hod held $\rho = 0$, the ρ equipotential surfaces were ellipsoidal near the surface of the hod. The streamlines terminated on and perpendicular to the surface. Farther from the hod, the equipotential surfaces become spheres and the hod appears as a point at the center of the sphere. Newtonian equations apply at larger distance. The $\rho \geq 0$ always including the scale of photons and, possibly, atomic nuclei[4]. Because the hod has extent, the effect on ρ from a distance is treated as a negative ρ value at the center of the hod. Therefore, the $\rho < 0$ in matter equations reflects the hod having a surface area.

This Paper proposes a model of light that postulates the necessary characteristics of photons to satisfy observations and yield diffraction phenomenon. The model combines Newton's speculations, Democritus's speculations, the Bohm interpretation of quantum mechanics, and the fractal philosophy. The wave–like behavior of light results from the photons changing the Ψ–field that guides the path of the photons. The resulting model is tested by numerical simulation of diffraction and interference, with application to the Afshar experiment. Therefore, the wave characteristics of light may be obtained from the interaction of photons and Ψ–field.

In section 2, the model of photons is described and the equations are derived. Section 3 describes the computer simulation. Section 4 describes the computer simulation of photons traveling a long distance. Section 5 describes the computer simulation of Young's experiment. Section 6 describes the computer simulation of laser light. Section 7 describes the model application to the Afshar experiment. The Discussion and Conclusion are in section 8 and section 9, respectively.

2 Model

Newton's third law suggests that if the Ψ–field acts on photons, the photons and other matter should act on the Ψ–field.

Compare coherent light passing through multiple slits in a mask to light passing through a single slit. The slits may be viewed as transmitters of light that produces a diffraction pattern in the Ψ–field if the incoming light is coherent. The fractal philosophy of self-similarity suggests that the photons passing through a slit have similar elements that emit waves. If light is a particle, then the energy of a photon is smaller than the other known particles. Differing energy levels (color) among photons suggests light is an assembly of smaller, identical particles (hods). Like the Democritus's argument, the proportionality of mass and energy suggests mass is also an assembly of hods. Each photon must be coherent, must be emitting a diffraction pattern into the Ψ–field, and

[4]The action of the plenum on atomic nuclei at the largest was suggested by the differing Hα and HI galactic rotation curves (Hodge 2006c).

be composed of smaller emitters. That is, moving photons emit coherent forward waves in the Ψ–field like multiple slits. The analogy of a photon, of a slit, and of multiple slits is analogous to a linear antenna array of smaller emitters in a photon.

Photons exhibit gravitational effects caused by matter perpendicular to their direction of travel. The Michelson-Morley experiment suggests the hods and the photon have no resistance or drag in the Ψ–field in the direction of movement. This suggests the hod has zero thickness. Asymmetry in photon behavior near mass suggests asymmetry in photon structure. Therefore, the hod is a two dimensional object. The simplest structure that can produce a coherent wave is one hod next to another. Additional hods create a column of hods.

Each hod of a photon attracts Ψ at its surface to its maximum value. Between these hod surfaces and very near the hods of a photon, a $\Psi = 0$ surface is created. Beyond this surface, $\vec{\nabla}\Psi$ is directed first toward then away from a hod by the other hods. This holds the hods together. The Ψ–field cavitates as hods travel through the Ψ–field like an air foil traveling at the maximum speed through a gas.

If a hod transmits Ψ changes into the Ψ–field, the hod also receives changes in Ψ. The resulting force $\vec{F}_{\rm s}$ of the Ψ–field acts on a hod to change $\vec{v}$ and to rotate the hod. The $\vec{v}$ is the result of a force $\vec{F}_{\rm s}$ acting on the hod cross–section proportional to $\vec{n} \bullet \vec{\nabla}\Psi$ on the cross-section $m_{\rm s}$ of matter where $\vec{}$ indicates a vector in 3–space,

$$\vec{F}_{\rm s} \propto m_{\rm s}(\vec{n} \bullet \vec{\nabla}\Psi)\vec{n}, \tag{1}$$

where $\vec{n}$ is the surface, normal, unit vector. A vector without the $\vec{}$ indicates the value of the vector.

The Ψ change that results from the effect of other matter M on the Ψ–field is

$$\Psi = \sum_{i}^{N} GM_i/R_i, \tag{2}$$

where N is the number of bodies used in the calculation and R is the distance from the center of mass of M to the point where Ψ is measured. The R is large enough that the equipotential surfaces are approximately spherical. The M is linearly related to $N_{\rm h}$ where $N_{\rm h}$ is the number of hods in the body.

The gravitational lens phenomena suggest the photon's response to $\vec{\nabla}\Psi$ changes relative to the direction of $\vec{c}$. The Shapiro delay phenomena could be modeled as a decrease of c in a gravitational field. This also suggests c is a function of the gravitational field. The lower c suggested by the Shapiro delay in a lower Ψ–field causes a refraction in the direction of $\vec{c}$. The matter caused variation of the Ψ–field and the action of the Ψ–field on light causes $\vec{c}$ to change. The plane of the Michelson-Morley experiment was perpendicular to the Earth's surface and, therefore, in a nearly constant Ψ–field.

The $\vec{\nabla}\Psi$ produces the effect of gravity. The speed of the gravity wave (Ψ wave) that are changes in the force exerted by $\vec{\nabla}\Psi$ is much greater than c (van Flandern 1998). The development of Special Relativity was done using

electromagnetic experimental results. The speed limit of electromagnetic force changes is c. The higher speed of the $\vec{\nabla}\Psi$ wave is consistent with the quantum entanglement phenomena and is necessary if photons are to project the changing Ψ–field forward.

The redshift and blueshift of the Pound-Rebka experiment suggests a dependence of photon energy shift on the changing Ψ–field. That is, the energy of a photon is partly composed of the Ψ–field between the hods of the column as well as the number of hods in a photon. If the hods in a photon have no Ψ–field between them, the photon would have no third dimensional characteristics. Because a blueshift increase of energy occurs when Ψ decreases, because c decreases with Ψ, and if number N_{h} of hods in a photon remains constant, the inertial mass m_{I} of kinetic energy is posited to be

$$m_{\mathrm{I}} = N_{\mathrm{h}}(1 + K_{\Psi}\frac{\Psi_{\max}}{\Psi}), \tag{3}$$

where K_{Ψ} is a proportionality constant.

If c is much lower than the wave velocity of the Ψ–field, then the limiting factor of c may be like a terminal velocity. The Shapiro delay and the well known $E = mc^2$ relation combined with the lower Ψ near mass suggest for short range experiments,

$$c = K_{\mathrm{c}}(\frac{\Psi}{\Psi_{\max}})^2, \tag{4}$$

where K_{c} is the proportionality constant and $\Psi_{\max}$ is the maximum Ψ in the path of light wherein c is measured.

2.1 Hod action on Ψ–field

The action of all matter (hods) causes the Ψ to change. Therefore, the motion of a hod is relative to the local Ψ–field variation.

Matter causes the Ψ to decrease. Therefore the hod motion which pulls the Ψ to zero displaces the Ψ–field. The Ψ–field then flows back when the hod passes a point. That is, the hod's motion within a volume neither increases nor decreases the amount of Ψ–field in that volume. The cavitation in the Ψ–field produces a wave in the Ψ–field. The cavitation limits the c. The cavitation depends on the density of the Ψ–field, higher density allows a faster refresh of the Ψ round the hod and, therefore, a faster c than in a low Ψ–field. The wave has a $-\cos(kr)$ affect from the point of the hod center. Because the Ψ–field wave travels faster than the hod, the Ψ–field flows around the hod to fill in the depression behind the hod. If the speed of the wave in the Ψ–field is much larger than c, the harmonic wave is transmitted forward and the Ψ level behind the wave reaches a stable, non-oscillating level very rapidly and is the effect of gravity. This can be modeled as a $\cos(K_{\beta}\beta)$ decrease of Ψ's value, where β is the angle between the direction $\vec{c}$ of a hod and the direction of the point from the hod center where Ψ is calculated. The angle at which the Ψ–field no longer oscillates has the $\cos(K_{\beta}\beta) = 0$. This is analogous to the Fresnel model that

secondary wavelets are emitted in only the forward direction. Therefore, the effect on the Ψ–field of a single hod is

$$\begin{aligned} \Psi_{\text{single}} = & \quad -\tfrac{K_r}{r}\cos\left(\tfrac{2\pi r}{\lambda_{\text{T}}}\right)\cos(K_\beta\beta)\exp^{-j(\omega t)} \qquad K_\beta\beta < \pi/2 \\ \Psi_{\text{single}} = & \quad -K_r r^{-1} \qquad\qquad\qquad\qquad\qquad\quad K_\beta\beta \geq \pi/2, \end{aligned} \tag{5}$$

where K_r is a proportionality constant, $j = \sqrt{(-1)}$, $\exp^{-j(\omega t)}$ is the wave time dependent component in the Ψ–field, and λ_{T} is the radial wavelength of the forward wave.

2.2 Ψ–field action on a hod

If energy may be transferred to the hods from the Ψ–field, then the motion of the hods may be dampened by the Ψ–field. Posit the dampening force is proportional to the Ψ, m_{s}, and $\vec{v}$ in analogy to non–turbulent movement through a medium. Non–turbulent flow is because the Ψ–field is ubiquitous and the speed of changes in the Ψ–field is much greater than c. Therefore,

$$m_{\text{I}}\dot{v} = K_{\text{v}}F_{\text{st}} - K_{\text{d}}m_{\text{s}}\Psi v, \tag{6}$$

where the over dot means a time derivative, K_{v} and K_{d} are proportionality constants, $F_{\text{st}} \propto m_{\text{s}}|\vec{\nabla}\Psi|\sin(\theta)$, and θ is the angle between $\vec{\nabla}\Psi$ and $\vec{c}$.

Solving for v yields:

$$v_{\text{f}} = \frac{(A_{\text{vh}} + B_{\text{vh}}v_{\text{i}})\exp^{B_{\text{vh}}\Delta t} - A_{\text{vh}}}{B_{\text{vh}}}, \tag{7}$$

where t is the time of measurement, $v_{\text{i}} = v(\text{time} = t)$, $v_{\text{f}} = v(\text{time} = t + \Delta t)$,
$A_{\text{vh}} = K_{\text{v}}m_{\text{s}}\vec{\nabla}\Psi\sin(\theta)/m_{\text{I}}$, and
$B_{\text{vh}} = -K_{\text{d}}m_{\text{s}}\Psi/m_{\text{I}}$.

Posit the action of the Ψ–field on a single hod is acting with the same radial r component of the hod. The $\vec{r}$ is directed along the hod surface in the direction of $\vec{c}$. If the hod has a m_{s} and circular shape, $\dot{r} = 0$. The m_{s} and r are constant and the same for all hods. Because the hod is free to move, a change in orientation of the hod results in the hod changing direction. A differing $\vec{F}_{\text{s}}$ from one side of the hod surface to the other along $\vec{r}$ will cause a rotation of the hod. Assuming the hod is rigid is unnecessary, but convenient. If the $\vec{F}_{\text{s}}$ is close to $\vec{c}$, the rotation will be to align $\vec{c}$ with $\vec{F}_{\text{s}}$. If the $\vec{F}_{\text{s}}$ is close to the axis of the photon, the rotation will be to align the axis of the photon with $\vec{F}_{\text{s}}$. Define α as the critical angle between $\vec{c}$ and the angle at which $\vec{F}_{\text{s}}$ changes from acting to align with the axis to acting to align with $\vec{c}$. Because the dampening is in only the axial direction, the dampening caused by rotation is also proportional to the Ψ, m_{s}, and rotational velocity $\dot{\theta}$ in analogy to non–turbulent movement through a medium.

The torque from $\vec{F}_{\text{s}}$ is $rF_{\text{sa}} = K_{\theta\text{s}}m_{\text{s}}r|\vec{\nabla}\Psi|\sin(\theta)$ or

$$\frac{\text{d}}{\text{d}t}(r^2 m_{\text{I}}\dot{\theta}) = rF_{\text{sa}} - K_{\theta\text{d}}rm_{\text{s}}\Psi\dot{\theta}, \tag{8}$$

Solving for $\dot{\theta}$ yields:

$$\dot{\theta}_{\rm f} = \frac{(A_{\theta{\rm h}} + B_{\theta{\rm h}}\dot{\theta}_{\rm i})\exp^{B_{\theta{\rm h}}\Delta t} - A_{\theta{\rm h}}}{B_{\theta{\rm h}}}, \tag{9}$$

where $\dot{\theta}_{\rm i} = \dot{\theta}({\rm time} = t)$, $\dot{\theta}_{\rm f} = \dot{\theta}({\rm time} = t + \Delta t)$,
$A_{\theta{\rm h}} = K_{\theta{\rm s}} m_{\rm s} |\vec{\nabla}\Psi| \sin(\theta)/r m_{\rm I}$,
$B_{\theta{\rm h}} = -K_{\theta{\rm d}} \frac{m_{\rm s}}{r} \frac{\Psi}{m_{\rm I}} - (\dot{m}_{\rm I}/m_{\rm I})$, and
$\dot{m}_{\rm I}/m_{\rm I}$ is small and nearly constant.

2.3 Photon action on Ψ–field

The suggested structure of the photon in Hodge (2004) implies each hod is positioned at a minimum Ψ because the hod surface holds the $\Psi = 0$. Therefore, the distance between hods in a photon is one wavelength $\lambda_{\rm T}$ of the emitted Ψ–field wave. Also, the hods of a photon emit in phase. Further, the number of hods in a photon and the Ψ–field potential Ψ_{max} due to all other causes around the hod effects this distance. Therefore, the transmitted potential Ψ_T from a photon is

$$\Psi_T = -\frac{K_r}{r} N_{\rm effT}, \tag{10}$$

where

$$\begin{aligned} N_{\rm effT} = &\quad \cos\left(\tfrac{2\pi r}{\lambda_{\rm T}}\right)\cos(K_\beta\beta)\left|\tfrac{\sin[N_{\rm hT}\pi\sin(\beta)]}{\sin[\pi\sin(\beta)]}\right| &\quad K_\beta\beta < \pi/2 \\ N_{\rm effT} = &\quad N_{\rm hT} &\quad K_\beta\beta \geq \pi/2 \end{aligned} \tag{11}$$

and $\lambda_{\rm T} = K_\lambda/(\Psi_{max} N_{\rm hT})$ and $N_{\rm hT}$ is the number of hods in the transmitting photon.

These equations apply to the plane that includes the center of the photon, the direction vector, and the axis of the photon.

2.4 Ψ–field action on a photon

The Ψ–field wave from each cause impinges on a photon through the hods. Because the photon is linearily extended, the photon is analogous to the receiving antenna. Therefore,

$$\vec{\nabla}\Psi_{{\rm eff}i} = \vec{\nabla}\Psi_i \left| \frac{\sin[\frac{N_{\rm hR}\pi\lambda_{{\rm T}i}}{\lambda_{\rm R}}\sin(\beta + \frac{2\pi\lambda_{{\rm T}i}}{\lambda_{\rm R}})]}{\sin[\pi\sin(\beta + \frac{2\pi\lambda_{{\rm T}i}}{\lambda_{\rm R}})]} \right|, \tag{12}$$

where $N_{\rm hR}$ is the number of hods in the receiving photon; $\lambda_{\rm R}$ is the wavelength, which is the distance between hods, of the receiving photon; and Ψ_i, $\lambda_{{\rm T}i}$, and $\vec{\nabla}\Psi_{{\rm eff}j}$ is the effective Ψ, $\lambda_{\rm T}$, $\vec{\nabla}\Psi$, respectively, for the $i^{\rm th}$ cause.

Using Eq. (12) in Eqns. (7) and (9) and summing the effect of all causes yields the total change in a Δt.

3 Simulation

A particle in the computer experiment was a set of numbers representing the points in position and momentum space whose motion was advanced in discrete time intervals of one calculation cycle. The particles position at the previous interval determines Ψ–field effect by the equations developed in section 2. The limitations of a computer experiment are caused by (1) noise from using a relatively small number of particles and the relative distances used, (2) bias caused by the approximate solution limitation of the number of digits, (3) the discrete time interval rather than continuous time progression, and (4) other possible effects. Understanding the results requires insight that can be obtained by experimental observation in the real world.

The simulation was conducted using a PC running a Visual Basic V program on the Windows XP platform. The horizontal axis (Y-axis) was the initial direction of the photons. The vertical axis (X-axis) was perpendicular to the initial photon direction. The n^{th} photon p_n had X and Y position coordinates, $P_\mathrm{x}(p_\mathrm{n})$ and $P_\mathrm{y}(p_\mathrm{n})$, respectively, and the parameter values required by the equations. "Masks" were simulated by a range of y values at specified Y coordinates for all x values, a zone wherein photons were deleted, and a gap wherein photons were unchanged that defined the "slit". Each calculation was an "interval". Each "step" was one unit of distance measure along the X-axis or Y-axis. The values of $\Delta t = m_\mathrm{s} = 1$ were used. The $y = 0$ was the initial value of introduced photons. The $x = 0$ was the center of the initial photon distribution, the center of the one slit gap, and the center of the two slit gaps.

A "screen" was a virtual, vertical column positioned at a constant Y wherein the lower X coordinate x_s of the bin range determines the value that was recorded. Each X step was divided into 50 bins. As photons pass through the screen, the number $B(x_\mathrm{s})$ of photons at each bin was counted. Each photon was eliminated after being counted. Therefore, the accumulation of the number of photons through each bin over several intervals was the intensity of light at each bin. After 1000 photons were counted, x_s and $B(x_\mathrm{s})$ were recorded in an Excel file.

Because of the sampling error and of calculation errors noted above, the calculation of the best-fit Fresnel curve used an average

$$\bar{B}(x_\mathrm{s}) = \sum_{i=x_\mathrm{s}-0.1}^{i=x_\mathrm{s}+0.1} B(i)/11, \tag{13}$$

where i increments by 0.02 steps.

All the photons in these experiments were identical. That is, $N_\mathrm{hT} = N_\mathrm{hR} = 10$ and $\lambda_\mathrm{T} = \lambda_\mathrm{R}$. The constants in the equations are listed in Table 1 and were determined by trial and error. The difficulty in determining the constants was so the angle (x_s/L), where L (step) is the distance from the last mask to the screen, corresponded to the Fresnel equations. That is, so that $x = 1$ step and $y = 1$ step was the same distance of photon movement for the calculations, which is a Cartesian space.

Table 1: The values of the constants.

Parameter	value	units
$K_{\rm c}$	1×10^{-8}	step interval^{-1}
K_{Ψ}	1×10^{1}	gr. hod^{-1}
K_{β}	1.1	
$K_{\rm r}$	-8×10^{-11}	erg step
$K_{\rm v}$	1×10^{-5}	
$K_{\rm d}$	2.4×10^{-4}	gr. step^{-2} erg^{-1} interval^{-1}
α	5×10^{-1}	radian
$K_{\theta {\rm s}}$	6×10^{1}	gr. interval^{-2} erg^{-1}
$K_{\theta {\rm d}}$	2×10^{1}	gr. step^{-2} erg^{-1} interval^{-1}
K_{λ}	1×10^{-4}	erg step hod
$\Psi_{\rm max}$	1×10^{3}	erg

Table 2 lists the curve fit measurements for the plots shown in the referenced figures. The first column is the figure number showing the curve. The second column is the number $N_{\rm c}$ of photons counted at the screen. The third column is the center of the theoretical curve ($K_{\rm cent}$). The fourth column is the asymmetry $A_{\rm sym}$ of the data points (number of photons counted greater than $K_{\rm cent}$ minus number of photons counted less than $K_{\rm cent}$)/$N_{\rm c}$. The fifth column is the sum of the least squares $L_{\rm sq}$ between the $\bar{B}(x_{\rm s})$ and the theoretical plot for -5 steps$\leq$ $x_{\rm s} \leq$ 5 steps divided by $N_{\rm c}$. The sixth column is the correlation coefficient between $\bar{B}(x_{\rm s})$ and the theoretical plot for -5 steps $\leq x_{\rm s} \leq$ 5 steps.

4 Photons traveling a long distance

The initial distribution of photons was within a 60 steps by 60 steps section. One photon was randomly placed in each 1 step by 1 step part. The equations were applied to each of the photons in the section. Because the section has an outer edge, additional virtual photons were necessary to calculate the Ψ. Therefore, $[P_{\rm x}(p_{\rm n}),\ P_{\rm y}(p_{\rm n})]$, $[P_{\rm x}(p_{\rm n}),\ P_{\rm y}(p_{\rm n}) + 60]$, $[P_{\rm x}(p_{\rm n}),\ P_{\rm y}(p_{\rm n}) - 60]$, $[P_{\rm x}(p_{\rm n}) + 60,$ $P_{\rm y}(p_{\rm n})]$, $[P_{\rm x}(p_{\rm n})-60, P_{\rm y}(p_{\rm n})]$ $[P_{\rm x}(p_{\rm n})+60, P_{\rm y}(p_{\rm n})+60]$, $[P_{\rm x}(p_{\rm n})-60, P_{\rm y}(p_{\rm n})+60]$, $[P_{\rm x}(p_{\rm n})$-60, $P_{\rm y}(p_{\rm n}) - 60]$, and $[P_{\rm x}(p_{\rm n})+60,\ P_{\rm y}(p_{\rm n}) - 60]$ were included in the calculation of Ψ. Only photons and virtual photons within a radius of 30 steps were used in the calculation of Ψ at a point.

The equations were developed without consideration of photons colliding. That is, some photons were too close which generated very large, unrealistic Ψ values. Another characteristic of the toy model is that each interval moves a photon a discrete distance that occasionally places photons unrealistically close. When this happened, one of the close photons was eliminated from consideration. The initial distribution started with 3600 photons from which 111 were eliminated because of collision.

Figure 1(left) is a plot of the $N_{\rm effT}$ versus β with $N_{\rm hT} = 10$. The first five

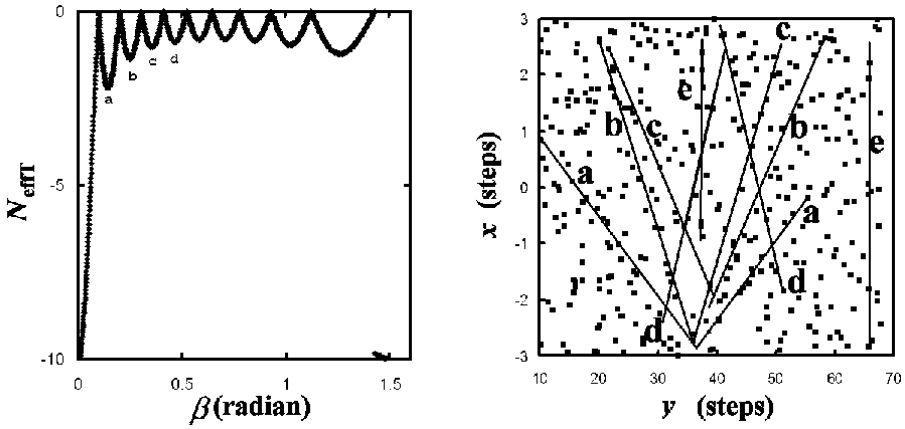


Figure 1: The left figure is a plot of the N_{effT} versus angle from the direction of the photon β $N_{\mathrm{hT}} = 10$. The first six minima are at $\beta = 0$ rad, 0.144 rad (a), 0.249 rad (b), 0.355 rad. (c), and 0.466 rad (d). The right figure is a plot of the longitudinal vs. latitudinal position of photons after 1000 intervals. The line labeled "e" is $\beta = \pi/2$. The lines are labeled as the angles in the left plot. The position of photons along lines corresponding to minima of a photons transmission pattern is what determines coherence.

peaks are at $\beta = 0$ rad, 0.144 rad (a), 0.249 rad (b), 0.355 rad. (c), 0.466 rad (d), and $\pi/2$ rad (e). This pattern is similar to the antenna emission pattern.

After 1000 intervals, a pattern of photon position developed as seen in Fig.1(right). The photons positions were recorded. The photons were organizing themselves into recognizable patterns of lines with angles to the direction of travel (Y axis) corresponding to the minima of Fig. 1(left).

A mask with a single slit with a width $W_{\mathrm{s}} = 1$ step was placed at $y = 100$ steps and a screen was placed at $y = 140$ steps ($L = 40$ steps). The positions of the photons were read from the recording. The group of photons were placed in 60 steps increments rearward from the mask. The photons were selected from the recording to form a beam width W_{in} (step) centered on the $x = 0$ step axis. Because the incoming beam had edges, the calculation for $P_{\mathrm{yold}}(p_{\mathrm{n}}) < 100$ was $P_{\mathrm{ynew}}(p_{\mathrm{n}}) = P_{\mathrm{yold}}(p_{\mathrm{n}}) + K_{\mathrm{c}} * 1$ interval, where $P_{\mathrm{yold}}(p_{\mathrm{n}})$ is the position of the n^{th} photon from the last calculation and $P_{\mathrm{ynew}}(p_{\mathrm{n}})$ is the newly calculated position. The $P_{\mathrm{x}}(p_{\mathrm{n}})$ remained the same. If $P_{\mathrm{yold}}(p_{\mathrm{n}}) \geq 100$, the $P_{\mathrm{ynew}}(p_{\mathrm{n}})$ and $P_{\mathrm{x}}(p_{\mathrm{n}})$ were calculated according to the model.

Figure 2 shows the resulting patterns for varying W_{in}. The thicker, solid line in each figure is the result of a Fresnel equation fit to the data points. Although each plot shows a good fit to the Fresnel equation, the fit differs among the plots and depends on W_{in}. Because the calculation includes all photons, the photons that were destined to be removed by the mask have an affect on the diffraction pattern beyond the mask.

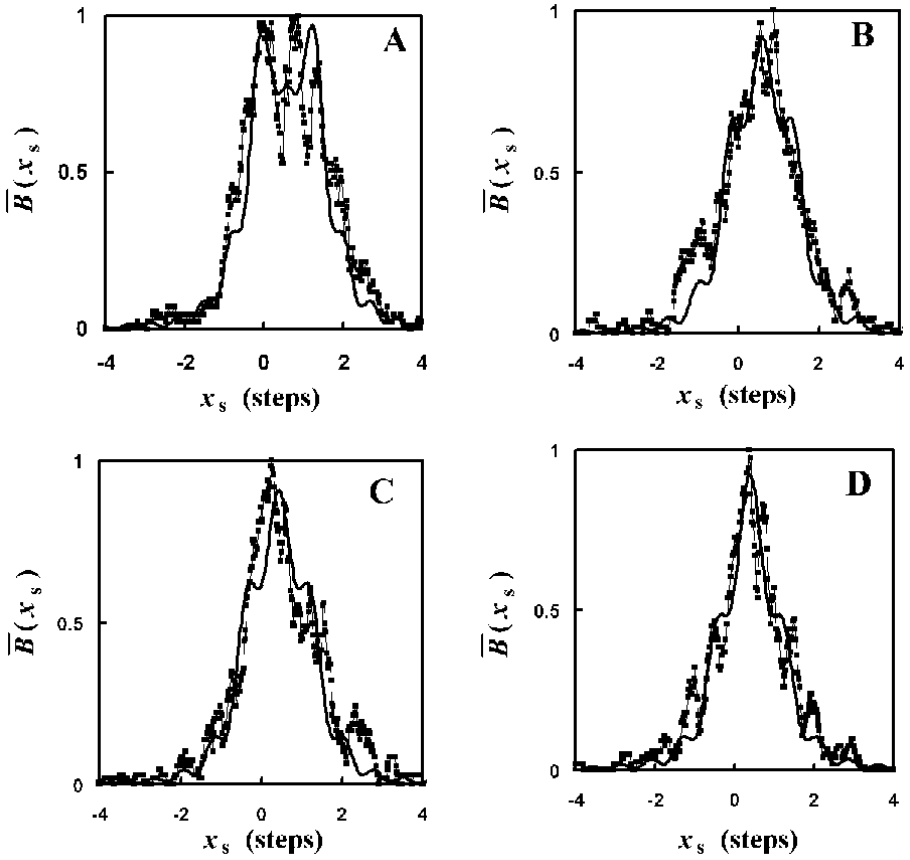


Figure 2: The single slit width $W_s = 1.0$ step screen patterns for $L = 40$ steps: (A) input beam width $W_{in} = 1.0$ step which is the same as the slit width, (B) $W_{in} = 2.0$ steps, (C) $W_{in} = 4.0$ steps, and (D) $W_{in} = 6.0$ steps. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation. Although each plot shows a good fit to the Fresnel equation, the fit differs among the plots and depends on W_{in}. Because the calculation includes all photons, the photons that were destined to be removed by the mask have an effect on the diffraction pattern beyond the mask.

Table 2: The measurements of the curve fit.

Fig.	N_{c}[a]	K_{cent}[b]	A_{sym}[c]	L_{sq}[d]	C_{c}[e]
2A	1000	0.585	0.014	0.219	0.96
2B	1000	0.605	-0.028	0.199	0.97
2C	1000	0.408	0.000	0.247	0.96
2D	1000	0.383	0.006	0.259	0.96
3A	1000	0.259	0.084	0.360	0.97
3B	1000	0.426	0.050	0.021	0.92
3C[f]	1000	0.271	0.060	0.185	0.98
3D[f]	1000	0.535	0.006	0.246	0.83
5A	438	0.187	-0.046	0.178	0.92
5B	1000	0.314	-0.070	0.717	0.82
8	400	0.850	-0.320	0.221	0.68
9	1000	0.145	-0.122	0.909	0.79
11	1000	0.273	-0.074	0.895	0.81

[a] The number of photons counted at the sceen.
[b] The center of the theoretical curve.
[c] The (number of photons counted greater than K_{cent} minus number of photons counted less than K_{cent})/N_{c}.
[d] The sum of the least squares between the $\bar{B}(x_{\mathrm{s}})$ and the theoretical plot for -5 steps $\leq x_{\mathrm{s}} \leq$ 5 steps.
[e] The correlation coefficient between $\bar{B}(x_{\mathrm{s}})$ and the theoretical plot for -5 steps $\leq x_{\mathrm{s}} \leq$ 5 steps.
[f] This curve is relative to a Fraunhofer equation fit.

Figure 3 shows the resulting patterns for varying L. The mask, screen, and photon input was the same as the previous experiment with $W_{\rm in} = 6$ steps. Comparing Fig. 3(A), Fig. 2(D), and Fig. 3(B) shows the evolution of the diffraction pattern with $L = 30$ steps, $L = 40$ steps, and $L = 50$ steps, respectively. Fig. 3(C) and Fig. 3(D) show the Fraunhofer equation fits. The greater L produces a closer match between the Fresnel and Fraunhofer equation fits.

Figure 4 shows an expanded view of Fig. 3(B). The center area, first ring, and second ring of Fig. 3(B) and Fig. 4 has 954 photons, 20 photons, and 4 photons, respectively, of the 1000 photons counted.

Figure 5 shows the screen pattern with the mask from the previous experiment replaced by a double slit mask. Figure 5(A) was with the slits placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step. The best two slit Fresnel fit (a cosine term multiplied be the one slit Fresnel equation) is expected for slits with a ratio of the width b of one slit to the width d between the centers of the slits (the "d/b ratio"). Figure 5(B) shows the screen pattern with the slits placed from 0.75 step to 1.25 steps and from -1.25 steps to -0.75 step.

Figure 6 shows the paths traced by 10 consecutive photons through the slits at two different intervals that form part of the distribution of Fig. 5A. The traces are from the mask to the screen. The θ for each photon is established after $y = 130$ steps. Before $y = 120$, there is considerable change in θ, which is consistent with Fig. 3. That is, the photon paths do not start at the slit and follow straight lines to the screen. The Fresnel equation applies only after some distance from the mask.

The numbers in Fig. 6 mark the following occurrences: (1) One photon follows another and traces the same path. The following photon travels a longer path before path merging. (2) One photon follows another and traces a parallel and close path. (3) A photon experiences an abrupt change in θ as it passes close to another photon. These events were probably a result of the discrete nature of the simulation like the collision condition noted previously. (4) A photon from one slit follows another from the other slit. The leading photon determines the $x_{\rm s}$ and θ at the screen.

5 Young's experiment

The input to Young's experiment was at 0 step $\leq y \leq$ 1 step and -3 steps $\leq x \leq$ 3 steps. A photon is placed at random in each 1 step X 1 step area. The input was repeated every 100 intervals. The first mask was place at y=100 steps with $W_{\rm s} = 1$ step centered on the X-axis. If a photon's $P_{\rm y}(p_{\rm n}) \leq 100$ step, then $P_{\rm ynew}(p_{\rm n}) = P_{\rm yold}(p_{\rm n}) + K_{\rm c} * 1$ interval. If a photon's $P_{\rm yold}(p_{\rm n}) > 100$ step, then the new position was calculated in accordance with the model.

A screen placed at y= 140 steps showed a rounded pattern between -4 steps $\leq x_{\rm s} \leq 4$ steps. Comparison with Fig. 2 diffraction patterns shows Fig. 7 is not a diffraction pattern for $L = 40$ steps.

Figure 7 (right) shows the distribution of photons between the first mask and the screen. The lines and the lower case letters are as in Fig. 1.

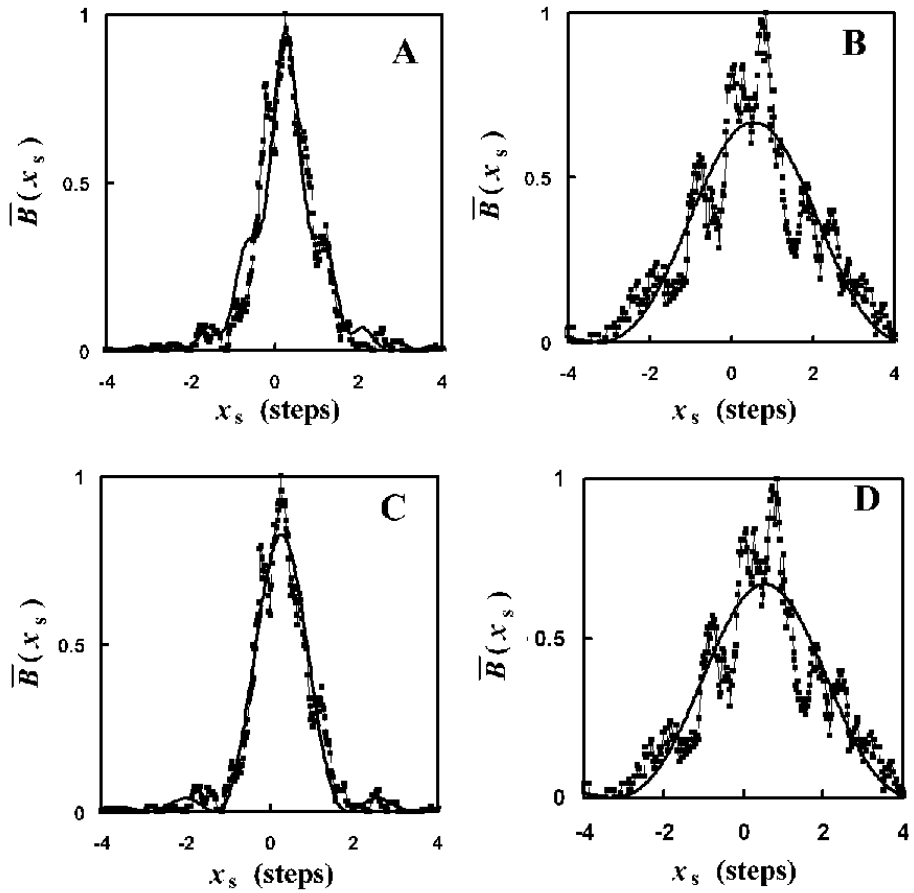


Figure 3: Resulting patterns for varying L. The mask, screen, and photon input was the same as the previous experiment with $W_{\text{in}} = 6$ steps. The single slit $W_{\text{s}} = 1.0$ step screen patterns for $L = 30$ steps (left figures A and C) and for $L = 50$ steps (right figures B and D). The top row is the Fresnel calculation plots and the bottom row is the Fraunhofer calculation plots. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation. Comparing Fig. 3(A), Fig. 2(D), and Fig. 3(B) shows the evolution of the diffraction pattern with $L = 30$ steps, $L = 40$ steps, and $L = 50$ steps, respectively. Fig. 3(C) and Fig. 3(D) show the Fraunhofer equation fits. The greater L produces a closer match between the Fresnel and Fraunhofer equation fits.

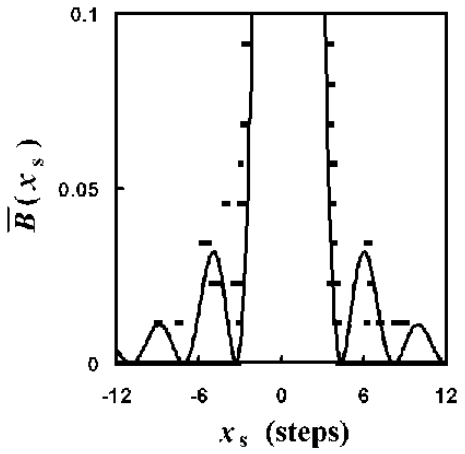


Figure 4: Plot of Fig. 3B with an expanded scale to show the second and third diffraction rings. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation. The center area, first ring, and second ring of Fig. 3(B) and Fig. 4 has 954 photons, 20 photons, and 4 photons, respectively, of the 1000 photons counted. The number of photons in each ring agrees with the theoretical calculation of the relative intensity of the diffraction rings.

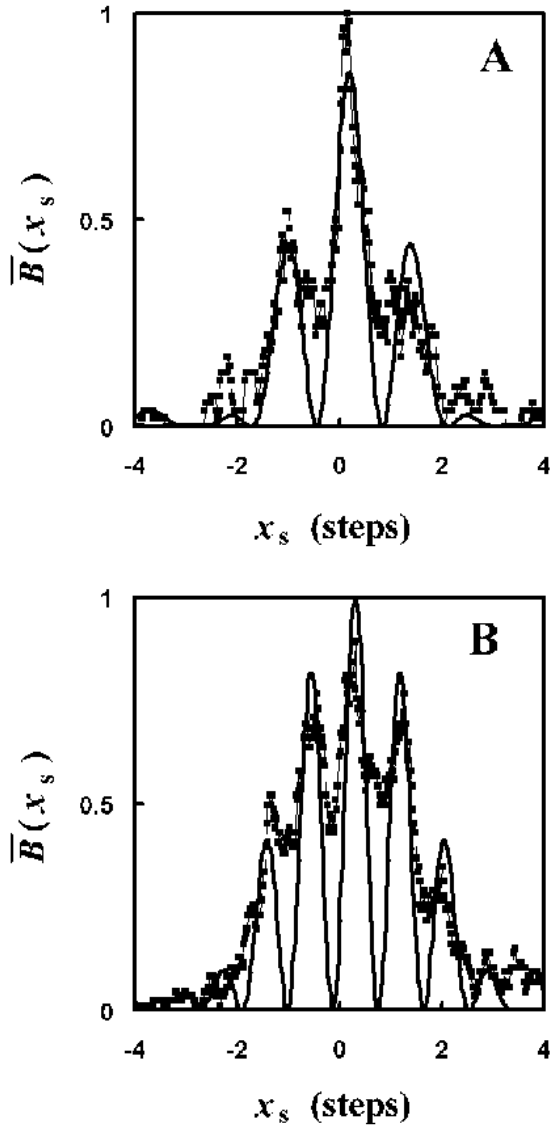


Figure 5: Plot of the double slit screen pattern at $L = 40$ steps and $W_{\text{in}} = 8$ steps. The A figure is with the slits placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step. The B figure is with the slits placed from 0.75 step to 1.25 steps and from -1.25 steps to -0.75 step. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation. The model produces the double slit interference pattern.

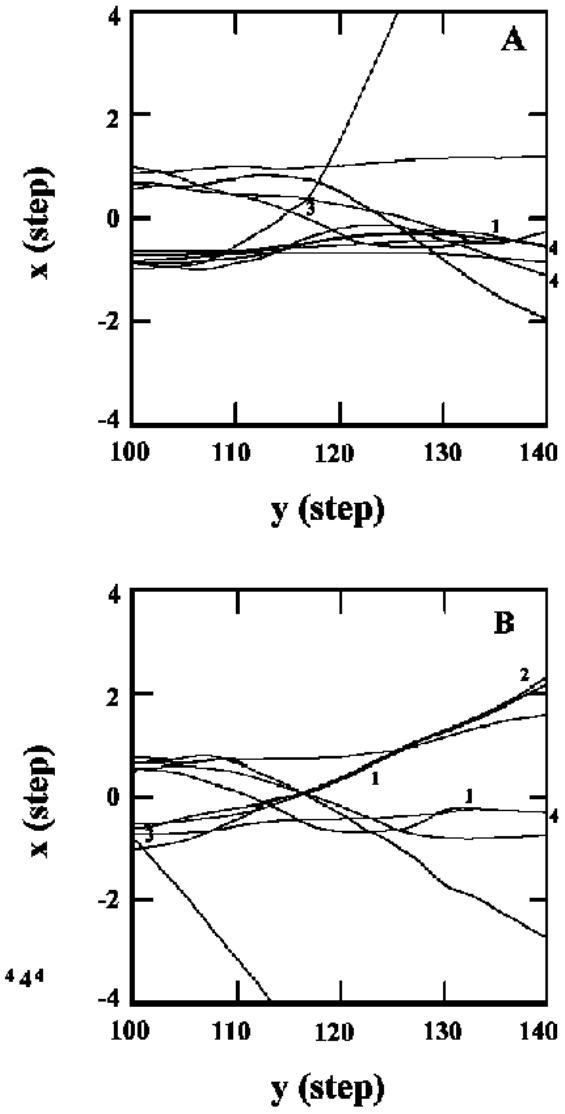


Figure 6: Traces of 10 consecutive photon paths between the mask and screen at two different intervals. The numbers mark the following occurrences: (1) One photon follows another and traces the same path. The following photon travels a longer path before path merging. (2) One photon follows another and traces a parallel and close path. (3) A photon experiences an abrupt change in θ as it passes close to another photon. These events were probably a result of the discrete nature of the simulation like a collision condition. (4) A photon from one slit follows another from the other slit. The leading photon determines the x_s and θ at the screen. The photon's path continues to change direction for a short distance after the mask.

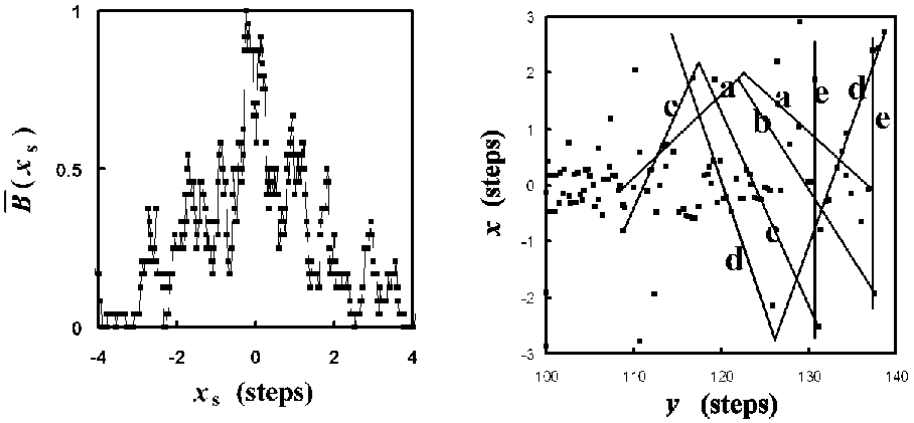


Figure 7: The right figure is a plot screen pattern of Young's experiment at $L = 40$ steps after the first mask and $W_{\text{in}} = 6$ steps. The filled squares are the data points. The thin line connects the data points. The Fresnel equation fit is poor. Therefore, the pattern is not a diffraction pattern. The right figure shows the distribution of photons from the first mask to the screen. The lines and the lower case letters are as in Fig. 1. Random photons through a first slit fail to produce a diffraction pattern that indicates incoherence. However, the position distribution shows coherence (see Fig. 1B).

The screen was removed and the second mask was placed at y= 140 steps. The second mask had two slits placed 0.5 step $\leq x \leq$ 1.0 step and at -1.0 step $\leq x \leq$ -0.5 step. The screen was placed at y= 180 steps ($L = 40$ steps). Figure 8 shows the resulting distribution pattern.

Although the statistics for Fig. 8 are poorer than previous screen interference patterns, inspection of Fig. 8 indicates an interference pattern was obtained. Figure 3(B and D) after 50 steps of calculation compared with Figure 3(A and C) after 30 steps of calculation indicates that the calculation errors are causing increased and noticeable scatter after 50 steps of calculation. Figure 8 is after 80 steps of calculation.

6 Laser

The initial, overall photon density in the previous sections was approximately uniform and incoherent. The photons in the distance simulation or the slit in the Young's simulation formed the coherent distribution. These coherent distributions resulted from an application of the model to an initial random distribution.

The popular model of a laser is that a seed photon in a medium stimulates the emission of more photons. Because the photons from a laser impinging

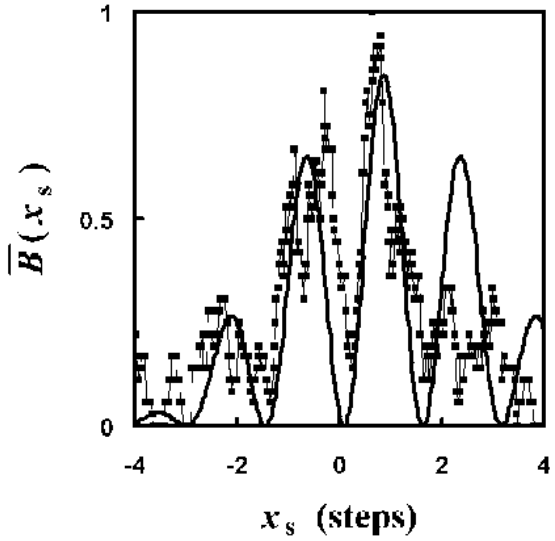


Figure 8: Plot of the double slit screen pattern at $L = 40$ steps from the second mask and 80 steps beyond the first mask. The slits were placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation. The position distribution after the first mask is coherent.

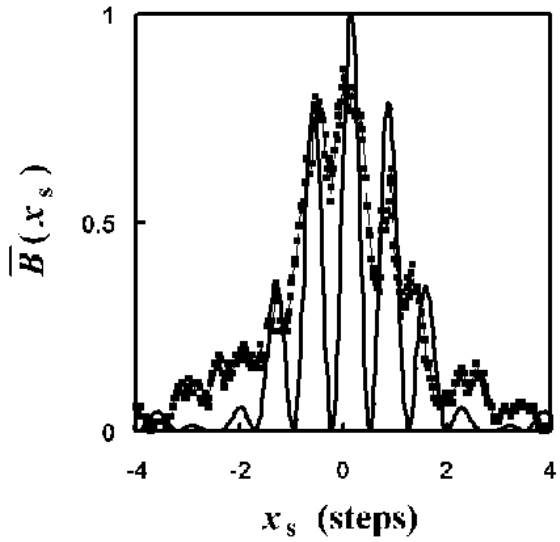


Figure 9: Plot of the pattern on a screen at $L = 40$ steps of a laser pulse input $W_{\text{in}} = 6$ steps through a double slit. The slits were placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation.

on a slit(s) produces a diffraction pattern, the laser simulation must produce the coherent distribution of photons. Because of the diversity of materials that produce laser emissions, the laser must form an ordered distribution within the light.

The Fourier transform of a triangular function is the sinc-type function. Another function with a Fourier transform of the sinc-type function is the rectangular function. If the slit acts as a Fourier transform on the stream of photons, a pulse pattern with high density and a low duty cycle may also produce the diffraction pattern.

A simple model of the simulation of the laser light is several photons followed by a delay between pulses. Figure 9 shows the result of passing a pulse through a double slit. The pulse is formed by positioning a photon randomly in half step x intervals and randomly within a 1.1 steps y interval. These pulses were three steps apart and $W_{\text{in}} = 6$ steps. Several other pulse configurations were tried and yielded a poorer fit. The parameters are unique. Also, the fit is inconsistent with observation of interference patterns for a d/b ratio of 3/1. That this is what lasers in general produce seems unlikely.

That the photons form lines with a set angle to $\vec{c}$ was noted in Fig. 1 and Fig. 7. Posit that a seed photon follows a free photon in the laser material. These photons form themselves at an angle noted in Fig. 1. The angles are

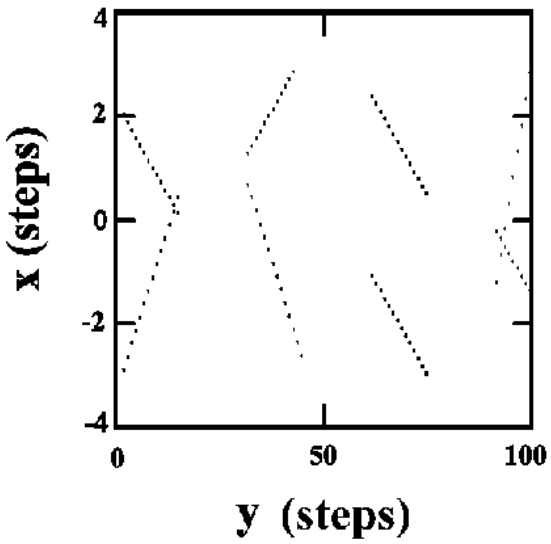


Figure 10: Plot of the position of photons between 0 step$\leq y \leq$ 100 steps and $W_{\rm in} = 6$ steps.

related to $N_{\rm h}$. These photon then exert a force along the angle to free weakly bound photons. Thus, a line of photons is formed. The lines are then emitted.

The experiment was to create two seed photons at $y = 0$ and randomly between -3 steps $\leq x \leq$ 3 steps. A line of 13 additional photons was introduced from each of the seed photons at one of the four angles, which was randomly chosen, progressing positively or negatively. Figure 10 depicts a plot of such a distribution at one interval. This model has the advantage of being dependent on $N_{\rm h}$ and the form of the distribution produced by distant travel and by the first slit in Young's experiment.

The photons were then directed to a mask at $y = 100$ with a double slit. The slits placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step.

The fit is consistent with observation of interference patterns for a d/b ratio of 3/1.

7 Afshar experiment

The Afshar experiment involves a wire grid and lenses in addition to a screen. Placing thin wires of one bin thickness at minima is possible in the present experiment. The $\bar{B}(x_{\rm s})$ values averaged 11 bins. There were bins with no photons near the minima.

The modeling of the photon interacting with matter and lenses is for a future paper. Therefore, the $x_{\rm s}$ and the angle $\phi_{\rm s}$ of $\vec{c}$ to the Y axis was also recorded

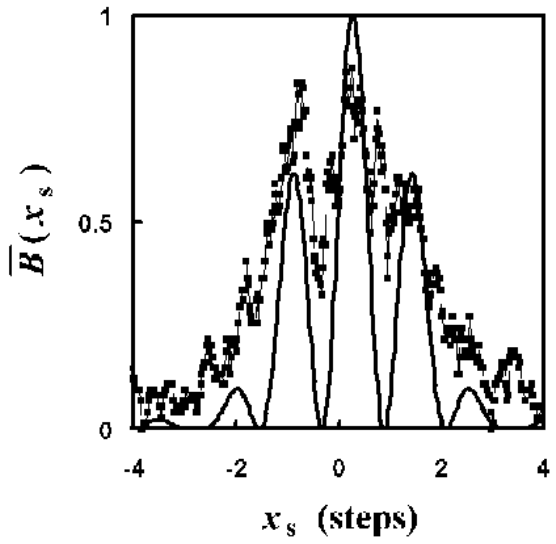


Figure 11: Plot of the pattern on a screen at $L = 40$ steps of a line laser input (see Fig. 10) $W_{\rm in} = 6$ steps through a double slit. The slits placed from 0.50 step to 1.00 step and from -1.00 step to -0.50 step. The filled squares are the data points, the thin line connects the data points, and the thick line marks the theoretical calculation.

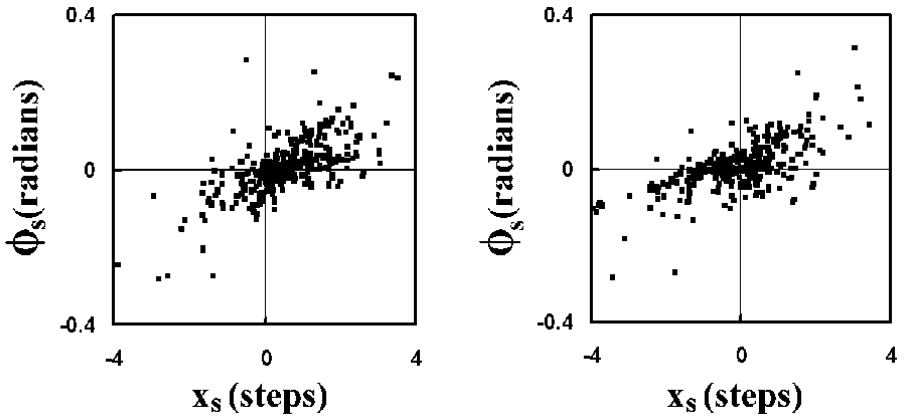


Figure 12: Plot of ϕ_s vs. x_s for the photons that passed through the Positive Slit (left) and the photons that passed through the Negative Slit (right). The groups of photons with -3 steps $\leq x_s \leq$ 3 steps to have a nearly linear distribution. A linear regression was done on the data of each of the groups. Photons existed outside this range. However, occurrences (1) and (2) of Fig. 6, which was considered an artifact of the simulation, caused errors. The distribution outside this range became non-linear. Over 86% of the recorded photons were in this range.

when a photon was counted at the screen.

Figure 12 shows a plot of ϕ_s vs. x_s for the photons that passed through the slit at 0.5 steps $\leq x_s \leq$ 1.0 steps (the "Positive Slit") and the photons that passed through the slit at -1.0 steps $\leq x_s \leq$ -0.5 steps (the "Negative Slit").

Figure 12 shows the groups of photons with -3 steps $\leq x_s \leq$ 3 steps to have a nearly linear distribution. A linear regression was done on the data of each of the groups. Photons existed outside this range. However, occurrences (1) and (2) of Fig. 6, which was considered an artifact of the simulation, caused errors. The distribution outside this range became non-linear. Over 86% of the recorded photons were in this range.

$$\phi_s = mx_s + b, \tag{14}$$

where m is the slope and b is the intercept of the linear regression.

Table 3 lists the resulting values of the linear regression equation for each of the data sets and the calculated ϕ_s at $x_s = 2$ steps and $x_s = 2$ steps. The majority of the photons impinge on the screen at angles that would cause a condensing lens to focus them at different points associated with the slits. Figure 6, occurrence (4) showed some photons from one slit follow another photon from the other slit and, therefore, were recorded with the ϕ_s as if from the wrong slit. This occurred for both slits. Therefore, the statistical effect would balance.

The majority of the photons impinge on the screen at angles that would

Table 3: The values of the constants in the linear regression Eq. 14.

slit	m rad. step^{-1}	b rad.	$\phi_s(x_s = 2$ steps) rad.	$\phi_s(x_s = -2$ steps) rad.
Positive Slit	0.0428	-0.0177	0.068	-0.10
Negative Slit	0.0348	0.00512	0.074	-0.065

cause a condensing lens to focus them at different points associated with the slits. The model produces the Afshar experiment.

8 Discussion

The constants were determined iteratively and with few significant figures. The solution presented may not be unique or optimal.

The $E = mc^2$ relation was used to derive Eq. (3) and (4). This suggests a way to relate measurable quantities to the constants $E = mc^2 = h\nu$. Further, Ψ is responsible for inertial mass. Thus, Ψ is a wave in a "real" physical entity.

The "wave" in quantum mechanics is a wave in the Ψ–field. The hod causes the wave and the wave directs the hod. The speed of the wave in the Ψ–field is much greater than c. Because the number of hods in a *moving* photon determines the wavelength of the Ψ–field wave, the photon causally interacts with other similar hods. Because the wave is a sine or cosine function, matter producing equal wavelength in the Ψ–field can "tune" into each other. This produces the interference pattern. Therefore, quantum entanglement may be a causal and connected observation.

This paper suggests the transverse and longitudinal position of photons undergo forces that may be treated as Fourier transforms with each encounter with more massive particles. The varying Ψ–field experienced by photons causes a Fourier transform on the distribution of photons. Therefore, the probability distribution of the position and movement of the large number of photons may be treated as in quantum mechanics.

The flow of photons through a volume with matter would produce a pattern of waves from each matter particle. Therefore, the Huygen's model of each point being a re-emitter of waves is justified if "each point" means each matter particle such as atoms.

Fourier mathematics assumes an infinite stream of particles obeying a given function for all time. Each encounter with other matter produces a different function. The mathematics of the Fourier transform includes the integration in both time and distance from $-\infty$ to $+\infty$. Therefore, observations made over a region or over a shorter interval allows for the uncertainty of waves. This non-uniformity of the time and distance of the particle stream distribution is Fourier transformed into the Heisenberg Uncertainty Principle (Tang 2007, see, for example, Section 2.9).

The change in the diffraction pattern upon the change in the width of the photon stream that the mask blocks (see Fig. 2) suggests these photons have an influence on the photons that go through the slit. This differs from the traditional wave model of diffraction. It also suggests the photon diffraction experiment is an experiment of quantum entanglement. Indeed, the photons blocked by the mask are non-local to the transmitted photons beyond the mask.

Bell's inequality includes the assumption of locality (Dürr,et al. 2009; Goldstein 2009). Because the present model is intrinsically nonlocal, it avoids Bell's inequality.

The calculation equations allow several negative feedback loops. For example, c is dependent on Ψ. If a photon is at a high Ψ region, c is high. This causes the photon to be faster then the photon producing the wave and move to a lower Ψ. The lower Ψ slows the photon to match the speed of the photon producing the wave. This mechanism exists in θ and v_{t}.

The present concept of "coherent" differs from the traditional model. The photons interact through the Ψ–field and tend toward lower Ψ. Coherence in the sense of interaction of photons is when the photons are maintained at a position and momentum relative to other photons through the feedback mechanisms. This occurs when a photon distribution causes a constant relative, moving minima. That is when $\cos(K_\beta\beta)/r < 1/r$. This also implies there are constant relative forbidden zones where $\cos(K_\beta\beta)/r > 1/r$ and $\Psi \approx \Psi_{\max}$. Thus, position and momentum are quantized.

Knowledge of the density of particles is insufficient to determine the Bohmian quantum potential as noted in Young's experiment. The structure of hods must also be known to provide the Ψ–field. For example, the Ψ–field wave caused by a photon structure of $N_{\mathrm{h}1}$ hods differs from the Ψ–field wave caused by a photon structure of $N_{\mathrm{h}2}$ hods where $N_{\mathrm{h}1} \neq N_{\mathrm{h}2}$.

Gravity is another manifestation of the Ψ–field–hod interaction. Moving particles produce "pilot waves" (gravity waves) in the Ψ–field. The wave–particle duality of the Copenhagen interpretation may be viewed as which of the two entities (Ψ–field or particle) predominates in an experiment.

For $\theta > \alpha$ the photon's tendency is to align its axis along streamlines of the Ψ–field and axes of other photons. Because of the symmetry of the photon, a converging Ψ–field such as from Sinks and matter and a diverging Ψ–field such as from Sources and galaxy clusters have the same angle changing effect on a photon.

Comparing the SPM and the present model yields the conclusions that the plenum is the Ψ–field and the ρ is the Ψ. A possible addition to the concept of the plenum is that the Ψ–field in the present model supports a transverse wave.

The SPM of the gravitational lens phenomena is that of gravitational attraction changing v_{t} and of $\dot{\theta}$ changing the direction of $\vec{c}$.

Conceptualizing the macroscopic scale of everyday experience, the galactic scale, and the quantum scale can now be done with a single concept as fractal cosmology suggests. The SPM view strengthens the deterministic philosophy. Therefore, the Schrödinger equation represents our lack of knowledge or inability to measure the initial and evolving parameters. The SPM solution for the

Bohmian trajectory(Goldstein 2009) is unique for each particle. The uncertainty of the measurement of position and momentum produces other possible trajectories.

Our measuring instruments measure particles that consist of hods and bound Ψ–field. That is, we measure only particles and their behavior. The properties of the Ψ–field are inferred from how the particles are guided by the Ψ–field and not by measurement of the Ψ–field itself.

Because the plenum is real and because the structure of the photon is consistent with the Michealson-Morley and diffraction experiments, there is no need to introduce action-at-a-distance, length contraction, time dilation, or a single big bang. Because of the unity of concepts between the big and the small, the SPM may be a Theory of Everything.

9 Conclusion

Newton's speculations, Democritus's speculations, the Bohm interpretation, and the fractal philosophy were combined with the cosmological Scalar Potential Model (SPM). The resulting model of photon structure and dynamics was tested by a toy computer experiment.

The simulation of light from a distance produced the diffraction pattern after passing through one and two slit masks. The screen patterns were determined to be diffraction patterns by fitting the pattern on a screen to the Fresnel equation. The distribution that was interpreted as coherent was formed by several lines of photons at angles to $\vec{c}$ consistent with the antenna pattern for the photons with the given $N_{\rm h}$. The photons impinging on the opaque portion of the mask were necessary in the calculation. Tracing the path of the photons showed the Fresnel pattern forms after a number of steps from the mask. Also, by varying the distance between the mask and the screen, the Fresnel pattern became the Fraunhofer pattern.

The simulation of Young's experiment showed randomly distributed photons were not coherent but became coherent after passing through one slit. The distribution of photons after the first slit resembled the line pattern of the light from a distance. This pattern was shown to be coherent after passing through a double slit. Young's experiment was duplicated.

The simulation of laser light examined two possible photon distributions. One considered the laser to be emitting pulses of random composition. A Fresnel fit was found. However, the number of minima was inconsistent with the physical structure of the mask. The second posited seed photons formed lines of photons at angles related the $N_{\rm h}$. This distribution was also fit by the Fresnel equation and the number of minima was consistent with the mask construction.

Because a model for photon interaction with lenses is lacking, the slit the photon passed through and the position and angle the photons strike the screen were recorded. The average difference of angle depending on the slit the photon passed through could produce the Afshar result. The model is consistent with the Afshar experiment.

Acknowledgments

I appreciate the financial support of Maynard Clark, Apollo Beach, Florida, while I was working on this project.

References

Afshar, S.S., Proceedings of SPIE 5866 (2005): 229-244. preprint http://www.arxiv.org/abs/quant-ph/0701027v1.

Baryshev, Y., Teerikorpi, P., 2004. Discovery of Cosmic Fractals. World Scientific Press.

Bertolami, O., Páramos, J., 2004. Clas. Quantum Gravity 21, 3309.

Descartes, R, 1644. Les Principes de la philosophie. Miller, V. R. and R. P., trans., 1983. Principles of Philosophy. Reidel. http://plato.stanford.edu/entries/descartes-physics/ .

Dürr, D., et al.., 2009. preprint http://arxiv.org/abs/0903.2601.

Eddington, A. S., Space, Time and Gravitation: An Outline of the General Relativity Theory (Cambridge University Press, London, UK).

Goldstein, S., 2009. preprint http://arxiv.org/abs/0907.2427.

Goldstein, S.,Tumulka, R., and Zanghi, N., 2009. preprint http://arxiv.org/abs/0912.2666.

Hodge, J.C., 2004. preprint http://arxiv.org/abs/astro-ph?0409765v1.pdf .

Hodge, J.C., 2006a. NewA 11, 344.

Hodge, J.C., 2006c. preprint http://arxiv.org/abs/astro-ph?0603140v1.pdf .

Hodge, J.C., 2006b. preprint http://arxiv.org/abs/astro-ph?/0611029v2.pdf .

Hodge, J.C., 2006c. preprint http://arxiv.org/abs/astro-ph?0611699v1.pdf .

Hodge, J.C., 2006c. preprint http://arxiv.org/PS_cache/astro-ph/pdf/0612/0612567v2.pdf .

Hodge, J.C., Theory of Everything: Scalar Potential Model of the big and the small(ISBN 9781469987361, available through Amazon.com, 2012).

Lauscher, O. and Reuter, M. 2005. Asymptotic Safety in Quantum Gravity. preprint http://arxiv.org/abs/hep-th?0511260v1.pdf .

Newton, I., Opticks based on the 1730 edition (Dover Publications, Inc., New York, 1952).

REFERENCES

Sartori, L., Understanding Relativity(Univ. of California Press, 1996).

Sommerfield, A., 1954. Optics, chaps 5 and 6. Acedemic Press, New York, NY.

Tang, K.T. 1999. Mathematical Methods for Engineers and Scientists 3, Berlin, Germany: Springer–Verlag.

von Flandern, T., 1998. Physics Letters A 250, 1.

Will, C.M., 2001. "The Confrontation between General Relativity and Experiment". Living Rev. Rel. 4: 4-107. http://www.arxiv.org/abs/ gr-qc?0103036.

Algebrizing friction: a brief look at the Metriplectic Formalism

Massimo Materassi (1)
Emanuele Tassi (2)

(1) Istituto dei Sistemi Complessi ISC-CNR, Sesto Fiorentino, Firenze, Italy
E-mail: massimo.materassi@fi.isc.cnr.it
(2) Centre de Physique Théorique, CNRS -Aix-Marseille Universités, Campus de Luminy, Marseille, France
E-mail: tassi@cpt.univ-mrs.fr

Abstract: The formulation of Action Principles in Physics, and the introduction of the Hamiltonian framework, reduced dynamics to bracket algebræ of observables. Such a framework has great potentialities, to understand the role of symmetries, or to give rise to the quantization rule of modern microscopic Physics.
Conservative systems are easily algebrized via the Hamiltonian dynamics: a conserved observable H generates the variation of any quantity f via the Poisson bracket $\{f,H\}$.
Recently, dissipative dynamical systems have been algebrized in the scheme presented here, referred to as *metriplectic framework*: the dynamics of an isolated system with dissipation is regarded as the sum of a Hamiltonian component, generated by H via a Poisson bracket algebra; plus dissipation terms, produced by a certain quantity S via a new symmetric bracket. This S is in involution with any other observable and is interpreted as the *entropy* of those degrees of freedom statistically encoded in friction.
In the present paper, the metriplectic framework is shown for two original "textbook" examples. Then, dissipative Magneto-Hydrodynamics (MHD), a theory of major use in many space physics and nuclear fusion applications, is reformulated in metriplectic terms.
Keywords: Dissipative systems, Hamiltonian systems, Magneto-Hydrodynamics.

1. Introduction

Hamiltonian systems play a key role in Physics, since the dynamics of elementary particles appear to be Hamiltonian. Hamiltonian systems are endowed with a bracket algebra (that of quantum commutators, or classically of Poisson brackets): such a scheme is of exceptional clarity in terms of symmetries [1], offering the opportunity of retrieving most of the information about the system without even trying to solve the equations of motion.
Despite their central role, Hamiltonian systems are far from covering the main part of real systems: indeed, Hamiltonian systems are intrinsically conservative and reversible, while, as soon as one zooms out from the level of elementary particles, the real world appears to be made of dissipative, irreversible processes [2]. In most real systems there are couplings bringing energy from processes at a certain time- or space-scale, treated deterministically, to processes evolving at much "smaller" and "faster" scales, to be treated statistically, as "noise". This is exactly what *friction* does, and this transfer appears to be *irreversible*.

A promising attempt of algebrizing the classical Physics of dissipation appears to be the *Metriplectic Formalism* (MF) exposed here [3, 4]. The MF applies to *closed systems with dissipation*, for which the energy conservation and entropy growth hold: the MF satisfies these two conditions [5]. The first important ingredient of the MF is the *metriplectic bracket* (MB):

$$\langle\langle f, g \rangle\rangle = \{f, g\} + (f, g),$$

where the first term $\{f,g\}$ is a Poisson bracket, while the term (f,g) is a *symmetric bracket*, bilinear and semi-definite. The total energy is represented by a Hamiltonian H which has *zero symmetric bracket with any quantity* (i.e. $(f,H) = 0$ for all f). The total entropy is mimicked by an observable S that has *zero Poisson bracket with any quantity* (i.e. $\{f,S\} = 0$ for all f). Then, a free energy F is defined as

$$F = H + \alpha S,$$

α being a coefficient that will disappear from the equations of motion, due to the suitable definition of (f,g); it coincides with minus the equilibrium temperature of the system (see below in the examples). The dynamics of any f reads:

$$\dot{f} = \langle\langle f, F \rangle\rangle = \{f, H\} + \alpha(f, S).$$

This dynamics conserves H and gives a monotonically varying (increasing) S. Metriplectic systems admit asymptotic equilibria (due to dissipation) in correspondence to extrema of F.
In this paper the MF is applied to some examples of isolated dissipative systems: two "textbook" examples and, more significantly, to visco-resistive magneto-hydrodynamics (MHD).

2. Two "textbook" examples

In order to illustrate how the MF works, two simple systems are considered.
The first one is a particle of mass m dragged by the conservative force of a potential V throughout a viscous medium. A viscous friction force, proportional to the minus velocity of the particle via a coefficient λ, converts its kinetic energy into internal energy U of the medium, with entropy S. The equations of motion of the system read:

$$\dot{\boldsymbol{x}} = \frac{\boldsymbol{p}}{m}, \quad \dot{\boldsymbol{p}} = -\nabla V - \lambda \frac{\boldsymbol{p}}{m}, \quad \dot{S} = \frac{\lambda p^2}{m^2 T}.$$

T is the temperature of the medium, simply defined as the derivative of *U* with respect to *S*. If the MB is defined as follows:

$$\dot{f} = \langle\langle f, F \rangle\rangle, \quad \langle\langle f, g \rangle\rangle = \{f, g\} + (f, g) \quad /$$
$$\{f, g\} = \frac{\partial f}{\partial \mathbf{x}} \cdot \frac{\partial g}{\partial \mathbf{p}} - \frac{\partial g}{\partial \mathbf{x}} \cdot \frac{\partial f}{\partial \mathbf{p}},$$
$$(f, g) = \Gamma^{ij} \frac{\partial f}{\partial \psi^i} \frac{\partial g}{\partial \psi^j} \quad / \quad \underline{\psi} = (\mathbf{x}, \mathbf{p}, S),$$
$$\Gamma = \alpha^{-1} \begin{pmatrix} (\nabla V)^2 \mathbf{1}_{3,3} - \nabla V \otimes \nabla V & 0 & 0 \\ 0 & \lambda T \mathbf{1}_{3,3} & -m^{-1} \lambda \mathbf{p} \\ 0 & -m^{-1} \lambda \mathbf{p}^{\mathrm{T}} & \frac{\lambda p^2}{m^2 T} \end{pmatrix},$$

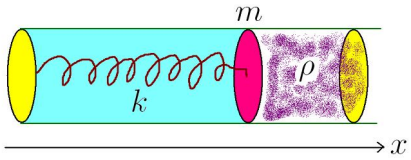
it is easy to show that these ODEs are given by the MB of $\mathbf{x}$, $\mathbf{p}$ and *S*, with a free energy *F* constructed as:

$$F(\mathbf{x}, \mathbf{p}, S) = H(\mathbf{x}, \mathbf{p}, S) + \alpha S,$$
$$H(\mathbf{x}, \mathbf{p}, S) = \frac{p^2}{2m} + V(\mathbf{x}) + U(S).$$

The matrix Γ is semi-definite with the same sign as α. The foregoing framework conserves *H* and increases *S*, driving the system to the asymptotic equilibrium:

$$\mathbf{p}_{eq} = 0, \quad \nabla V(\mathbf{x}_{eq}) = 0, \quad T_{eq} = -\alpha.$$

At the equilibrium the point particle stops at a stationary point of *V* once its kinetic energy has been fully dissipated into heat by friction.
The second rather simple example of metriplectic system is a piston of mass *m* and area *A*, running along a horizontal guide pushed by a spring of elastic constant *k*. It works against a viscous gas of pressure *P* and mass *M*. The system is depicted in the following Figure.



Piston moved by the spring of elastic constant *k* and mass *m*, working against a viscous gas of density ρ.

If ℓ is the rest-length of the spring, then the equations of motion of the system read:

$$\dot{x} = \frac{p}{m}, \quad \dot{p} = -PA - k(x - \ell) - \lambda \frac{p}{m}, \quad \dot{S} = \frac{\lambda p^2}{m^2 T}.$$

These equations of motion may be obtained out of a metriplectic scheme assigned as

$$\dot{f} = \langle\langle f, F \rangle\rangle, \quad \langle\langle f, g \rangle\rangle = \{f, g\} + (f, g) \quad /$$
$$\{f, g\} = \frac{\partial f}{\partial x}\frac{\partial g}{\partial p} - \frac{\partial g}{\partial x}\frac{\partial f}{\partial p}, \quad (f, g) = \Gamma^{ij}\frac{\partial f}{\partial \psi^i}\frac{\partial g}{\partial \psi^j} \quad / \quad \underline{\psi} = (x, p, S),$$
$$\Gamma = \alpha^{-1}\begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda T & -m^{-1}\lambda p \\ 0 & -m^{-1}\lambda p & \frac{\lambda p^2}{m^2 T} \end{pmatrix},$$

provided the following free energy is defined

$$F(x, p, S) = H(x, p, S) + \alpha S,$$
$$H(x, p, S) = \frac{p^2}{2m} + \frac{k}{2}(x - \ell)^2 + U(\rho(x), S).$$

Again, this Γ is semi-definite with the same sign as α. The asymptotic equilibrium of the foregoing *F* read

$$x_{eq} = \ell - \frac{PA}{k}, \quad p_{eq} = 0, \quad T_{eq} = -\alpha$$

(the temperature *T* is still defined as the derivative of *U* with respect to *S*): the piston stops where the spring equilibrates the gas pressure, its kinetic energy all dissipated by friction.

3. Dissipative MHD

Dissipative MHD is expected to describe many plasma processes, wherever its fundamental hypotheses apply to a highly conductive plasma interacting with its own magnetic field [6, 7]. Ideal MHD has already been cast into Hamiltonian formalism [8], here the metriplectic extension of the Poisson algebra, and the free energy extension of the Hamiltonian, is proposed to include dissipative effects [9].

The 3D visco-resistive MHD equations read:

$$\partial_t \boldsymbol{v} = -(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} - \tfrac{\nabla p}{\rho} - \tfrac{\nabla B^2}{2\rho} + \tfrac{(\boldsymbol{B}\cdot\nabla)}{\rho}\boldsymbol{B} - \nabla V_{grav} + \tfrac{\nabla\cdot\sigma}{\rho},$$
$$\partial_t \boldsymbol{B} = -(\boldsymbol{B}\cdot\nabla)\boldsymbol{v} - (\nabla\cdot\boldsymbol{v})\boldsymbol{B} - (\boldsymbol{v}\cdot\nabla)\boldsymbol{B} + \mu\nabla^2\boldsymbol{B},$$
$$\partial_t \rho = -\nabla\cdot(\rho\boldsymbol{v}),$$
$$\partial_t s = -(\boldsymbol{v}\cdot\nabla)s + \tfrac{\sigma:(\nabla\otimes\boldsymbol{v})}{\rho T} + \tfrac{\mu(\nabla\times\boldsymbol{B})^2}{\rho T} + \tfrac{\kappa\nabla^2 T}{\rho T}.$$

The MHD, defined on a 3D domain **D**, with suitable boundary conditions on $\partial\mathbf{D}$, is a *complete system* described by: plasma bulk velocity $\boldsymbol{v}$, magnetic induction $\boldsymbol{B}$, plasma density ρ and plasma mass-specific entropy s. In the foregoing field equations, p is plasma pressure, V_{grav} is an external gravitational potential; σ is plasma stress tensor, containing (linearly) the fluid viscosity coefficients η and ζ (see below), while μ is resistivity. κ is thermal conductivity, and T is temperature of the plasma. The system conserves the total energy:

$$H = \int_{\mathbf{D}} d^3x\left(\tfrac{\rho v^2}{2} + \tfrac{B^2}{2} + \rho V_{grav} + \rho U(\rho, s)\right),$$

which is the Hamiltonian, being U the mass-specific internal energy of the plasma. In the non-dissipative limit $\sigma = 0$, $\mu = 0$ and $\kappa = 0$, the whole physics is given by H and the following noncanonical Poisson brackets:

$$\{f,g\} = -\int_{\mathbf{D}} d^3x\Big[\tfrac{\delta f}{\delta\rho}\nabla\cdot\left(\tfrac{\delta g}{\delta\boldsymbol{v}}\right) + \tfrac{\delta g}{\delta\rho}\nabla\cdot\left(\tfrac{\delta f}{\delta\boldsymbol{v}}\right) +$$
$$-\tfrac{1}{\rho}(\nabla\times\boldsymbol{v})\cdot\left[\left(\tfrac{\delta f}{\delta\boldsymbol{v}}\right)\times\left(\tfrac{\delta g}{\delta\boldsymbol{v}}\right)\right] + \tfrac{1}{\rho}\left[\left(\tfrac{\delta f}{\delta\boldsymbol{v}}\right)\times\boldsymbol{B}\right]\cdot\left[\nabla\times\left(\tfrac{\delta g}{\delta\boldsymbol{B}}\right)\right] +$$
$$+\tfrac{\delta f}{\delta\boldsymbol{B}}\cdot\left[\nabla\times\left(\tfrac{\boldsymbol{B}}{\rho}\times\tfrac{\delta g}{\delta\boldsymbol{v}}\right)\right] + \tfrac{\nabla s}{\rho}\cdot\left(\tfrac{\delta f}{\delta s}\tfrac{\delta g}{\delta\boldsymbol{v}} - \tfrac{\delta g}{\delta s}\tfrac{\delta f}{\delta\boldsymbol{v}}\right)\Big]$$

(here $\delta f/\delta\varphi$ is the Fréchet derivative of the functional f with respect to the field φ). When dissipation is considered, the Hamiltonian must be extended to free energy adding a suitable entropic term:

$$S[\rho,s] = \int_{\mathbf{D}} \rho s\, d^3x, \quad \{f,S\} = 0 \quad \forall \quad f = f[\boldsymbol{v},\boldsymbol{B},\rho,s],$$
$$F[\boldsymbol{v},\boldsymbol{B},\rho,s] = H[\boldsymbol{v},\boldsymbol{B},\rho,s] + \alpha S[\rho,s];$$

the symmetric bracket to be used to form a complete MB, together with the Poisson bracket defined before, reads:

$$(f,g)=\alpha^{-1}\int_{\mathbf{D}} Td^3x\Big[\kappa T\nabla\Big(\tfrac{1}{\rho T}\tfrac{\delta f}{\delta s}\Big)\cdot\nabla\Big(\tfrac{1}{\rho T}\tfrac{\delta g}{\delta s}\Big)+$$
$$+\Lambda::\Big[\Big(\nabla\otimes\Big(\tfrac{1}{T}\tfrac{\delta f}{\delta \boldsymbol{v}}\Big)-\tfrac{1}{\rho T}\tfrac{\delta f}{\delta s}\nabla\otimes\boldsymbol{v}\Big)\otimes\Big(\nabla\otimes\Big(\tfrac{1}{T}\tfrac{\delta g}{\delta \boldsymbol{v}}\Big)-\tfrac{1}{\rho T}\tfrac{\delta g}{\delta s}\nabla\otimes\boldsymbol{v}\Big)\Big]+$$
$$+\Theta::\Big[\Big(\nabla\otimes\Big(\tfrac{\delta f}{\delta \boldsymbol{B}}\Big)-\tfrac{1}{\rho T}\tfrac{\delta f}{\delta s}\nabla\otimes\boldsymbol{B}\Big)\otimes\Big(\nabla\otimes\Big(\tfrac{\delta g}{\delta \boldsymbol{B}}\Big)-\tfrac{1}{\rho T}\tfrac{\delta g}{\delta s}\nabla\otimes\boldsymbol{B}\Big)\Big]\Big].$$

Note the strict analogy between the dissipative ***v***-terms and ***B***-terms, which are so alike because in the equations of motion dissipation terms appear as quadratic in the gradients of ***v*** and ***B***, respectively through the rank-4 tensors Λ and Θ (quadratic dissipation, see [9]):

$$\Lambda_{ikmn}=\eta\big(\delta_{ni}\delta_{mk}+\delta_{nl}\delta_{mi}-\tfrac{2}{3}\delta_{ik}\delta_{mn}\big)+\zeta\delta_{ik}\delta_{mn}\,,\quad \sigma=\Lambda:(\nabla\otimes\boldsymbol{v}),$$
$$\Theta_{ikmn}=\mu\varepsilon^{ikj}\varepsilon^{\ j}{}_{mn}\,.$$

Due to the symmetry properties of Λ and Θ, the symmetric bracket (f,g) just defined is semi-definite with the same sign of α; the functional gradient of H is a null mode of it. Finally, the quantities related to the space-time symmetries, generating the Galileo transformations

$$\boldsymbol{P}=\int_{\mathbf{D}}\rho\boldsymbol{v}d^3x,\quad \boldsymbol{L}=\int_{\mathbf{D}}\rho(\boldsymbol{x}\times\boldsymbol{v})d^3x,\quad \boldsymbol{G}=\int_{\mathbf{D}}\rho(\boldsymbol{x}-t\boldsymbol{v})d^3x$$

via the Poisson bracket algebra given in [8] and reported above, are conserved by the metriplectic dynamics:

$$\dot f=\{f,H[\boldsymbol{v},\boldsymbol{B},\rho,s]\}+\alpha(f,S[\rho,s]),$$

provided suitable boundary conditions are assigned to all the fields.
In the above Eulerian description of MHD, the bracket is noncanonical, depends on s, and the entropy S appears as a Casimir of the bracket which, by definition, belongs to the kernel of the co-symplectic form associated to the bracket [10], while in the "textbook" cases the Poisson bracket was canonical and was independent on the entropy-related variable S.
The free energy $F[\boldsymbol{v},\boldsymbol{B},\rho,S]$ constructed before is able to predict the asymptotic equilibrium state:

$$\boldsymbol{v}_{eq}=0,\quad \boldsymbol{B}_{eq}=0,\quad T_{eq}=-\alpha,\quad p_{eq}+\rho_{eq}V_{grav}=\rho_{eq}(Ts-U)_{eq}.$$

Such an equilibrium configuration has zero bulk velocity and magnetization, while pressure and gravity equilibrates the thermodynamic free energy of the gas.

4. Conclusions

In metriplectic formalism friction forces, acting within isolated systems, are algebrized. The dissipative terms in the equations of motion are given by a suitable symmetric, semi-definite bracket of the variables with the entropy of the degrees of freedom to which friction drains energy.

Two simple "textbook" examples are reported: the point particle moving through a viscous medium; a piston, moved by a spring against a viscous gas in a rigid cylinder. In both the examples the evolution is generated via the metriplectic bracket with the free energy $F = H + \alpha S$, where H is the conserved Hamiltonian and S is the monotonically growing entropy. α appears to coincide with the equilibrium temperature.

The same formalism is then applied to an isolated magnetized plasma, represented by the dissipative (i.e. viscous and resistive) MHD with suitable boundary conditions. A Hamiltonian scheme already exists for the non-dissipative limit; furthermore, the full MF had been introduced for the neutral fluid version. In this paper, we report the extension of the latter formalisms to include the magnetic forces and the dissipation due to Joule Effect [9]. The "macroscopic" level of plasma physics is described by the fluid variable $\boldsymbol{v}$, but a "microscopic" level exists too, encoded effectively in the thermodynamical field s. The energy attributed to the macroscopic degrees of freedom $\boldsymbol{v}$ is passed to the microscopic ones by friction, while the electric dissipation of Joule Effect consumes the energy pertaining to the magnetic degrees of freedom $\boldsymbol{B}$. Notice that the metriplectic formulation for dissipative MHD that we found, does not require $\mathrm{div}.\boldsymbol{B} = 0$.

Dissipative MHD is mathematically much more complicated than the two "textbook" examples, nevertheless its essence is rigorously the same: the MF algebraically generates asymptotically stable motions for closed systems. At the equilibrium, mechanical and electromagnetic energies are turned into internal energy of the microscopic degrees of freedom: the asymptotic equilibria found here for the three examples are essentially entropic deaths.

Let's conclude with few more observations.

MF is a deterministic description, but it must be possible to obtain it as an effective representation of a scenario where the superposition of the Hamiltonian and the entropic motion mirrors the Physics of a deterministic Hamiltonian system under the action of noise [8].

The appearance of MF offers potentially great chances because it drives the algebraic Physics out of the realm of Hamiltonian systems: many interesting processes in nature (as the apparent self-organization of space physics systems [12], not to mention biological or learning processes) are not expected to be even conceptually Hamiltonian. It is very stimulating to imagine dealing with algebraic formalisms describing them. MF, however, is not able to compound such processes, because it pertains to *complete*, i.e. closed, systems, while the processes just mentioned take place in open ones. Adapting MF to open systems will then be a necessary step to face this challenge.

Before concluding, let's underline again the dynamical role of entropy in MF: entropy may be interpreted as an information theory quantity [13, 14], and here we find information directly included in the algebraic dynamics. Furthermore: irreversible biophysical processes appear to have something in common with learning processes [15], i.e. processes in which the information is constructed or degraded, and having a formalism where "information" is an essential function appears to offer hopes in this branch.

References

1. L. D. Landau, E. M. Lifshitz, Mechanics. Course of Theoretical Physics.Vol.1, Butterworth-Heinemann, 1982.
2. B. Misra, I. Prigogine, M. Courbage, From deterministic dynamics to probabilistic description, Physica A, vol. 98, 1-26, 1979.
3. P. J. Morrison, Some Observations Regarding Brackets and Dissipation, Center for Pure and Applied Mathematics Report PAM--228, University of California, Berkeley (1984).
4. P. J. Morrison, Thoughts on brackets and dissipation: old and new, Journal of Physics Conference Series, 169, 012006 (2009).
5. P. J. Morrison, A paradigm for joined Hamiltonian and dissipative systems, Physics D, vol. 18, 410-419, 1986.
6. A. Raichoudhuri, The Physics of Fluids and Plasmas – an introduction for astrophysicists, Cambridge University Press, 1998.
7. D. Biskamp, Nonlinear Magnetohydrodynamics, Cambridge University Press, 1993.
8. P. J. Morrison, J. M. Greene, Noncanonical Hamiltonian Density Formulation of Hydrodynamics and Ideal Magnetohydrodynamics, Physical Review Letters, vol. 45, 10 (1980).
9. M. Materassi, E. Tassi, Metriplectic Framwork for the Dissipative Magneto-Hydrodynamics, submitted to Physica D.
10. P. J. Morrison, Hamiltonian description of the ideal fluid, Reviews of Modern Physics, Vol. 70, No. 2, 467-521, April 1998.
11. T. D. Frank, T. D., Nonlinear Fokker-Planck Equations, Springer-Verlag Berlin-Heidelberg, 2005.
12. T. Chang, Self-organized criticality, multi-fractal spectra, sporadic localized reconnections and intermittent turbulence in the magnetotail, Phys. Plasmas, 6, 4137-4145, 1999.
13. E. T. Janes, Information Theory and Statistical Mechanics, Phys. Rev., 106, 4, 620-630, 1957.
14. E. T. Janes, Information Theory and Statistical Mechanics II, Phys. Rev., 108, 2, 171-190, 1957.
15. G. Careri, La fisica della vita (Physics of Life), Sapere, Agosto 2002.

Antigravity in AFT

Diego Marin*

July 25, 2012

Abstract

We show how antigravity effects emerge from arrangement field theory. AFT is a proposal for an unifying theory which joins gravity with gauge fields by using the Lie group $E6$ or $Sp(6)$. Details of theory have been exposed in the ArXiv papers 1206.3663 and 1206.5665 (2012).

*dmarin.math@gmail.com

Contents

1 Introduction

Arrangement field theory is a quantum theory defined by means of probabilistic spin-networks. These are spin-networks where the existence of an edges is regulated by a quantum amplitude. AFT is a proposal for an unifying theory which joins gravity with gauge fields. See [1] and [2] for details. The unifying group is $U(6, \mathbf{H})$ for the euclidean theory ($Sp(6)$ in other notation), while it is $E6$ for the lorentzian theory. The unifying group contains three indistinguishable copies of gauge fields, mixed by gravitational field. Moreover, commutators between gravitational and gauge fields are non null and give new terms for the Einstein equations. In what follows we focus on the term which mixes gravity with electromagnetism, showing that its contribution to Einstein equations could generate antigravity. In the end we verify that new interactions don't affect the making of nucleus and nucleons.

2 Antigravity

The term which mixes gravity with electromagnetism is given by space-time integration of the following expression:

$$-\frac{1}{4} f^{(G)(EM1)(EM2)} A_{\mu}^{(G)} A_{\nu}^{(EM1)} \left(F^{(EM2)\mu\nu} + \alpha f^{(EM3)(EM1)(EM2)} A^{(EM3)\mu} A^{(EM1)\nu} \right) \tag{1}$$

Remember that AFT includes three indistinguishable electro-magnetic fields, with non-trivial commutators. In this way $A^{(G)}$ is the gravitational gauge field, $A^{(EMn)}$ is the n-th electromagnetic field and α is the fine structure constant. In the realistic case of null torsion, the gravitational gauge field can be rewritten in function of the tetrad field:

$$A_{\mu}^{(G)bc} = \frac{1}{2} e^{\nu[b} \partial_{[\mu} e_{\nu]}^{c]} + \frac{1}{4} e_{\mu d} e^{\nu b} e^{\sigma c} \partial_{[\sigma} e_{\nu]}^{d}$$

From now we take a low energy limit so defined: $e_{ii} = 1$ with $i = 1, 2, 3$, $e_{00} = \theta(x)$ and $\partial_0 \theta(x) = 0$. Varying with respect to e we obtain:

$$\frac{\delta A_\mu^{(G)bc}}{\partial e_\tau^s} = \frac{1}{2} e^{\nu[b} \delta_s^{c]} \delta_{[\nu}^\tau \partial_{\mu]} + \frac{1}{4} e_{\mu s} e^{\nu b} e^{\sigma c} \delta_{[\nu}^\tau \partial_{\sigma]}$$

$$\frac{\delta A_\mu^{(G)bc}}{\partial g_{\omega\tau}} = 2e^{\omega s} \frac{\delta A_\mu^{(G)bc}}{\partial e_\tau^s} = e^{\omega[c} e^{b]\nu} \delta_{[\nu}^\tau \partial_{\mu]} + \frac{1}{2} \delta_\mu^\omega e^{\nu b} e^{\sigma c} \delta_{[\nu}^\tau \partial_{\sigma]}$$

The component with $c = \omega = \tau = 0$ and $b \neq 0$ results:

$$\frac{\delta A_\mu^{(G)b0}}{\partial g_{00}} = -\theta^{-1} \delta_\mu^0 \partial_b - \frac{1}{2} \theta^{-1} \delta_\mu^0 \partial_b = -\frac{3}{2\theta} \delta_\mu^0 \partial_b$$

$$A^{(EM)\rho} A_\rho^{(EM)} A^{(EM)\mu} \frac{\delta A_\mu^{(G)b0}}{\partial g_{00}} = \frac{3}{2\theta} \partial_b A^{(EM)0} A^{(EM)\rho} A_\rho^{(EM)}$$

The minus sign has disappeared because we have reversed the derivative. The quartic term in (1) becomes:

$$-\frac{\alpha}{4} f^b \frac{3}{2\theta} \partial_b A^{(EM)0} A^{(EM)\rho} A_\rho^{(EM)} = -\partial_b f^b \frac{3\alpha}{8\theta} V(\theta^2 V^2 - A^2)$$

$$f^b = \sum_{cade} f^{(bo)ca} f^{dea} \approx 4 \frac{x^b}{r}.$$

Here we have indicated with V the electric potential and with A the magnetic vector potential. The sum inside f is over the three electromagnetic fields.

It's so clear that varying the complete action with respect to $g_{\mu\nu}$ we obtain a new term for Einstein equations. In the Newtonian limit we can substitute $g_{00} = -(1 - 2\phi)$ and $R_{00} - (1/2) R g_{00} = \nabla^2 \phi$ where ϕ is the newtonian potential. Hence:

$$\begin{aligned} 2\nabla^2 \phi &\approx 8\pi T^{00} = 8\pi \frac{-2}{\sqrt{-g}} \frac{\delta \sqrt{-g} L_{matt}}{\delta g_{00}} \\ &\approx \partial_b \frac{x^b}{r} 24\pi\alpha V(\theta V^2 - \theta^{-1} A^2) \end{aligned} \tag{2}$$

For radial potential we have

$$\partial_b \phi = \frac{x^b}{r} \partial_r \phi.$$

In such case

$$C_G = \partial_r \phi \approx 12\pi\alpha V(\theta V^2 - \theta^{-1} A^2)$$

Now we insert the appropriate universal constants and approximate θ with 1:

$$C_G \approx 12\pi\alpha \frac{(G\varepsilon_0)^{3/2}}{c^4 L_p} V(V^2 - c^2 A^2) = kV(V^2 - c^2 A^2) \tag{3}$$

Here L_p is the Planck length, equal to $\sqrt{\hbar G/c^3}$. The multiplicative constant is

$$k = \frac{12\pi}{137} \cdot \frac{(6,67 \cdot 10^{-11} \cdot 8,85 \cdot 10^{-12})^{3/2}}{(3 \cdot 10^8)^4 \cdot (1,62 \cdot 10^{-35})} = 30,27 \cdot 10^{-33} \left(\frac{C^3 s^4}{K g^3 m^5} \right).$$

This means that for having a weight variation (on Earth) of about 10% ($\Delta C_G = 1$) we need an electrical potential of 10^{11} Volts. These are 100 billions of Volts. For $V = Q/r$ and $A = 0$ we have:

$$C_G = \frac{k}{(4\pi\varepsilon_0)^3} \cdot \frac{Q^3}{r^3} = 2,198 \cdot 10^{-2} \left(\frac{m^4}{s^2 C^3} \right) \frac{Q^3}{r^3}$$

Note that the sign of C_G is the sign of Q and then we obtain antigravity for negative Q. We associate to this interaction an equivalent mass m, substituting $C_G = Gm/r^2$. We have

$$m = \frac{k}{G} V^3 r^2 = \frac{k}{G(4\pi\varepsilon_0)^3} \frac{Q^3}{r} = 3,293 \cdot 10^8 \left(\frac{K g\, m}{C^3} \right) \frac{Q^3}{r}$$

which is a negative mass for negative Q. Negative mass implies negative energy via the relation $E = mc^2$. Intuitively, if we search a similar relation for gravi-magnetic field (which is $\nabla \times (g^{0i})$, $i = 1, 2, 3$), we should find the same formula (3) with an exchange between V and cA.

We calculate now at what distance the gravitational attraction between two protons is equal to their electromagnetic repulsion.

$$G\frac{m^2}{r^2} = \frac{k^2}{G^2(4\pi\varepsilon_0)^6}\frac{Q_p^6}{r^4} = \frac{1}{4\pi\varepsilon_0}\frac{Q_p^2}{r^2}$$

$$\frac{k^2 Q_p^4}{G^2(4\pi\varepsilon_0)^5} = r^2$$

$$\Longrightarrow r^2 = 79,49 \cdot 10^{-70} m^2 \Longrightarrow r = 8,916 \cdot 10^{-35} m = 5,516\, L_p$$

Note that we are 20 orders of magnitude under the range of strong force and 23 orders of magnitude under the range of weak force. In this way the gravitational force doesn't affect the making of nucleus and nucleons.

3 Conclusion

We have seen that a potential of 10^{11} Volts can induce relevant gravitational effects. They are too many for notice variations in the experiments with particles accelerators. However they sit at the border of our technological capabilities. The possibility to rule gravitation is very attractive and constitutes a good reason for try experiments with high electric potentials. Such experiments can be connected to the work of Nikola Tesla and can also be a good test for the arrangement field theory.

References

[1] Marin, D.: The arrangement field theory (AFT). ArXiv: 1206.3663 (2012).

[2] Marin, D.: The arrangement field theory (AFT) - PART. 2 -. Arxiv: 1206.5665 (2012)

[3] Barton, C., H., Sudbery A.: Magic squares and matrix models of Lie algebras. ArXiv: 0203010 (2002).

[4] De Leo, S., Ducati, G.: Quaternionic differential operators. Journal of Mathematical Physics, Volume 42, pp.2236-2265. ArXiv: math-ph/0005023 (2001).

[5] Zhang, F.: Quaternions and matrices of quaternions. Linear algebra and its applications, Volume 251 (1997), pp.21-57. Part of this paper was presented at the AMS-MAA joint meeting, San Antonio, January 1993, under the title "Everything about the matrices of quatemions".

[6] Farenick, D., R., Pidkowich, B., A., F.: The spectral theorem in quaternions. Linear algebra and its applications, Volume 371 (2003), pp.75102.

[7] Hartle, J. B., Hawking, S. W.: Wave function of the universe. Physical Review D (Particles and Fields), Volume 28, Issue 12, 15 December 1983, pp.2960-2975.

[8] Minkowski, H., Die Grundgleichungen fur die elektromagnetischen Vorgange in bewegten Korpern. Nachrichten von der Gesellschaft der Wissenschaften zu Gottingen, Mathematisch-Physikalische Klasse, 53-111 (1908).

[9] LaFave, N.J., A Step Toward Pregeometry I.: Ponzano-Regge Spin Networks and the Origin of Spacetime Structure in Four Dimensions. ArXiv:gr-qc/9310036 (1993).

[10] Reisenberger, M., Rovelli, C., 'Sum over surfaces' form of loop quantum gravity. Phys. Rev. D, 56, 3490-3508 (1997).

[11] Engle, G., Pereira, R., Rovelli, C., Livine, E., LQG vertex with finite Immirzi parameter. Nucl. Phys. B, 799, 136-149 (2008).

[12] Banks, T., Fischler, W., Shenker, S.H., Susskind, L., M Theory As A Matrix Model: A Conjecture. Phys. Rev. D, 55, 5112-5128 (1997). Available at URL http://arxiv.org/abs/hep-th/9610043 as last accessed on May 19, 2012.

[13] Garrett Lisi, A., An Exceptionally Simple Theory of Everything. ArXiv:0711.0770 (2007). Available at URL http://arxiv.org/abs/0711.0770 as last accessed on March 29, 2012.

[14] Nastase, H., Introduction to Supergravity ArXiv:1112.3502 (2011). Available at URL http://arxiv.org/abs/1112.3502 as last accessed on March 29, 2012.

[15] Penrose, R., On the Nature of Quantum Geometry. Originally appeared in Wheeler. J.H., Magic Without Magic, edited by J. Klauder, Freeman, San Francisco, 333-354 (1972).

The Theory of existences

A.K.Balamuruhan[1]

http://www.akbmurugan.com

Abstract: Some new concepts on the perception of physical existences, based on new interpretations of quantum mechanical wave functions are presented. These new concepts remove the earlier imbalance in the notion towards the physical existences. A new understanding about Gravity is also presented. This understanding explains why gravity does not have a force-carrying particle. These concepts also lead to decipher the cosmological concepts, dark matter and dark energy. Further, It is found that time can exist even before the "big bang".

Preface

It is true that the existence of the physical universe is real. If asked whether the existence of the physical universe is positive, negative or both, one may reply that it is positive. However, every physical existence in the universe is composed of two components, of which one is a positive existence and the other is a negative existence. This may sound startling. However, after reading the reasoning discussed below, one may get convinced. Moreover, this understanding leads to a clue on the cosmological observations, dark matter and dark energy.

One interprets negative quantities mostly in a simple sense. For example, by declaring a distance to be negative, one means that its direction is opposite to that of a distance that is taken to be positive. There is nothing fundamentally different between these positive and negative distances. Similarly, a negative time would mean the past with reference to a point of time, if positive time means future. Arithmetic sum of these positive and negative quantities simply means a shift in the corresponding ray of distance, time or any other quantity concerned. This indicates the obsession with existence that it is eternal and concrete while its origin is not understood. One normally takes the existence of everything physical, for granted, to be positive. There is a problem in this assumption. If the physical existences are positive, where did they come from? If they came from nothing, what happen to the conservation? This is an imbalance or asymmetry in the common notion of existence. I present my solution to this problem, below.

The norm of a complex quantity, $x + iy$ is represented as $x^2 + y^2$. The square of real part is x^2 and the square of imaginary part is $(iy)^2 = -y^2$. However, we expect something 'positive definite' as the norm of

1 Author's email: akbmuruhan@gmail.com

the complex quantity. The definition of the norm as the product of itself with its complex conjugate satisfies this expectation. This change of the sign of the square of the imaginary part in the norm is factitious in the case of quantum wave functions and it conceals many facts.

Quantum Wave Functions

The norm of a wave function represents the probability of existence of the system that it represents. However, why should the norm be obtained by the combination of the absolute squares of the real and imaginary parts? I posit that the squares of the two parts should be considered separately to interpret the existences and that the two parts describe independently, the existences of two components of the system that the wave function represents. The squares of the real and imaginary parts are positive and negative respectively. This means that the existences that they represent are positive and negative existences. Positive and negative existences mean that they are opposite to each other in all aspects, like the mass, space, charge etc attributed to them. This concept should be understood carefully. This is very fundamental to the other discussions in the article. One would commonly think of positive existence as a normal existence and a negative existence as something weird. However, a new picture is presented here. According to this, both the existences are equivalent, however, not the same; neither of these two existences is more privileged by any means than the other is. From the concepts introduced here, one cannot identify any priority given to either of the existences. Any privilege identified with the positive existence could be because of a biased thought. In fact, the two existences are similar. One does *not* have to attribute the positive existence that is mentioned here to some *normal* existence and the negative existence that is mentioned here to something *imaginary*. Both are equivalent. In other words, **the existences are bipolar and the two polarities of existences are equivalent**, in contrast to the conventional notion of mono-polar existences. The advantages exclusive to this bipolar-existence interpretation are discussed next.

Merits of the interpretation

There are many advantages from this interpretation, some of which are discussed in the following.

Understanding spin

The conventional interpretation says that spin is an intrinsic property of elementary particles [1], which comes as an outcome of quantum mechanical analysis. However, it does not give any physical interpretation for spin and warns that one cannot imagine spin as the revolution of elementary particle. Neither is it advocated here, to interpret it so. A new interpretation is presented here that according to the bipolar-existence picture, the spin is the time dependent cycling between the two polarities of the existence-states. This interpretation also leads to an interesting interpretation of time that is discussed later in the article. The spin of, say, an s-electron is the time dependent cycling of its existence-states with a period of π/ω. In general, **if $\psi(r, t)$ is the wave function of a particle spinning in a particular**

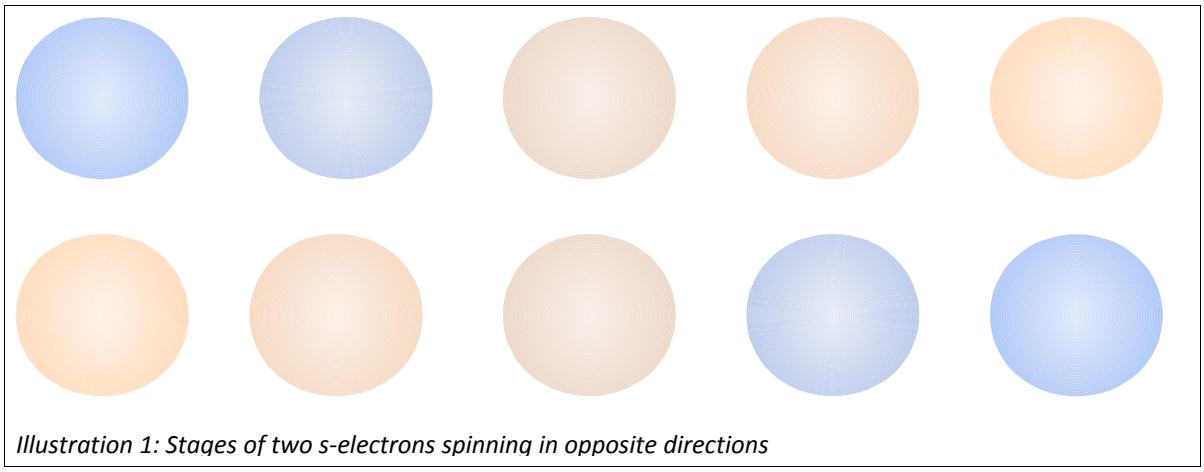


Illustration 1: Stages of two s-electrons spinning in opposite directions

direction, then, ψ(r, t)exp(iπ/2) is the wave function of an identical particle spinning in the opposite direction. Illustration 1 shows the different stages of two s-electrons spinning in opposite directions. The two colors, light blue and orange represent the two existence-states. The gradual change of colors in the illustration represents the gradual transformation of the states from purely positive to purely negative state and vice versa, the intermediate ones being mixtures of both states. It is to be remembered that either of the two existences can be taken as either positive or negative; there is no preference for one polarity over the other. Thus, a negative mass, for example, is present all over the universe, in contrast to the normal anticipation that it could be in some remote space. In fact, there is no preference in any respect, including the abundance, for positive or negative existence over the other. Both are equivalently present.

The old interpretation of wave functions allows the arithmetic summation of the squares of the two parts of the wave function, of course, after making both as positive. As a result, we get monotonously time independent pictures of, for example, the s-orbitals and p-orbitals as static, spherical and ring-like shells. On the other hand, since we deal the two parts separately in the bipolar picture presented here, we are able to see the dynamism called as spin.

Angular motion of electrons in atoms

Similar interpretation on a p-electron with a non-zero value for the quantum number 'm' gives rise to a more interesting scene. The electron in an atom, with a non-zero value for ‘m’ has a magnetic moment as if it has an angular motion. However, our old quantum mechanical interpretations impose that orbital motion also, like the spin, is an inherent property of the electron, nothing is moving in the classical sense and one should not interpret an angular motion of the electron. However, one logically expects an angular motion since the electron is having a magnetic moment. According to the bipolar interpretation, the angular motion is evident, as shown by the factors $\cos^2(m\phi+\omega t)$ and $-\sin^2(m\phi+\omega t)$ in the squares of the real and imaginary parts of the wave function of a p-electron. Illustration 2 shows a few stages of two p-electrons spinning in opposite directions and revolving in the same direction because of the same 'm' value. The revolution appears to be a consequential effect of spinning.

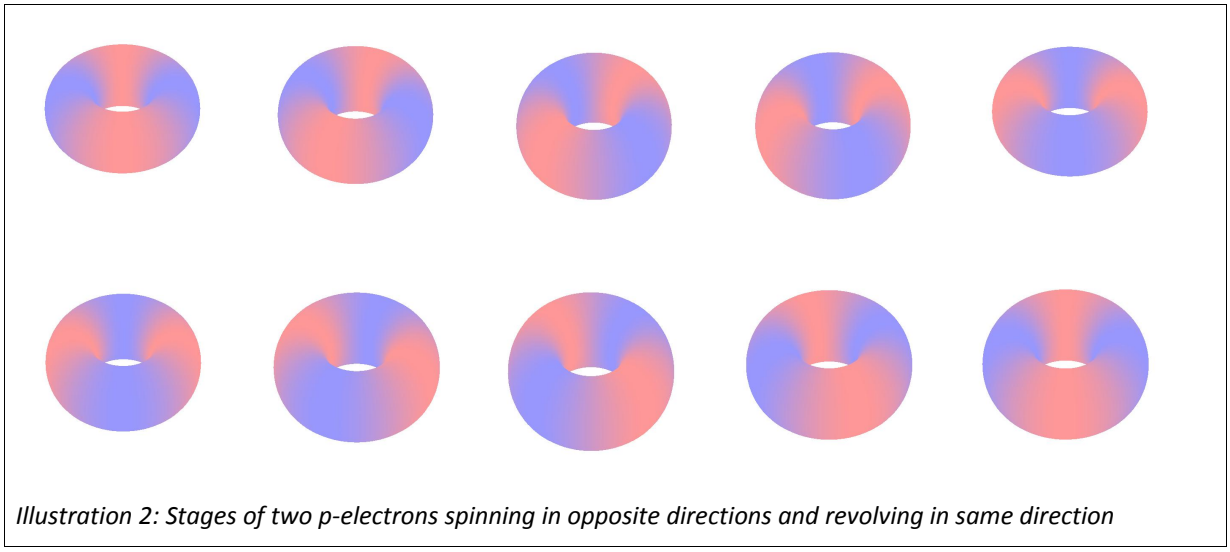


Illustration 2: Stages of two p-electrons spinning in opposite directions and revolving in same direction

From this simple case, where the spin of a p-electron manifests as a physical motion, I deduce that all the physical motions including macroscopic motions are the results of coordinated spins (polarity transformations) of all the particles involved in that motion.

In the old picture, where the norm is taken as the product of the wave function with its complex conjugate, these factors are summed up with suitable change of sign to get unity independent of time and hence, these physical interpretations are missed.

Pauli Exclusion Principle

According to the bipolar-existence picture, an electron is always in a state that is a combination of the two mutually opposite existence-states. Whatever the state this electron is in, the state complementary to it is available for another identical electron to occupy. However, its own state cannot be occupied by the second electron. This understanding provides the insight into the Pauli Exclusion Principle, which states that no two identical fermions can occupy the same quantum state simultaneously.

The compliance of this picture with the more rigorous statement of the Pauli exclusion principle that the total of the wave functions of two identical fermions is anti-symmetric with respect to the exchange of the particles, can be verified as follows. According to the bipolar-existence, the flip of spin is achieved by multiplying the wave function by the phase factor $\exp(i\pi/2)$. The exchange of particles in Pauli Exclusion Principle involves flipping of the spins of the two particles considered. This dual flipping amounts to multiplying the wave function by the factor $\exp(i\pi)$. Multiplying by $\exp(i\pi)$ causes rotation of the wave function in the complex plane through π, which means that the resulting wave function is anti-symmetric with respect to the original function. Thus, the bipolar-existence picture and the Pauli Exclusion Principle comply with each other.

Source of the creations

There are both physical and subtle creations in the world. Physical matter is an example of physical existence. Mind is an example of subtle existence. Creation of physical existence is discussed in this article. Is the source of all creations some 'nothing'? How can the source of the universe be nothing while something from nothing means violation of conservation? **I propose that the source of all creations is a subtle thing, it contains all the creations within it in a subtle form and all the physical creations are brought in to physical existence from this subtle source during the time of creation.** This subtle source is found to correspond to the 'Mayai' or Maya described in Indian scripture. My interpretation of wave functions discussed above avoids the violation of conservation during creation, since it tells that every physical creation is composed of two mutually opposite existences. These two opposite existences together can be assumed to amount to null since their polarities are opposite. *However, these two opposite existences are REAL although their source of creation is subtle.* This idea is discussed, next.

The origin of universe

Before “the big bang”, there was no space, mass etc, in short, no physical universe. However, the subtle source always exists. The 'big bang' marks the start of creation from the subtle source. The subtle source 'does not have' any physical dimension since these dimensions are the properties of its creations. The subtle source gives birth to space, particles (mass charge, etc), etc. and **exists** with the creations. This means that the subtle source is spreading with the creations created from it. This can be visualized as analogous to expanding water gel crystals. This is not an exact analogy since the crystal has some initial physical dimensions. There is a speck of subtle source at the center of every physical particle or a region of space created from it. The order of creation may be that the subtle source gives birth to space first and then the other creations since all other physical creations have to be held in the space created. This order of creation is described in the Indian scripture.

Here, we have to remember that each of these creations is constituted by the two mutually opposite components of existences as discussed above. Thus, the creation maintains conservation. If the wave function of a created particle is $x(\mathbf{r}, t)+iy(\mathbf{r}, t)$, then $x^2(\mathbf{r}, t)$ represents one existence-state of the particle and $-y^2(\mathbf{r}, t)$ represents the other existence-state of the particle. However, these two terms should not be simply summed together, since it means counteracting the creation with a result that the particle's very existence grows and vanishes with time. The conventional interpretation does sum up the two terms, however by changing the sign of the second term, to get a positive definite quantity. Because of this, we get a function for existence that may be dependent on space and monotonously independent of time. It preserves the 'reality of existence', however, monotonously. However, it fails in explaining the conservation during the creation of everything. The bipolar interpretation also preserves the reality of existence, however, in a dynamic manner that it provides an insight into what spin is, which consequently demonstrates orbital revolution.

Thus, the conventional interpretation does well as long as the reality of existences is concerned without concerning the origin of the existences since it deals with both the polarities of the existences not only equivalently but also as one and the same. However, the new bipolar-existence interpretation is necessary to demystify the concepts like conservation during creation, quantum spin and more aspects to be discussed further.

Creation of time

The spinning of a particle is defined above as its transformation between its two states of existences. At a given time, the transformations of existence-states of two particles spinning in opposite directions, are opposite in direction. However, the only difference between them is in a shift in time by $\pi/(2\omega)$. The existence-states and the transformation of existence-states of two similar particles spinning in opposite directions, one considered at time t and the other considered at time $t + \pi/(2\omega)$ are the same. Then, what is unique about a particular direction of spin? Since the difference between the two transformations corresponding to the two directions of spins is in time, the time at which a transformation happens has to be unique. From this requirement and the requirement for the conservation during the creation of all physical quantities, I understand that time also, being a creation, has two mutually opposite existence-states. With this understanding about the time, the uniqueness of the direction of spin of other creations becomes evident. The transformations of existence-states of all other physical creations are 'synchronized with' or 'rooted at' the transformation of existence-states of time. The direction of the spin of a particle means the similarity or dissimilarity between the directions of transformation of its existence-states and the direction of transformation of existence-state of time. In the first instance, this disparity, between the two directions of spins with respect to the direction of transformation of time, may appear to contradict my earlier proclamation that the two existence-states are very similar. However, this is not a contradiction since we have the freedom to consider the polarity of either of the existence-states of time as either positive or negative so that either of the spins can be viewed to be similar or dissimilar with respect to the spin of time. This brings back the equal stature between the two spins as well as the existence-states.

From these discussions, it follows that time is a creation that is connected with all the instances of all other physical creations like space, mass etc, whereas each instance of those other creations (space etc) are individualistic, being connected with their corresponding quantum of subtle source. This indicates that the subtle source of time is something different from and subtler than the subtle source of the other physical creations. Time, being common to all other creations and having a subtler source of creation, time can exist before the “big bang” that marks the start of this universe. Hence, time is the first creation and the subtle source of time should be subtler than the subtle source from which the other physical existences are created. Thus, time is always ticking.

This picture of time resembles the clock pulse in a digital computer. In a digital computer, an event/command is executed in synchronization with every clock pulse. Similarly, in Nature, in synchronization with every spin of time, the other creations spin. Recall the hypothesis from the section,

'Angular motion of electrons in atoms' above, that spin is the fundamental entity responsible for the dynamics that are also more complex than the angular revolution of an electron.

The fact that time, being a single entity, shows different periods of its spin to the spins of different other existences, appears to imply some theory of relativity, probably the existing theories of relativity. This also indicates the subtlety of time and the inadequacy in our understanding about time.

Gravity

Gravity, according to me, is the force of attraction between the specks of subtle source and hence it is 'within the subtle source'. Hence, out of the four fundamental forces, gravity does not need a created particle to act as a carrier of force. Newton stated that the gravitational force exists between instances of physical mass. However, I infer that the gravitational force exists between the instances of subtle source (specks or quanta of subtle source). Is there any verification that gravitation is the force of attraction rather between the specks of subtle source from which physical creations are created than between the instances of mass that are creations? Yes, this can be verified by understanding the existence of “Dark matter”.

Dark Matter

Observational cosmologists have discovered Dark matter by the gravitational force that they exert [2], [3], [4] in the absence of any luminous matter (mass) at places where dark matter is envisaged.

The hypothesis

My hypothesis on gravitation agrees with Newton’s law of gravitation in the sense that every physical mass also has its subtle source with it. However, the latter is a special case of the former. The regions of space in the cosmos devoid of physical mass also have their portions of subtle source and hence should exert gravity, according to my hypothesis. This gravity is observed as caused by some matter called as 'dark matter'. According to the hypothesis, dark matter is just space and the gravitation observed is caused by the subtle source of that space. If this explanation is true, they need not be called as dark matter anymore since they are just the space. Cosmologists have mapped the distribution of dark matter. This actually could be the distribution of space in the physical universe.

Mass-Energy conversion

Creation of mass from energy or energy from mass in mass-energy conversion events is not equivalent to the creation of mass and energy from the subtle source. The latter is the real creation while the former are transformations within the creations. However, subtle source also plays a role in the mass –

energy conversions. The hypothesis proposed is that when a gamma ray photon of sufficient energy intercepts a quantum of subtle source (for example, at the center of an atom), a particle-antiparticle pair is produced. Thus, subtle source transform energy into mass in addition to giving birth to both of them. By this way, pair production can happen even in 'empty' space since empty space also has its subtle source with it, which we have called as dark matter! The experimental observation of these 'antiparticles' as originating from cosmic space can be an experimental verification of this hypothesis.

Conservation between source and creation

Can there be conservation between the subtle source and the creation? The 'dark matter effect' of the subtle source indicates that there could be some sort of conservation between a speck or quantum of subtle source and the creation from it. The indication is by the following observation.

The local dark matter density in the milky way galaxy (the density of subtle source giving birth to only space) is estimated [5] to be less than the physical matter density (the density of subtle source giving birth to physical matter, probably in addition to space) by more than 20 orders of magnitude. This indicates that the source gives birth to more space when giving birth to only space than when giving birth to other creations, probably in addition to space. In other words, subtle source density is more with physical matter than with 'empty' space. This amounts to an inference that there could be conservation between the subtle source and its creations.

Curvature of space

It is interesting to conceive the curvature of space from this picture. Each physical entity like a particle and a quantum of space has its existence radiating out from the corresponding quantum of subtle source located at its center. Moreover, the quanta of subtle source maintain the link among them, gravity being the force of attraction between them. Since the density of subtle source, as discussed previously, is orders of magnitude higher in celestial objects like planets and stars than in the space between them, we observe the space curving towards these celestial objects.

Interweave between source and creation

How are the subtle source and its creations interwoven with each other? If the source also exists with the creation, how is the infusion between the source and the creations? Various possibilities are assumed:

1. There is no infusion between the source and the creations. The quantum of source remains as an ideal point without occupying the space created from it.

2. The source evenly infuses with the creation.

3. The source has a graded infusion with the creation. This is a core-shell model. This can be visualized for an atomic system by comparing with the planetary system. Because of the finite mass of the sun and the consequent pull by the orbiting planets, the sun as well as the planets can be modeled to be orbiting in spheroidal shells, the result being that the motion of planets relative to the sun is elliptical as in a spirograph. An atom can also be expected to have similar spheroidal shells for the nucleus and the electrons. The intensities of these atomic shells would be vanishing gradually towards the core as well as outwards. We could observe this if we solve the hydrogen atom problem as a literally two-body problem without invoking the reduced-mass concept that reduces the problem to a one-body problem. We can expect the subtle source to occupy this central core of these shells in an atom with its intensity complementing the intensity of the creation (the nucleus and electrons).

Cold fusion

The third model above proposes a model for the cold fusion [6]. When two atoms come so close that there is a critical overlap between the intensities of the subtle sources of the two nuclei, the sources coalesce leading to the 'cold fusion'. The dimensions of this core-shell model can be computed and the corresponding 'active-surface' potentials that can facilitate this mutual approach of nuclei causing the LENR (Low Energy Nuclear Reactions) [7] can be studied.

Dark Energy

The expansion of universe had been earlier understood to be simply inertial [8]. Recently, it is observed that the rate of expansion of the universe is increasing [2], [9]. Physicists, generally, imagine a physical energy causing this acceleration of expansion of the universe. Since the source of this energy is not known, it is called as 'dark energy'.

The interpretation

Many natural laws have been discovered so far. These are the laws created with the physical universe and for the physical universe. However, these laws do not bind why and how the physical universe and these laws themselves were created.

Let us consider the creation that started with the 'big bang'. Is it still continuing? The observed accelerated expansion of the universe suggests that the creation of space is continuing at this accelerated pace. Thus, this is not a form of physical energy, contrary to the common belief. It is the 'energy' of creation. Our science deals only with the created physical universe. The act of creation is beyond the physical laws since those laws are also amongst those created. Hence, it is not necessary to consider 'dark energy' as a physical energy.

Conclusions

A new interpretation of the quantum mechanical wave functions is proposed. This new interpretation leads to a new interpretation of the spin of elementary particles. This also enables one to 'observe' orbital motion of electrons in atoms and proposes that spin is the base event of any more complex dynamics too. Consequent to this interpretation, the concept of subtle source is introduced, which is the source for the creation of the universe. This leads to the new definition of Gravity and explains why gravity does not require a force-carrying particle. This also deciphers the problem of dark matter. This analysis suggests that time is created before the other physical existences were created. The problem of dark energy associated with the accelerated expansion of the universe is deciphered by discerning the physical laws from laws of creation.

References

[1] http://en.wikipedia.org/wiki/Spin_(physics)#Elementary_particles

[2] http://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy/

[3] http://en.wikipedia.org/wiki/Dark_matter

[4] http://imagine.gsfc.nasa.gov/docs/science/know_l1/dark_matter.html

[5] http://arxiv.org/abs/0907.0018v2

[6] http://en.wikipedia.org/wiki/Cold_fusion

[7] http://www.lenr-canr.org/

[8] http://en.wikipedia.org/wiki/Metric_expansion_of_space

[9] Riess, A. et al. 1998, *Astronomical Journal*, 116, 1009

Bivariate least squares linear regression: towards a unified analytic formalism. II. Extreme structural models

R. Caimmi

Dipartimento di Fisica e Astronomia, Università di Padova, Vicolo Osservatorio 3/2, I-35122 Padova, Italy email: roberto.caimmi@unipd.it fax: 39-049-8278212

Abstract

Concerning bivariate least squares linear regression, the classical results obtained for extreme structural models in earlier attempts [14, 10] are reviewed using a new formalism in terms of deviation (matrix) traces which, for homoscedastic data, reduce to usual quantities leaving aside an unessential (but dimensional) multiplicative factor. Within the framework of classical error models, the dependent variable relates to the independent variable according to a variant of the usual additive model. The classes of linear models considered are regression lines in the limit of uncorrelated errors in X and in Y. The following models are considered in detail: (Y) errors in X negligible (ideally null) with respect to errors in Y; (X) errors in Y negligible (ideally null) with respect to errors in X; (C) oblique regression; (O) orthogonal regression; (R) reduced major-axis regression; (B) bisector regression. For homoscedastic data, the results are taken from earlier attempts and rewritten using a more compact notation. For heteroscedastic data, the results are inferred from a procedure related to functional models [23, 7]. An example of astronomical application is considered, concerning the [O/H]-[Fe/H] empirical relations deduced from five samples related to different stars and/or different methods of oxygen abundance determination. For low-dispersion samples and assigned methods, different regression models yield results which are

in agreement within the errors ($\mp\sigma$) for both heteroscedastic and homoscedastic data, while the contrary holds for large-dispersion samples. In any case, samples related to different methods produce discrepant results, due to the presence of (still undetected) systematic errors, which implies no definitive statement can be made at present. Asymptotic expressions approximate regression line slope and intercept variance estimators, for normal residuals, to a better extent with respect to earlier attempts. Related fractional discrepancies are not exceeding a few percent for low-dispersion data, which grows up to about 10% for large-dispersion data. An extension of the formalism to generic structural models is left to further investigation.

keywords: bivariate regression; linear regression; least squares regression;

1 Introduction

Linear regression is a fundamental and frequently used statistical tool in almost all branches of science, among which astronomy. The related problem is twofold: regression line slope and intercept estimators are expressed involving minimizing or maximizing some function of the data; on the other hand, regression line slope and intercept variance estimators are expressed requiring knowledge of the error distributions of the data. The complexity mainly arises from the occurrence of intrinsic dispersion in addition to the dispersion related to the measurement processes (hereafter quoted as instrumental dispersion), where the distribution corresponding to the former can be different from the distribution corresponding to the latter i.e. non Gaussian (non normal).

In statistics, problems where the true points have fixed but unknown coordinates are called functional regression models, while problems where the true points have random (i.e. obeying their own intrinsic distribution) and unknown coordinates are called structural regression models. Accordingly, functional regression models may be conceived as structural regression models where the intrinsic dispersion is negligible (ideally null) with respect to the instrumental dispersion. Conversely, structural regression models where the instrumental dispersion is negligible (ideally null) with respect to the intrinsic dispersion, can be defined as extreme structural models [7]. A distinction between functional and structural modelling is currently preferred, where the former can be affected by intrinsic scatter but with no or only min-

imal assumptions on related distributions, while the latter implies (usually parametric) models are placed on the above mentioned distributions. For further details refer to specific textbooks e.g., [8] Chap. 2, §2.1. In addition, models where the instrumental dispersion is the same from point to point for each variable, are called homoscedastic models, while models where the instrumental dispersion is (in general) different from point to point, are called heteroscedastic models. Similarly, related data are denoted as homoscedastic and heteroscedastic, respectively.

In general, problems where the true points lie precisely on an expected line relate to functional regression models, while problems where the true points are (intrinsically) scattered about an expected line relate to structural regression models [10, 11].

Bivariate least squares linear regression related to heteroscedastic functional models with uncorrelated and correlated errors, following Gaussian distributions, were analysed and formulated in two classical papers [23, 25], where regression line slope and intercept variance estimators are determined using the method of partial differentiation [25]. On the contrary, the method of moments estimator is used to this aim in later attempts e.g., [12] Chap. 1, §1.3.2, Eq. (1.3.7), [10].

Bivariate least squares linear regression related to extreme structural models, where the instrumental dispersion is negligible (ideally null) with respect to intrinsic dispersion, was exhaustively treated in two classical papers [14, 10] and extended to generic structural models in a later attempt [1].

The above mentioned papers provide the simplest description of linear regression. In reality, biases and additional effects must be taken into consideration, which implies much more complicated description and formulation, as it can be seen in specific monographies e.g., [12, 8, 4]. Restricting to the astronomical literature, a recent investigation [15] is particularly relevant in that it is the first example (in the field under discussion) where linear regression is considered following the modern (since about half a century ago) approach based on likelihoods rather than the old (up to about a century ago) least-squares approach. More specifically, a hierarchical measurement error model is set up therein, the complicated likelihood is written down, and a variety of minimum least-squares and Bayesan solutions are shown, which can treat functional, structural, multivariate, truncated and censored mesaurement error regression problems.

Even in dealing with the simplest homoscedastic (or heteroscedastic) func-

tional and structural models, still no unified analytic formalism has been developed (to the knowledge of the author) where (i) structural heteroscedastic models with instrumental and intrinsic dispersion of comparable order in both variables, are considered; (ii) previous results are recovered in the limit of dominant instrumental dispersion; and (iii) previous results are recovered in the limit of dominant intrinsic dispersion. A related formulation may be useful also for computational methods, in the sense that both the general case and limiting situations can be described by a single numerical code.

A first step towards a unified analytic formalism of bivariate least squares linear regression involving functional models was performed in an earlier attempt [7], where the least-squares approach developed in two classical papers [23, 25] was reviewed and reformulated by definition and use of deviation (matrix) traces. The current investigation aims at making a second step along the same direction, in dealing with extreme structural models.

More specifically, the results found in two classical papers [14, 10] shall be reformulated in terms of deviation traces for homoscedastic models, and extended to the general case of heteroscedastic models by analogy with their counterparts related to functional models, within the framework of classical error models where the dependent variable relates to the independent variable according to a variant of the classical additive error model.

In this view, homoscedastic structural models are conceived as models where both the instrumental and the intrinsic dispersion are the same from point to point. Conversely, models where the instrumental and/or the intrinsic dispersion are (in general) different from point to point, are conceived as heteroscedastic structural models.

Regression line slope and intercept estimators, and related variance estimators, are expressed in terms of deviation traces for different homoscedastic models [14, 10] in section 2, where an extension to corresponding heteroscedastic models is also performed, and both normal and non normal residuals are considered. An example of astronomical application is outlined in section 3. The discussion is presented in section 4. Finally, the conclusion is shown in section 5. Some points are developed with more detail in the Appendix. An extension of the formalism to generic structural models is left to further investigation.

2 Least-squares fitting of a straight line

2.1 General considerations

Attention shall be restricted to the classical problem of least-squares fitting of a straight line, where both variables are measured with errors. Without loss of generality, structural models can be conceived as related to an ideal situation where the variables obey a linear relation[1], as:

$$y_i^* = a x_i^* + b \quad ; \qquad 1 \leq i \leq n \quad ; \tag{1}$$

in connection with true points, $\mathsf{P}_i^* \equiv (x_i^*, y_i^*)$, $1 \leq i \leq n$. The occurrence of random (measure independent) processes makes true points shift outside or along the ideal straight line, inferred from Eq. (1), towards actual points, $\mathsf{P}_{\mathrm{S}i} \equiv (x_{\mathrm{S}i}, y_{\mathrm{S}i})$. The occurrence of mesaurement processes makes the actual points shift towards the observed points, $\mathsf{P}_i \equiv (X_i, Y_i)$.

In this view, the least squares fitting of a straight line is conceptually similar for functional (in absence of intrinsic scatter) and structural (in presence of intrinsic scatter) models: "What is the best line fitting to a sample of observed points, P_i, $1 \leq i \leq n$?" It is worth noticing the correspondence between true points, P_i^*, and observed points, P_i, is not one-to-one unless it is assumed all points are shifted along the same direction. More specifically, two observed points, P_i, P_j, with equal coordinates, $(X_i, Y_i) = (X_j, Y_j)$, relate to true points, P_i^*, P_j^*, with (in general) different coordinates, $(x_i^*, y_i^*) \neq (x_j^*, y_j^*)$, both in presence and in absence of intrinsic scatter. The least-square estimator and the loss function have the same formal expression for functional and structural models, but in the latter case the "statistical distances" e.g., [12] Chap. 1, §1.3.3, depend on the total (instrumental + intrinsic) scatter.

The observed points and the actual points are related as:

$$Z_i = z_{\mathrm{S}i} + (\xi_{\mathrm{F}_z})_i \quad ; \quad Z = X, Y \quad ; \quad z = x, y \quad ; \quad 1 \leq i \leq n \quad ; \tag{2}$$

where $(\xi_{\mathrm{F}_x})_i$, $(\xi_{\mathrm{F}_y})_i$, are the instrumental (i.e. due to the intrumental scatter) errors on $x_{\mathrm{S}i}$, $y_{\mathrm{S}i}$, respectively, assumed to obey Gaussian distributions with null expectation values and known variances, $[(\sigma_{xx})_{\mathrm{F}}]_i$, $[(\sigma_{yy})_{\mathrm{F}}]_i$, and covariance, $[(\sigma_{xy})_{\mathrm{F}}]_i$.

[1]The Italian convenction shall be adopted here, according to which the slope and the intercept of a straight line on the Cartesian plane, are denoted as a, b, respectively.

The actual points and the true points on the ideal straight line are related as:

$$z_{\mathrm{S}i} = z_i^* + (\xi_{\mathrm{S}_z})_i \quad ; \quad z = x, y; \quad 1 \leq i \leq n \quad ; \tag{3}$$

where $(\xi_{\mathrm{S}_x})_i$, $(\xi_{\mathrm{S}_y})_i$, are the intrinsic (i.e. due to the intrinsic scatter) errors on x_i^*, y_i^*, respectively, assumed to obey specified distributions with null expectation values and finite variances, $[(\sigma_{xx})_{\mathrm{S}}]_i$, $[(\sigma_{yy})_{\mathrm{S}}]_i$, and covariance, $[(\sigma_{xy})_{\mathrm{S}}]_i$.

The observed points and the true points on the ideal straight line are related as:

$$Z_i = z_i^* + \xi_{z_i} \quad ; \quad Z = X, Y; \quad z = x, y \quad ; \quad 1 \leq i \leq n \quad ; \tag{4}$$

where the (instrumental + intrinsic) errors, ξ_{x_i}, ξ_{y_i}, are defined as:

$$\xi_{z_i} = (\xi_{\mathrm{F}_z})_i + (\xi_{\mathrm{S}_z})_i \quad ; \quad z = x, y \quad ; \quad 1 \leq i \leq n \quad ; \tag{5}$$

which obey specified distributions with null expectation values and finite variances, $(\sigma_{xx})_i$, $(\sigma_{yy})_i$, and covariance, $(\sigma_{xy})_i$. The further restriction that $(\xi_{\mathrm{F}_z})_i$, $(\xi_{\mathrm{S}_z})_i$, $z = x, y$, $1 \leq i \leq n$, are independent, implies the relation [1]:

$$(\sigma_{zz})_i = [(\sigma_{zz})_{\mathrm{F}}]_i + [(\sigma_{zz})_{\mathrm{S}}]_i \quad ; \qquad (\sigma_{xy})_i = [(\sigma_{xy})_{\mathrm{F}}]_i + [(\sigma_{xy})_{\mathrm{S}}]_i \quad ; \tag{6}$$

where the intrinsic covariance matrixes are unknown and must be assigned or estimated, which will be supposed in the following.

Then the error model is defined by Eqs. (1)-(6), where both instrumental errors, $(\xi_{\mathrm{F}_z})_i$, and intrinsic errors, $(\xi_{\mathrm{S}_z})_i$, are assumed to be independent of true values, z_i^*, for given instrumental covariance matrixes, $(\Sigma_{\mathrm{F}})_i = || \, [(\sigma_{xy})_{\mathrm{F}}]_i \, ||$, intrinsic covariance matrixes, $(\Sigma_{\mathrm{S}})_i = || \, [(\sigma_{xy})_{\mathrm{S}}]_i \, ||$, respectively, and (total) covariance matrixes, $\Sigma_i = || \, (\sigma_{xy})_i \, ||$, hence $\Sigma_i = (\Sigma_{\mathrm{F}})_i + (\Sigma_{\mathrm{S}})_i$, $1 \leq i \leq n$. It may be considered as a variant of the classical additive error model e.g., [1], [8] Chap. 1, §1.2, Chap. 3, §3.2.1, [15, 16], [4] Chap. 4, §4.3.

In the case under discussion, the regression estimator minimizes the loss function, defined as the sum (over the n observations) of squared residuals e.g., [25], or statistical distances of the observed points, $\mathsf{P}_i \equiv (X_i, Y_i)$, from the estimated line in the unknown parameters, $a, b, x_1, ..., x_n$, e.g., [12] Chap. 1, §1.3.3. Under restrictive assumptions, the regression estimator is the functional maximum likelihood estimator e.g., [8] Chap. 3, §3.4.2.

The coordinates, (x_i, y_i), may be conceived as the adjusted values of related observations, (X_i, Y_i), on the estimated regression line [23, 25] and, in addition, as estimators of the coordinates, (x_i^*, y_i^*), on the true regression line i.e. the ideal straight line. The line of adjustment, $\overline{\mathsf{P}_i\hat{\mathsf{P}}_i}$ e.g., [25], may be conceived as an estimator of the statistical distance, $\overline{\mathsf{P}_i\mathsf{P}_i^*}$ e.g., [12] Chap. 1, §1.3.3, where $\hat{\mathsf{P}}_i(x_i, y_i)$ is the adjusted point on the estimated regression line, which implies the relation:

$$y_i = \hat{a}x_i + \hat{b} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{7}$$

where, in general, estimators are denoted by hats, and $\mathsf{P}_i^*(x_i^*, y_i^*)$ is the true point on the ideal straight line, Eq. (1).

To the knowledge of the author, only classical error models are considered for astronomical applications, and for this reason different error models such as Berkson models and mixture error models e.g., [8] Chap. 3, Sect. 3.2, shall not be dealt with in the current attempt. From this point on, investigation shall be limited to extreme structural models and least-squares regression estimators for the following reasons. First, they are important models in their own right, furnishing an approximation to real world situations. Second, a careful examination of these simple models helps for understanding the theoretical underpinnings of methods for other models of greater complexity such as hierarchical models e.g., [15, 16].

2.2 Extreme structural models

With regard to extreme structural models, bivariate least squares linear regression were analysed in two classical papers in the special case of oblique regression i.e. constant variance ratio, $(\sigma_{yy})_i/(\sigma_{xx})_i = c^2$, $1 \leq i \leq n$, and constant correlation coefficients, $r_i = r$, $1 \leq i \leq n$. More specifically, orthogonal ($c^2 = 1$) and oblique regression were analysed in the earlier [14] and in the latter [10] paper, respectively. In absence of additional information, homoscedastic models are used [14] unless the intrinsic dispersion is estimated [1], from which related weights may be determined and the least squares estimator together with the loss function may be expressed for both homoscedastic and heteroscedastic models [1, 16].

The (dimensionless) squared weighted residuals can be defined as in the

case of functional models [25]:

$$(\widetilde{R}_i)^2 = \frac{w_{x_i}(X_i - x_i)^2 + w_{y_i}(Y_i - y_i)^2 - 2r_i\sqrt{w_{x_i}w_{y_i}}(X_i - x_i)(Y_i - y_i)}{1 - r_i^2} \quad ; \tag{8a}$$

$$r_i = \frac{(\sigma_{xy})_i}{[(\sigma_{xx})_i(\sigma_{yy})_i]^{1/2}} \quad ; \qquad |r_i| \leq 1 \quad ; \qquad 1 \leq i \leq n \quad ; \tag{8b}$$

where w_{x_i}, w_{y_i}, are the weights of the various measurements (or observations) and r_i the correlation coefficients. The terms, $w_{x_i}(X_i - x_i)^2$, $w_{y_i}(Y_i - y_i)^2$, r_i, $1 \leq i \leq n$, are dimensionless by definition. An equivalent formulation in matrix formalism can be found in specific textbooks, where weighted true residuals are conceived as (dimensionless) "statistical distances" from data points to related points on the regression line e.g., [12] Chap. 1, §1.3.3, Eq. (1.3.16).

Accordingly, the least-squares regression estimator and the loss function can be expressed as in the case of functional models [7] but the weights, w_{x_i}, w_{y_i}, and the correlation coefficients, r_i, are related to intrinsic scatter instead of instrumental scatter. Then the regression line slope and intercept estimators take the same formal expression with respect to their counterparts related to functional models, while (in general) the contrary holds for regression line slope and intercept variance estimators.

Classical results on extreme structural models [14, 10] are restricted to oblique regression for homoscedastic data with constant correlation coefficients ($w_{x_i} = w_x$, $w_{y_i} = w_y$, $r_i = r$, $1 \leq i \leq n$). In the following subsections, the above mentioned results extended to heteroscedastic data shall be expressed in terms of weighted deviation (matrix) traces [7]:

$$\widetilde{Q}_{pq} = \sum_{i=1}^{n} Q_i(w_{x_i}, w_{y_i}, r_i)(X_i - \widetilde{X})^p(Y_i - \widetilde{Y})^q \quad ; \tag{9}$$

$$\widetilde{Q}_{00} = \sum_{i=1}^{n} Q_i(w_{x_i}, w_{y_i}, r_i) = n\overline{Q} \quad ; \tag{10}$$

where $\widetilde{Q}_{pq}$ are the (weighted) pure ($p = 0$ and/or $q = 0$) and mixed ($p > 0$ and $q > 0$) deviation traces, and $\widetilde{X}$, $\widetilde{Y}$, are weighted means:

$$\widetilde{Z} = \frac{\sum_{i=1}^{n} W_i Z_i}{\sum_{i=1}^{n} W_i} \quad ; \qquad Z = X, Y \quad ; \tag{11}$$

$$W_i = \frac{w_{x_i}\Omega_i^2}{1 + a^2\Omega_i^2 - 2ar_i\Omega_i} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{12}$$

$$\Omega_i = \sqrt{\frac{w_{y_i}}{w_{x_i}}} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{13}$$

in the limit of homoscedastic data with equal correlation coefficients, $w_{x_i} = w_x, w_{y_i} = w_y, r_i = r, 1 \leq i \leq n$, which implies $Q_i(w_{x_i}, w_{y_i}, r_i) = Q(w_x, w_y, r) = Q$, Eqs. (9), (10), (11), (12), and (13) reduce to:

$$\widetilde{Q}_{pq} = QS_{pq} \quad ; \tag{14}$$

$$S_{pq} = \sum_{i=1}^{n}(X_i - \overline{X})^p(Y_i - \overline{Y})^q \quad ; \tag{15}$$

$$\widetilde{Q}_{00} = QS_{00} \quad ; \tag{16}$$

$$S_{00} = n \quad ; \tag{17}$$

$$\tilde{Z} = \overline{Z} \quad ; \qquad Z = X, Y \quad ; \tag{18}$$

$$W_i = W = \frac{w_x\Omega^2}{1 + a^2\Omega^2 - 2ar\Omega} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{19}$$

$$\Omega_i = \Omega = \sqrt{\frac{w_y}{w_x}} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{20}$$

where S_{pq} are the (unweighted) pure ($p = 0$ and/or $q = 0$) and mixed ($p > 0$ and $q > 0$) deviation traces.

Turning to the general case and using the weighted squared error loss function, $T_{\widetilde{R}} = \sum_{i=1}^{n}(\widetilde{R}_i)^2$, yields for regression line slope and intercept estimators the same expression with respect to functional models [7]. Accordingly, regression line slope and intercept estimators may be conceived similarly to state functions in thermodynamics: for an assigned true point, $\mathsf{P}_i^* \equiv (x_i^*, y_i^*)$, what is relevant is the related observed point, $\mathsf{P}_i \equiv (X_i, Y_i)$, regardless of the path followed via instrumental and/or intrinsic scatter. More specifically, the regression line intercept estimator obeys the equation e.g., [25, 7]:

$$\hat{b} = \widetilde{Y} - \hat{a}\widetilde{X} \quad ; \tag{21}$$

which implies the "barycentre" of the data, $\widetilde{\mathsf{P}} \equiv (\widetilde{X}, \widetilde{Y})$, lies on the estimated regression line, inferred from Eq. (7), and the regression line slope estimator is one among three real solutions of a pseudo cubic equation or two real solutions of a pseudo quadratic equation, where the coefficients are weakly

dependent on the unknown slope. For further details refer to earlier attempts [23, 25, 7]. The above mentioned equations have the same formal expression for functional and structural models, which also holds for the regression line slope and intercept estimators.

The regression line slope and intercept variance estimators for functional models, calculated using the method of partial differentiation e.g., [25] and the method of moments estimators e.g., [12] Chap. 1, §1.3.2, Eq. (1.3.7), yield, in general, different results [7]. The same is expected to hold, a fortiori, for structural models, for which the method of moments estimators and the δ-method have been exploited in classical investigations e.g., [14, 10]. Accordingly, related results shall be considered and expressed in terms of unweighted deviation traces for homoscedastic data with equal correlation coefficients and extended in terms of weighted deviation traces for heteroscedastic data, with regard to a number of special cases considered in earlier attempts in the limit of uncorrelated errors in X and in Y [14, 10]. With this restriction, the pseudo cubic equation reduces to:

$$\widetilde{V}_{20}a^3 - 2\widetilde{V}_{11}a^2 - (\widetilde{W}_{20} - \widetilde{V}_{02})a + \widetilde{W}_{11} = 0 \quad ; \tag{22}$$

where the deviation traces are defined by Eq. (9), via Eq. (12) and $V_i = W_i^2/w_{x_i}$. For further details refer to the parent paper [23] and to a recent attempt [7]. A formulation of Euclidean and statistical squared residual sum for homoscedastic and heteroscedastic data is expressed in A.

2.3 Errors in X negligible with respect to errors in Y

In the limit of errors in X negligible with respect to errors in Y, $a^2(\sigma_{xx})_i \ll (\sigma_{yy})_i$, $a(\sigma_{xy})_i \ll (\sigma_{yy})_i$, $1 \leq i \leq n$. Ideally, $(\sigma_{xx})_i \to 0$, $(\sigma_{xy})_i \to 0$, $1 \leq i \leq n$, which implies $r_i \to 0$, $w_{x_i} \to +\infty$, $\Omega_i \to 0$, $W_i \to w_{y_i}$, $1 \leq i \leq n$. Accordingly, the errors in X and in Y are uncorrelated.

For homoscedastic data, $w_{x_i} = w_x$, $w_{y_i} = w_y$, $1 \leq i \leq n$, the regression line slope and intercept estimators are [14, 7]:

$$\hat{a}_{\mathrm{Y}} = \frac{S_{11}}{S_{20}} \quad ; \tag{23}$$

$$\hat{b}_{\mathrm{Y}} = \overline{Y} - \hat{a}_{\mathrm{Y}}\overline{X} \quad ; \tag{24}$$

where the index, Y, stands for OLS(Y|X) i.e. ordinary least square regression or, in general, WLS(Y|X) i.e. weighted least square regression of the depen-

dent variable, Y, against the independent variable, X [14]. Accordingly, related models shall be quoted as Y models.

The regression line slope and intercept variance estimators, in the special case of normal residuals may be calculated using different methods and/or models e.g., [12] Chap. 1, §1.3.2, Eq. (1.3.7), [10, 7]. The result is:

$$\begin{aligned}[(\hat{\sigma}_{\hat{a}_\mathrm{Y}})_\mathrm{N}]^2 &= \frac{(\hat{a}_\mathrm{Y})^2}{n-2}\left[\frac{(n-2)R_\mathrm{Y}}{\hat{a}_\mathrm{Y}S_{11}} + \Theta(\hat{a}_\mathrm{Y},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X})\right] \\ &= \frac{(\hat{a}_\mathrm{Y})^2}{n-2}\left[\frac{\hat{a}_\mathrm{X}-\hat{a}_\mathrm{Y}}{\hat{a}_\mathrm{Y}} + \Theta(\hat{a}_\mathrm{Y},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X})\right] \quad ; \end{aligned} \tag{25}$$

$$[(\hat{\sigma}_{\hat{b}_\mathrm{Y}})_\mathrm{N}]^2 = \left[\frac{1}{\hat{a}_\mathrm{Y}}\frac{S_{11}}{S_{00}} + (\overline{X})^2\right][(\hat{\sigma}_{\hat{a}_\mathrm{Y}})_\mathrm{N}]^2 - \frac{\hat{a}_\mathrm{Y}}{n-2}\frac{S_{11}}{S_{00}}\Theta(\hat{a}_\mathrm{Y},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X}) \quad ; \tag{26}$$

where the index, N, denotes normal residuals, R is defined in A, and $\hat{a}_\mathrm{X} = S_{02}/S_{11}$. The funcion, $\Theta(\hat{a}_\mathrm{Y},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X})$, is a special case of a more general function, $\Theta(\hat{a}_\mathrm{C},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X})$ which, in turn, depends on the method and/or model used. For further details refer to B.

The regression line slope and intercept variance estimators, in the general case of non normal residuals may be calculated using the δ-method [14]. The result is:

$$(\hat{\sigma}_{\hat{a}_\mathrm{Y}})^2 = \frac{S_{22} + (\hat{a}_\mathrm{Y})^2 S_{40} - 2\hat{a}_\mathrm{Y}S_{31}}{(S_{20})^2} \quad ; \tag{27}$$

$$(\hat{\sigma}_{\hat{b}_\mathrm{Y}})^2 = \frac{\hat{a}_\mathrm{Y}}{n}\frac{\hat{a}_\mathrm{X}-\hat{a}_\mathrm{Y}}{\hat{a}_\mathrm{Y}}\frac{S_{11}}{S_{00}} + (\overline{X})^2(\hat{\sigma}_{\hat{a}_\mathrm{Y}})^2 - \frac{2}{n}\overline{X}\hat{\sigma}_{\hat{b}_\mathrm{Y}\hat{a}_\mathrm{Y}} \quad ; \tag{28}$$

$$\hat{\sigma}_{\hat{b}_\mathrm{Y}\hat{a}_\mathrm{Y}} = \frac{S_{12} + (\hat{a}_\mathrm{Y})^2 S_{30} - 2\hat{a}_\mathrm{Y}S_{21}}{S_{20}} \quad ; \tag{29}$$

where Eqs. (27)-(29) are equivalent to their counterparts expressed in the parent paper [14].

The application of the δ-method provides asymptotic formulae which underestimate the true regression coefficient uncertainty in samples with low ($n \lesssim 50$) or weakly correlated population [10]. In the special case of normal and data-independent residuals, $\Theta(\hat{a}_\mathrm{Y},\hat{a}_\mathrm{Y},\hat{a}_\mathrm{X}) \to 0$, Eqs. (27), (28), must necessarily reduce to (25), (26), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (27)-(29). For further details refer to C.

The expression of the regression line slope and intercept estimators and related variance estimators for normal residuals, Eqs. (23), (24), (25), (26),

coincide with their counterparts determined for Y models in classical and recent attempts e.g., [10] Eq. (4) therein in the limit $c^2 = \sigma_{yy}/\sigma_{xx} \to +\infty$, [17] Eqs. (3)-(7) therein.

For heteroscedastic data, the regression line slope and intercept estimators are [7]:

$$\hat{a}_{\rm Y} = \frac{(\widetilde{w_y})_{11}}{(\widetilde{w_y})_{20}} \quad ; \tag{30}$$

$$\hat{b}_{\rm Y} = \widetilde{Y} - \hat{a}_{\rm Y}\widetilde{X} \quad ; \tag{31}$$

where the weighted means, $\widetilde{X}$ and $\widetilde{Y}$, are defined by Eqs. (11)-(13).

For functional models, regression line slope and intercept variance estimators in the general case of heteroscedastic data reduce to their counterparts in the special case of homoscedastic data, as $\{\hat{\sigma}_{\hat{a}_{\rm Y}}[(\widetilde{w_y})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{a}_{\rm Y}}(w_y S_{pq})]^2$, $\{\hat{\sigma}_{\hat{b}_{\rm Y}}[(\widetilde{w_y})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{b}_{\rm Y}}(w_y S_{pq})]^2$, via Eq. (9) where $Q_i = (w_y)_i = w_y$, $1 \leq i \leq n$. For further details refer to an earlier attempt [7].

Under the assumption that the same holds for extreme structural models, Eqs. (25)-(29) take the general expression:

$$\begin{aligned}[(\hat{\sigma}_{\hat{a}_{\rm Y}})_{\rm N}]^2 &= \frac{(\hat{a}_{\rm Y})^2}{n-2}\left[\frac{n-2}{n}\frac{R_{\rm Y}}{\hat{a}_{\rm Y}}\frac{(\widetilde{w_y})_{00}}{(\widetilde{w_y})_{11}} + \Theta(\hat{a}_{\rm Y}, \hat{a}_{\rm Y}, \hat{a}'_{\rm X})\right] \\ &= \frac{(\hat{a}_{\rm Y})^2}{n-2}\left[\frac{\hat{a}'_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm Y}, \hat{a}_{\rm Y}, \hat{a}'_{\rm X})\right] \quad ; \end{aligned} \tag{32}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm Y}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm Y}}\frac{(\widetilde{w_y})_{11}}{(\widetilde{w_y})_{00}} + (\widetilde{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm Y}})_{\rm N}]^2 - \frac{\hat{a}_{\rm Y}}{n-2}\frac{(\widetilde{w_y})_{11}}{(\widetilde{w_y})_{00}}\Theta(\hat{a}_{\rm Y}, \hat{a}_{\rm Y}, \hat{a}'_{\rm X}) \quad ; \tag{33}$$

$$(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 = \frac{(\widetilde{w_y})_{00}}{n}\frac{(\widetilde{w_y})_{22} + (\hat{a}_{\rm Y})^2(\widetilde{w_y})_{40} - 2\hat{a}_{\rm Y}(\widetilde{w_y})_{31}}{[(\widetilde{w_y})_{20}]^2} \quad ; \tag{34}$$

$$(\hat{\sigma}_{\hat{b}_{\rm Y}})^2 = \frac{\hat{a}_{\rm Y}}{n}\frac{\hat{a}'_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\frac{(\widetilde{w_y})_{11}}{(\widetilde{w_y})_{00}} + (\widetilde{X})^2(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 - \frac{2}{n}\widetilde{X}\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm Y}} \quad ; \tag{35}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm Y}} = \frac{(\widetilde{w_y})_{12} + (\hat{a}_{\rm Y})^2(\widetilde{w_y})_{30} - 2\hat{a}_{\rm Y}(\widetilde{w_y})_{21}}{(\widetilde{w_y})_{20}} \quad ; \tag{36}$$

where $\hat{a}'_{\rm X} = (\widetilde{w_y})_{02}/(\widetilde{w_y})_{11}$, R is defined in A and Θ is expressed in terms of $n(\widetilde{w_y})_{pq}/(\widetilde{w_y})_{00}$ instead of S_{pq}.

In the special case of normal and data-independent residuals, $\Theta(\hat{a}_{\rm Y}, \hat{a}_{\rm Y}, \hat{a}'_{\rm X}) \to 0$, Eqs. (34), (35), must necessarily reduce to (32), (33), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (34)-(36).

In absence of a rigorous proof, Eqs. (32)-(36) must be considered as approximate results.

2.4 Errors in Y negligible with respect to errors in X

In the limit of errors in Y negligible with respect to errors in X, $(\sigma_{yy})_i \ll a^2(\sigma_{xx})_i$, $(\sigma_{xy})_i \ll a(\sigma_{xx})_i$, $1 \leq i \leq n$. Ideally, $(\sigma_{yy})_i \to 0$, $(\sigma_{xy})_i \to 0$, $1 \leq i \leq n$, which implies $r_i \to 0$, $w_{y_i} \to +\infty$, $\Omega_i \to +\infty$, $W_i \to w_{x_i}$, $1 \leq i \leq n$. Accordingly, the errors in X and in Y are uncorrelated. As outlined in an earlier paper [7], the model under discussion can be related to the inverse regression, which has a large associate literature e.g., [18, 13, 19, 2, 17].

For homoscedastic data, $w_{x_i} = w_x$, $w_{y_i} = w_y$, $1 \leq i \leq n$, the regression line slope and intercept estimators are ([14, 7]:

$$\hat{a}_{\mathrm{X}} = \frac{S_{02}}{S_{11}} \quad ; \tag{37}$$

$$\hat{b}_{\mathrm{X}} = \overline{Y} - \hat{a}_{\mathrm{X}}\overline{X} \quad ; \tag{38}$$

where the index, X, stands for OLS(X|Y) i.e. ordinary least square regression or, in general, WLS(X|Y) i.e. weighted least square regression of the dependent variable, X, against the independent variable, Y [14]. Accordingly, related models shall be quoted as X models.

The regression line slope and intercept variance estimators, in the special case of normal residuals may be calculated using different methods and/or models e.g., [12] Chap. 1, §1.3.2, Eq. (1.3.7), [10, 7]. The result is:

$$\begin{aligned}
[(\hat{\sigma}_{\hat{a}_{\mathrm{X}}})_{\mathrm{N}}]^2 &= \frac{(\hat{a}_{\mathrm{X}})^2}{n-2}\left[\frac{(n-2)R_{\mathrm{X}}}{\hat{a}_{\mathrm{X}}S_{11}} + \Theta(\hat{a}_{\mathrm{X}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}})\right] \\
&= \frac{(\hat{a}_{\mathrm{X}})^2}{n-2}\left[\frac{\hat{a}_{\mathrm{X}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \Theta(\hat{a}_{\mathrm{X}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}})\right] \quad ;
\end{aligned} \tag{39}$$

$$[(\hat{\sigma}_{\hat{b}_{\mathrm{X}}})_{\mathrm{N}}]^2 = \left[\frac{1}{\hat{a}_{\mathrm{X}}}\frac{S_{11}}{S_{00}} + (\overline{X})^2\right][(\hat{\sigma}_{\hat{a}_{\mathrm{X}}})_{\mathrm{N}}]^2 - \frac{\hat{a}_{\mathrm{X}}}{n-2}\frac{S_{11}}{S_{00}}\Theta(\hat{a}_{\mathrm{X}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}}) \quad ; \tag{40}$$

where the index, N, denotes normal residuals, R is defined in A, and $\hat{a}_{\mathrm{Y}} = S_{11}/S_{20}$. The funcion, $\Theta(\hat{a}_{\mathrm{X}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}})$, is a special case of a more general function, $\Theta(\hat{a}_{\mathrm{C}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}})$ which, in turn, depends on the method and/or model used. For further details refer to B.

The regression line slope and intercept variance estimators, in the general case of non normal residuals may be calculated using the δ-method [14]. The result is:

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{S_{04} + (\hat{a}_{\rm X})^2 S_{22} - 2\hat{a}_{\rm X} S_{13}}{(S_{11})^2} \quad ; \tag{41}$$

$$(\hat{\sigma}_{\hat{b}_{\rm X}})^2 = \frac{\hat{a}_{\rm X}}{n}\frac{\hat{a}_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\frac{S_{11}}{S_{00}} + (\overline{X})^2(\hat{\sigma}_{\hat{a}_{\rm X}})^2 - \frac{2}{n}\overline{X}\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm X}} \quad ; \tag{42}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm X}} = \frac{S_{03} + (\hat{a}_{\rm X})^2 S_{21} - 2\hat{a}_{\rm X} S_{12}}{S_{11}} \quad ; \tag{43}$$

where Eqs. (41)-(43) are equivalent to their counterparts expressed in the parent paper [14].

The application of the δ-method provides asymptotic formulae which underestimate the true regression coefficient uncertainty in samples with low ($n \lesssim 50$) or weakly correlated population [10]. In the special case of normal and data-independent residuals, $\Theta(\hat{a}_{\rm X}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \to 0$, Eqs. (41), (42), must necessarily reduce to (39), (40), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (41)-(43). For further details refer to C.

For heteroscedastic data, the regression line slope and intercept estimators are [7]:

$$\hat{a}_{\rm X} = \frac{(\widetilde{w_x})_{02}}{(\widetilde{w_x})_{11}} \quad ; \tag{44}$$

$$\hat{b}_{\rm X} = \widetilde{Y} - \hat{a}_{\rm X}\widetilde{X} \quad ; \tag{45}$$

where the weighted means, $\widetilde{X}$ and $\widetilde{Y}$, are defined by Eqs. (11)-(13).

For functional models, regression line slope and intercept variance estimators in the general case of heteroscedastic data reduce to their counterparts in the special case of homoscedastic data, as $\{\hat{\sigma}_{\hat{a}_{\rm X}}[(\widetilde{w_x})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{a}_{\rm X}}(w_x S_{pq})]^2$, $\{\hat{\sigma}_{\hat{b}_{\rm X}}[(\widetilde{w_x})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{b}_{\rm X}}(w_x S_{pq})]^2$, via Eq. (9) where $Q_i = (w_x)_i = w_x$, $1 \le i \le n$. For further details refer to an earlier attempt [7].

Under the assumption that the same holds for extreme structural models, Eqs. (39)-(43) take the general expression:

$$[(\hat{\sigma}_{\hat{a}_{\rm X}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm X})^2}{n-2}\left[\frac{n-2}{n}\frac{R_{\rm X}}{\hat{a}_{\rm X}}\frac{(\widetilde{w_x})_{00}}{(\widetilde{w_x})_{11}} + \Theta(\hat{a}_{\rm X}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X})\right]$$

$$= \frac{(\hat{a}_{\rm X})^2}{n-2}\left[\frac{\hat{a}_{\rm X}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}} + \Theta(\hat{a}_{\rm X},\hat{a}'_{\rm Y},\hat{a}_{\rm X})\right] \quad ; \tag{46}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm X}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm X}}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm X}})_{\rm N}]^2 - \frac{\hat{a}_{\rm X}}{n-2}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}}\Theta(\hat{a}_{\rm X},\hat{a}'_{\rm Y},\hat{a}_{\rm X}) \quad ; \tag{47}$$

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{(\widetilde{w_x})_{00}}{n}\frac{(\widetilde{w_x})_{04} + (\hat{a}_{\rm X})^2(\widetilde{w_x})_{22} - 2\hat{a}_{\rm X}(\widetilde{w_x})_{13}}{[(\widetilde{w_x})_{11}]^2} \quad ; \tag{48}$$

$$(\hat{\sigma}_{\hat{b}_{\rm X}})^2 = \frac{\hat{a}_{\rm X}}{n}\frac{\hat{a}_{\rm X}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2(\hat{\sigma}_{\hat{a}_{\rm X}})^2 - \frac{2}{n}\widetilde{X}\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm X}} \quad ; \tag{49}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm X}} = \frac{(\widetilde{w_x})_{03} + (\hat{a}_{\rm X})^2(\widetilde{w_x})_{21} - 2\hat{a}_{\rm X}(\widetilde{w_x})_{12}}{(\widetilde{w_x})_{11}} \quad ; \tag{50}$$

where $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$, R is defined in A, and Θ is formulated in terms of $n(\widetilde{w_x})_{pq}/(\widetilde{w_x})_{00}$ instead of S_{pq}.

In the special case of normal and data-independent residuals, $\Theta(\hat{a}_{\rm X},\hat{a}'_{\rm Y},\hat{a}_{\rm X}) \to 0$, Eqs. (48), (49), must necessarily reduce to (46), (47), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (48)-(50).

In absence of a rigorous proof, Eqs. (46)-(50) must be considered as approximate results.

2.5 Oblique regression

In the limit of constant y to x variance ratios and constant correlation coefficients, the following relations hold:

$$\frac{(\sigma_{yy})_i}{(\sigma_{xx})_i} = c^2 \; ; \quad \frac{w_{x_i}}{w_{y_i}} = \Omega_i^{-2} = c^2 \; ; \quad \frac{(\sigma_{xy})_i}{(\sigma_{xx})_i} = r_i c = rc \quad ; \quad 1 \le i \le n \quad ; \tag{51}$$

$$W_i = \frac{w_{x_i}}{a^2 + c^2 - 2rac} \quad ; \qquad 1 \le i \le n \quad ; \qquad \frac{(\widetilde{w_y})_{pq}}{(\widetilde{w_y})_{rs}} = \frac{(\widetilde{w_x})_{pq}}{(\widetilde{w_x})_{rs}} \quad ; \tag{52}$$

where the weights are assumed to be inversely proportional to related variances, $w_{z_i} \propto 1/(\sigma_{zz})_i$, $z = x, y$, as usually done e.g., [10]. By definition, c has the dimensions of a slope, which highly simplifies dimension checks throughout equations, and for this reason it has been favoured with respect to different choices exploited in earlier attempts e.g., [23], [12] Chap. 1, §1.3, [10].

It is worth noticing that Eq. (51) holds for both homoscedastic and heteroscedastic data. It can be seen that the lines of adjustment are oriented

along the same direction [24] but are perpendicular to the regression line only in the special case of orthogonal regression, $c^2 = 1$ e.g., [8] Chap. 3, §3.4.2. Accordingly, the term "oblique regression" has been preferred with respect to "generalized orthogonal regression" used in an earlier attempt [7].

The variance ratio, c^2, may be expressed in terms of instrumental and intrinsic variance ratios, c_{F}^2, and c_{S}^2, respectively, as:

$$c^2 = \frac{[(\sigma_{xx})_i]_{\mathrm{F}}}{(\sigma_{xx})_i} c_{\mathrm{F}}^2 + \frac{[(\sigma_{xx})_i]_{\mathrm{S}}}{(\sigma_{xx})_i} c_{\mathrm{S}}^2 \quad ; \qquad 1 \leq i \leq n \quad ; \tag{53a}$$

$$c_{\mathrm{F}}^2 = \frac{[(\sigma_{yy})_i]_{\mathrm{F}}}{[(\sigma_{xx})_i]_{\mathrm{F}}} \quad ; \qquad c_{\mathrm{S}}^2 = \frac{[(\sigma_{yy})_i]_{\mathrm{S}}}{[(\sigma_{xx})_i]_{\mathrm{S}}} \quad ; \qquad 1 \leq i \leq n \quad ; \tag{53b}$$

where $c_{\mathrm{F}}^2 = c_{\mathrm{S}}^2$ implies $c_{\mathrm{F}}^2 = c_{\mathrm{S}}^2 = c^2$; $c^2 \to c_{\mathrm{F}}^2$ for functional models, $[(\sigma_{zz})_i]_{\mathrm{S}} \ll [(\sigma_{zz})_i]_{\mathrm{F}}$, $z = x, y$, $1 \leq i \leq n$; $c^2 \to c_{\mathrm{S}}^2$ for extreme structural models, $[(\sigma_{zz})_i]_{\mathrm{F}} \ll [(\sigma_{zz})_i]_{\mathrm{S}}$, $z = x, y$, $1 \leq i \leq n$.

For homoscedastic data, $w_{x_i} = w_x$, $w_{y_i} = w_y$, $1 \leq i \leq n$, the regression line slope and intercept estimators are [10, 7]:

$$\begin{aligned} \hat{a}_{\mathrm{C}} &= \frac{S_{02} - c^2 S_{20}}{2S_{11}} \left\{ 1 \mp \left[1 + c^2 \left(\frac{S_{02} - c^2 S_{20}}{2S_{11}} \right)^{-2} \right]^{1/2} \right\} \\ &= \frac{\hat{a}_{\mathrm{X}} \hat{a}_{\mathrm{Y}} - c^2}{2\hat{a}_{\mathrm{Y}}} \left\{ 1 \mp \left[1 + c^2 \left(\frac{\hat{a}_{\mathrm{X}} \hat{a}_{\mathrm{Y}} - c^2}{2\hat{a}_{\mathrm{Y}}} \right)^{-2} \right]^{1/2} \right\} \quad ; \end{aligned} \tag{54}$$

$$\hat{b}_{\mathrm{C}} = \overline{Y} - \hat{a}_{\mathrm{C}} \overline{X} \quad ; \tag{55}$$

where the index, C, denotes oblique regression, $\hat{a}_{\mathrm{Y}} = S_{11}/S_{20}$, $\hat{a}_{\mathrm{X}} = S_{02}/S_{11}$, and the double sign corresponds to the solutions of a second-degree equation, where the parasite solution must be disregarded. Accordingly, related models shall be quoted as C models or O models in the special case of orthogonal regression ($c^2 = 1$). For further details refer to an earlier attempt [7].

The regression line slope and intercept variance estimators, in the special case of normal residuals may be calculated using different methods and/or models e.g., [12] Chap. 1, §1.3.2, Eq. (1.3.7), [10, 7]. The result is:

$$\begin{aligned} [(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})_{\mathrm{N}}]^2 &= \frac{(\hat{a}_{\mathrm{C}})^2}{n-2} \left[\frac{(n-2) R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}} S_{11}} + \Theta(\hat{a}_{\mathrm{C}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}}) \right] \\ &= \frac{(\hat{a}_{\mathrm{C}})^2}{n-2} \left[\frac{\hat{a}_{\mathrm{X}} - \hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}} + \frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \Theta(\hat{a}_{\mathrm{C}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}}) \right] \quad ; \end{aligned} \tag{56}$$

Table 1: Explicit expression of the function, $\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})$, appearing in the slope and intercept variance estimator formula for oblique regression, Eq. (56) and (57), respectively, according to different methods and/or models. Symbol captions: $A_{\rm UV} = \hat{a}_{\rm U}/\hat{a}_{\rm V} - 1$; (U,V) = (X,C), (C,Y). Method captions: AFD - asymptotic formula determination; MME - method of moments estimators; LSE - least squares estimation; MPD - method of partial differentiation. Model captions: F - functional; S - structural; E - extreme structural. Case captions: HM - homoscedastic; HT - heteroscedastic.

$\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})$	method	model	case	source
0	AFD	E	HM	[11]
$A_{\rm XC}A_{\rm CY} + \frac{(A_{\rm CY})^2}{n-1}$	MME	E	HM	[10]
$A_{\rm XC}A_{\rm CY} + \frac{(A_{\rm CY})^2}{n-1}$	MME	S	HM	[12]
$\frac{A_{\rm XC}A_{\rm CY}}{n-1} + \left(\frac{A_{\rm CY}}{n-1}\right)^2$	LSE	E	HM	[11]
$2A_{\rm XC}A_{\rm CY}$	MPD	F	HT	[7]

$$[(\hat{\sigma}_{\hat{b}_{\rm C}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm C}}\frac{S_{11}}{S_{00}} + (\overline{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 - \frac{\hat{a}_{\rm C}}{n-2}\frac{S_{11}}{S_{00}}\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \quad ; \quad (57)$$

where R is defined in A and Θ depends on the method and/or model used, as shown in Table 1. For a formal demonstration, see B. The extreme situations, $a_{\rm C} \to a_{\rm Y}$, $a_{\rm C} \to a_{\rm X}$, are directly inferred from Table 1 as $\Theta(a_{\rm Y}, a_{\rm Y}, a_{\rm X}) = 0$, $\Theta(a_{\rm X}, a_{\rm Y}, a_{\rm X}) = 0$, respectively, in all cases. More specifically, the former relation rigorously holds while the latter has to be restricted to large ($n \gg 1$, ideally $n \to +\infty$) samples when appropriate. In general, Θ can be neglected with respect to the remaining terms in the asymptotic expressions of Eqs. (56) and (57). If the residuals are independent of the data, Θ also vanishes regardless of the sample population. For further details refer to B and C.

The regression line slope and intercept variance estimators, in the general case of non normal residuals, may be calculated using the δ-method [14, 10]. The result is:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = (\hat{a}_{\rm C})^2\frac{c^4(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + (\hat{a}_{\rm Y})^4(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2(\hat{a}_{\rm Y})^2c^2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{(\hat{a}_{\rm Y})^2[4(\hat{a}_{\rm Y})^2c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]} \quad ; \quad (58)$$

$$(\hat{\sigma}_{\hat{b}_{\rm C}})^2 = \frac{\hat{a}_{\rm C}}{n}\left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}} + \frac{\hat{a}_{\rm C} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2(\hat{\sigma}_{\hat{a}_{\rm C}})^2$$

$$- \frac{2}{n}\overline{X}(\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}}) \quad ; \tag{59}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{S_{13} + \hat{a}_{\rm Y}\hat{a}_{\rm X}S_{31} - (\hat{a}_{\rm Y} + \hat{a}_{\rm X})S_{22}}{S_{20}S_{11}} \quad ; \tag{60}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C}c^2}{\hat{a}_{\rm Y}[4(\hat{a}_{\rm Y})^2c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]^{1/2}} \frac{S_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm C}S_{30} - (\hat{a}_{\rm Y} + \hat{a}_{\rm C})S_{21}}{S_{20}} \quad ; \tag{61}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C}\hat{a}_{\rm Y}}{[4(\hat{a}_{\rm Y})^2c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]^{1/2}} \frac{S_{03} + \hat{a}_{\rm X}\hat{a}_{\rm C}S_{21} - (\hat{a}_{\rm X} + \hat{a}_{\rm C})S_{12}}{S_{11}} \quad ; \tag{62}$$

where Eqs. (58) and (59) in the special case, $c^2 = 1$, are equivalent to their counterparts expressed in the parent paper [14] provided absolute values appearing therein are removed. For a formal discussion refer to D. In addition, Eq. (58) is equivalent to its counterpart expressed in the parent paper [11].

The dependence on the variance ratio, c^2, in Eqs. (58), (61), (62), may be eliminated via Eq. (142), B. The result is:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = (\hat{a}_{\rm C})^2$$
$$\times \frac{(\hat{a}_{\rm C})^4(A_{\rm XC})^2(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + (\hat{a}_{\rm Y})^4(A_{\rm CY})^2(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY}\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{(\hat{a}_{\rm Y})^2\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY} + [\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}}; \tag{63}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} = \frac{(\hat{a}_{\rm C})^3A_{\rm XC}}{\hat{a}_{\rm Y}\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY} + [(\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}^{1/2}}$$
$$\times \frac{S_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm C}S_{30} - (\hat{a}_{\rm Y} + \hat{a}_{\rm C})S_{21}}{S_{20}} \quad ; \tag{64}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C}\hat{a}_{\rm Y}A_{\rm CY}}{\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY} + [(\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}^{1/2}}$$
$$\times \frac{S_{03} + \hat{a}_{\rm X}\hat{a}_{\rm C}S_{21} - (\hat{a}_{\rm X} + \hat{a}_{\rm C})S_{12}}{S_{11}} \quad ; \tag{65}$$

$$A_{\rm UV} = \frac{\hat{a}_{\rm U}}{\hat{a}_{\rm V}} - 1 \quad ; \qquad ({\rm U},{\rm V}) = ({\rm X},{\rm C}),({\rm C},{\rm Y}),({\rm X},{\rm Y}) \quad ; \tag{66}$$

in terms of slope estimators, variance slope estimators, and deviation traces.

The application of the δ-method provides asymptotic formulae which underestimate the true regression coefficient uncertainty in samples with low ($n \lesssim 50$) or weakly correlated population [10]. In the special case of normal and data-independent residuals, $\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \to 0$, Eqs. (58), (59), must necessarily reduce to (56), (57), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (27), (41), and (59)-(62). For further details refer to C.

For heteroscedastic data, the regression line slope and intercept estimators are [7]:

$$\hat{a}_{\rm C} = \frac{(\widetilde{w_x})_{02} - c^2(\widetilde{w_x})_{20}}{2(\widetilde{w_x})_{11}} \left\{ 1 \mp \left[1 + c^2 \left(\frac{(\widetilde{w_x})_{02} - c^2(\widetilde{w_x})_{20}}{2(\widetilde{w_x})_{11}} \right)^{-2} \right]^{1/2} \right\}$$

$$= \frac{\hat{a}_{\rm X}\hat{a}'_{\rm Y} - c^2}{2\hat{a}'_{\rm Y}} \left\{ 1 \mp \left[1 + c^2 \left(\frac{\hat{a}_{\rm X}\hat{a}'_{\rm Y} - c^2}{2\hat{a}'_{\rm Y}} \right)^{-2} \right]^{1/2} \right\} \quad ; \tag{67}$$

$$\hat{b}_{\rm C} = \widetilde{Y} - \hat{a}_{\rm C}\widetilde{X} \quad ; \tag{68}$$

where $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$; $\hat{a}_{\rm X} = (\widetilde{w_x})_{02}/(\widetilde{w_x})_{11}$; and the weighted means, $\widetilde{X}$, $\widetilde{Y}$, are defined by Eqs. (11)-(13).

For functional models, regression line slope and intercept variance estimators in the general case of heteroscedastic data reduce to their counterparts in the special case of homoscedastic data, as $\{\hat{\sigma}_{\hat{a}_{\rm C}}[(\widetilde{w_x})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{a}_{\rm C}}(w_x S_{pq})]^2$, $\{\hat{\sigma}_{\hat{b}_{\rm C}}[(\widetilde{w_x})_{pq}]\}^2 \to [\hat{\sigma}_{\hat{b}_{\rm C}}(w_x S_{pq})]^2$, via Eq. (9) where $Q_i = (w_x)_i = w_x$, $1 \le i \le n$. For further details refer to an earlier attempt [7].

Under the assumption that the same holds for extreme structural models, Eqs. (56)-(65) take the general expression:

$$[(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm C})^2}{n-2} \left[\frac{n-2}{n} \frac{R_{\rm C}}{\hat{a}_{\rm C}} \frac{(\widetilde{w_x})_{00}}{(\widetilde{w_x})_{11}} + \Theta(\hat{a}_{\rm C}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X}) \right]$$

$$= \frac{(\hat{a}_{\rm C})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}} + \frac{\hat{a}_{\rm C} - \hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}} + \Theta(\hat{a}_{\rm C}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X}) \right] \quad ; \tag{69}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm C}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm C}} \frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2 \right] [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 - \frac{\hat{a}_{\rm C}}{n-2} \frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} \Theta(\hat{a}_{\rm C}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X}) \quad ; \tag{70}$$

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = (\hat{a}_{\rm C})^2 \frac{c^4(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + (\hat{a}_{\rm Y})^4(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2(\hat{a}_{\rm Y})^2 c^2 \hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{(\hat{a}_{\rm Y})^2[4(\hat{a}_{\rm Y})^2 c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]} \quad ; \tag{71}$$

$$(\hat{\sigma}_{\hat{b}_{\rm C}})^2 = \frac{\hat{a}_{\rm C}}{n} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}} + \frac{\hat{a}_{\rm C} - \hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}} \right] \frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2 (\hat{\sigma}_{\hat{a}'_{\rm C}})^2$$

$$- \frac{2}{n} \widetilde{X} (\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}}) \quad ; \tag{72}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{(\widetilde{w_x})_{00}}{n} \frac{(\widetilde{w_x})_{13} + \hat{a}_{\rm Y}\hat{a}_{\rm X}(\widetilde{w_x})_{31} - (\hat{a}_{\rm Y} + \hat{a}_{\rm X})(\widetilde{w_x})_{22}}{(\widetilde{w_x})_{20}(\widetilde{w_x})_{11}} \quad ; \tag{73}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C} c^2}{\hat{a}_{\rm Y}[4(\hat{a}_{\rm Y})^2 c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]^{1/2}}$$

$$\times \frac{(\widetilde{w_x})_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm C}(\widetilde{w_x})_{30} - (\hat{a}_{\rm Y} + \hat{a}_{\rm C})(\widetilde{w_x})_{21}}{(\widetilde{w_x})_{20}} \quad ; \tag{74}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C}\hat{a}_{\rm Y}}{[4(\hat{a}_{\rm Y})^2c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]^{1/2}}$$
$$\times \frac{(\widetilde{w_x})_{03} + \hat{a}_{\rm X}\hat{a}_{\rm C}(\widetilde{w_x})_{21} - (\hat{a}_{\rm X} + \hat{a}_{\rm C})(\widetilde{w_x})_{12}}{(\widetilde{w_x})_{11}} \quad ; \tag{75}$$

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = (\hat{a}_{\rm C})^2$$
$$\times \frac{(\hat{a}_{\rm C})^4(A_{\rm XC})^2(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + (\hat{a}_{\rm Y})^4(A_{\rm CY'})^2(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY'}\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{(\hat{a}_{\rm Y})^2\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY'} + [\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY'} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}}; \tag{76}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm C}} = \frac{(\hat{a}_{\rm C})^3A_{\rm XC}}{\hat{a}_{\rm Y}\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY'} + [(\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY'} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}^{1/2}}$$
$$\times \frac{(\widetilde{w_x})_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm C}(\widetilde{w_x})_{30} - (\hat{a}_{\rm Y} + \hat{a}_{\rm C})(\widetilde{w_x})_{21}}{(\widetilde{w_x})_{20}} \quad ; \tag{77}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm C}} = \frac{\hat{a}_{\rm C}\hat{a}_{\rm Y}A_{\rm CY'}}{\{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2A_{\rm XC}A_{\rm CY'} + [(\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY'} - (\hat{a}_{\rm C})^2A_{\rm XC}]^2\}^{1/2}}$$
$$\times \frac{(\widetilde{w_x})_{03} + \hat{a}_{\rm X}\hat{a}_{\rm C}(\widetilde{w_x})_{21} - (\hat{a}_{\rm X} + \hat{a}_{\rm C})(\widetilde{w_x})_{12}}{(\widetilde{w_x})_{11}} \quad ; \tag{78}$$

and, in addition:

$$(\hat{\sigma}_{\hat{a}'_{\rm C}})^2 = (\hat{a}_{\rm C})^2\frac{c^4(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2 + (\hat{a}_{\rm Y})^4(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2(\hat{a}_{\rm Y})^2c^2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{(\hat{a}_{\rm Y})^2[4(\hat{a}_{\rm Y})^2c^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - c^2)^2]} \quad ; \tag{79}$$

$$(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2 = \frac{(\widetilde{w_x})_{00}}{n}\frac{(\widetilde{w_x})_{22} + (\hat{a}_{\rm Y})^2(\widetilde{w_x})_{40} - 2\hat{a}_{\rm Y}(\widetilde{w_x})_{31}}{[(\widetilde{w_x})_{20}]^2} \quad ; \tag{80}$$

where $\hat{a}_{\rm Y} = (\widetilde{w_y})_{11}/(\widetilde{w_y})_{20}$; $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$; $\hat{a}_{\rm X} = (\widetilde{w_x})_{02}/(\widetilde{w_x})_{11}$; $A_{\rm CY'} = \hat{a}_{\rm C}/\hat{a}'_{\rm Y} - 1$; R is defined in A, and Θ is formulated in terms of $n(\widetilde{w_x})_{pq}/(\widetilde{w_x})_{00}$ instead of S_{pq}.

In the special case of normal and data-independent residuals, $\Theta(\hat{a}_{\rm C}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X}) \to 0$, Eqs. (71), (72), must necessarily reduce to (69), (70), respectively, which implies an additional factor, $n/(n-2)$, in the first term on the right-hand side of Eqs. (34), (48), and (72)-(75).

In absence of a rigorous proof, Eqs. (69)-(75) must be considered as approximate results.

2.6 Reduced major-axis regression

The reduced major-axis regression may be considered as a special case

of oblique regression, where $c^2 = a_{\rm X} a_{\rm Y}$. Accordingly, Eqs. (51) and (52) also hold.

For homoscedastic data, $w_{x_i} = w_x$, $w_{y_i} = w_y$, $1 \leq i \leq n$, the regression line slope and intercept estimators, via Eqs. (54) and (55) are:

$$\hat{a}_{\rm R} = \mp\sqrt{\frac{S_{02}}{S_{20}}} = \mp\sqrt{\hat{a}_{\rm X}\hat{a}_{\rm Y}} \quad ; \tag{81}$$

$$\hat{b}_{\rm R} = \overline{Y} - \hat{a}_{\rm R}\overline{X} \quad ; \tag{82}$$

where the index, R, denotes reduced major-axis regression, $\hat{a}_{\rm Y} = S_{11}/S_{20}$; $\hat{a}_{\rm X} = S_{02}/S_{11}$; and the double sign corresponds to the solutions of the square root, where the parasite solution must be disregarded. Accordingly, related models shall be quoted as R models. For further details refer to an earlier attempt [7].

The regression line slope and intercept variance estimators may be directly inferred from Eqs. (56), (57), for normal residuals, in the limit, $\hat{a}_{\rm C} \to \hat{a}_{\rm R} = \sqrt{\hat{a}_{\rm X}\hat{a}_{\rm Y}}$. The result is:

$$\begin{aligned}
[(\hat{\sigma}_{\hat{a}_{\rm R}})_{\rm N}]^2 &= \frac{(\hat{a}_{\rm R})^2}{n-2}\left[\frac{(n-2)R_{\rm R}}{\hat{a}_{\rm R}S_{11}} + \Theta(\hat{a}_{\rm R},\hat{a}_{\rm Y},\hat{a}_{\rm X})\right] \\
&= \frac{(\hat{a}_{\rm R})^2}{n-2}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm R}}{\hat{a}_{\rm R}} + \frac{\hat{a}_{\rm R}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm R},\hat{a}_{\rm Y},\hat{a}_{\rm X})\right] \quad ;
\end{aligned} \tag{83}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm R}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm R}}\frac{S_{11}}{S_{00}} + (\overline{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm R}})_{\rm N}]^2 - \frac{\hat{a}_{\rm R}}{n-2}\frac{S_{11}}{S_{00}}\Theta(\hat{a}_{\rm R},\hat{a}_{\rm Y},\hat{a}_{\rm X}) \quad ; \tag{84}$$

and for non normal residuals the application of the δ-method yields [14]:

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = (\hat{a}_{\rm R})^2\left[\frac{1}{4}\frac{(\hat{\sigma}_{\hat{a}_{\rm Y}})^2}{(\hat{a}_{\rm Y})^2} + \frac{1}{4}\frac{(\hat{\sigma}_{\hat{a}_{\rm X}})^2}{(\hat{a}_{\rm X})^2} + \frac{1}{2}\frac{\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right] \quad ; \tag{85}$$

$$\begin{aligned}
(\hat{\sigma}_{\hat{b}_{\rm R}})^2 &= \frac{\hat{a}_{\rm R}}{n}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm R}}{\hat{a}_{\rm R}} + \frac{\hat{a}_{\rm R}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2(\hat{\sigma}_{\hat{a}_{\rm R}})^2 \\
&\quad - \frac{2}{n}\overline{X}(\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm R}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm R}}) \quad ;
\end{aligned} \tag{86}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm R}} = \frac{1}{2}\left(\frac{\hat{a}_{\rm X}}{\hat{a}_{\rm Y}}\right)^{1/2}\frac{S_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm R}S_{30} - (\hat{a}_{\rm Y}+\hat{a}_{\rm R})S_{21}}{S_{20}} \quad ; \tag{87}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm R}} = \frac{1}{2}\left(\frac{\hat{a}_{\rm Y}}{\hat{a}_{\rm X}}\right)^{1/2}\frac{S_{03} + \hat{a}_{\rm X}\hat{a}_{\rm R}S_{21} - (\hat{a}_{\rm X}+\hat{a}_{\rm R})S_{12}}{S_{11}} \quad ; \tag{88}$$

where $\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}$ is defined by Eq. (60) and Eqs. (85), (86), are equivalent to their counterparts expressed in the parent paper [14]. For further details refer to D.

The extension of the above results to heteroscedastic data via Eqs. (81)-(88) reads:

$$\hat{a}_{\rm R} = \mp\sqrt{\frac{(\widetilde{w_x})_{02}}{(\widetilde{w_x})_{20}}} = \mp\sqrt{\hat{a}_{\rm X}\hat{a}'_{\rm Y}} \quad ; \tag{89}$$

$$\hat{b}_{\rm R} = \widetilde{Y} - \hat{a}_{\rm R}\widetilde{X} \quad ; \tag{90}$$

$$\begin{aligned}[\hat{\sigma}_{\hat{a}_{\rm R}})_{\rm N}]^2 &= \frac{(\hat{a}_{\rm R})^2}{n-2}\left[\frac{n-2}{n}\frac{R_{\rm R}}{\hat{a}_{\rm R}}\frac{(\widetilde{w_x})_{00}}{(\widetilde{w_x})_{11}} + \Theta(\hat{a}_{\rm R}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X})\right] \\ &= \frac{(\hat{a}_{\rm R})^2}{n-2}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm R}}{\hat{a}_{\rm R}} + \frac{\hat{a}_{\rm R}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}} + \Theta(\hat{a}_{\rm R}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X})\right] \quad ; \end{aligned} \tag{91}$$

$$[\hat{\sigma}_{\hat{b}_{\rm R}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm R}}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm R}})_{\rm N}]^2 - \frac{\hat{a}_{\rm R}}{n-2}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}}\Theta(\hat{a}_{\rm R}, \hat{a}'_{\rm Y}, \hat{a}_{\rm X}) \quad ; \tag{92}$$

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = (\hat{a}_{\rm R})^2\frac{\hat{a}_{\rm Y}}{\hat{a}'_{\rm Y}}\left[\frac{1}{4}\frac{(\hat{\sigma}_{\hat{a}_{\rm Y}})^2}{(\hat{a}_{\rm Y})^2} + \frac{1}{4}\frac{(\hat{\sigma}_{\hat{a}_{\rm X}})^2}{(\hat{a}_{\rm X})^2} + \frac{1}{2}\frac{\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}}{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right] \quad ; \tag{93}$$

$$\begin{aligned}(\hat{\sigma}_{\hat{b}_{\rm R}})^2 &= \frac{\hat{a}_{\rm R}}{n}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm R}}{\hat{a}_{\rm R}} + \frac{\hat{a}_{\rm R}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}}\right]\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2(\hat{\sigma}_{\hat{a}'_{\rm R}})^2 \\ &\quad - \frac{2}{n}\widetilde{X}(\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm R}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm R}}) \quad ; \end{aligned} \tag{94}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm R}} = \frac{1}{2}\left(\frac{\hat{a}_{\rm X}}{\hat{a}_{\rm Y}}\right)^{1/2}\frac{(\widetilde{w_x})_{12} + \hat{a}_{\rm Y}\hat{a}_{\rm R}(\widetilde{w_x})_{30} - (\hat{a}_{\rm Y}+\hat{a}_{\rm R})(\widetilde{w_x})_{21}}{(\widetilde{w_x})_{20}} \quad ; \tag{95}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm R}} = \frac{1}{2}\left(\frac{\hat{a}_{\rm Y}}{\hat{a}_{\rm X}}\right)^{1/2}\frac{(\widetilde{w_x})_{03} + \hat{a}_{\rm X}\hat{a}_{\rm R}(\widetilde{w_x})_{21} - (\hat{a}_{\rm X}+\hat{a}_{\rm R})(\widetilde{w_x})_{12}}{(\widetilde{w_x})_{11}} \quad ; \tag{96}$$

and, in addition:

$$(\hat{\sigma}_{\hat{a}'_{\rm R}})^2 = \frac{1}{4}\left[\frac{\hat{a}_{\rm X}}{\hat{a}_{\rm Y}}(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2 + \frac{\hat{a}_{\rm Y}}{\hat{a}_{\rm X}}(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right] \quad ; \tag{97}$$

where $\hat{a}_{\rm Y} = (\widetilde{w_y})_{11}/(\widetilde{w_y})_{20}$; $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$; $\hat{a}_{\rm X} = (\widetilde{w_x})_{02}/(\widetilde{w_x})_{11}$; R is defined in A, $\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}$, $(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2$, are expressed by Eqs. (73), (80), respectively, and Θ is formulated in terms of $n(\widetilde{w_x})_{pq}/(\widetilde{w_x})_{00}$ instead of S_{pq}.

In absence of a rigorous proof, Eqs. (91)-(96) must be considered as approximate results.

2.7 Bisector regression

The bisector regression implies use of both Y and X models for determining the angle formed by related regression lines. The bisecting line is assumed to be the estimated regression line of the model.

Let $\alpha_{\rm Y}$, $\alpha_{\rm X}$, $\alpha_{\rm B}$, be the angles formed between Y, X, B, regression line, respectively, and x axis, and γ the angle formed between Y and X regression lines, as outlined in Fig. 1. Accordingly, $\gamma/2$ is the angle formed between Y or X and B regression lines.

The following relations can easily be deduced from Fig. 1: $\alpha_{\rm X} = \alpha_{\rm Y} + \gamma$; $\alpha_{\rm B} = \alpha_{\rm Y} + \gamma/2 = (\alpha_{\rm Y} + \alpha_{\rm X})/2$, and the dimensionless slope of the regression line is $\tan\alpha_{\rm B}$. Using the trigonometric formulae:

$$\tan(u+v) = \frac{\tan u + \tan v}{1 - \tan u \tan v} \quad ; \qquad \tan\frac{u}{2} = \frac{\sin u}{1+\cos u} \quad ;$$

and the identity:

$$\frac{X(1+S_{\rm Y}) + Y(1+S_{\rm X})}{(1+S_{\rm X})(1+S_{\rm Y}) - XY} = \frac{XY - 1 + S_{\rm X}S_{\rm Y}}{X+Y} \quad ;$$

$$X = \frac{a_{\rm X}}{a_u} \quad ; \quad Y = \frac{a_{\rm Y}}{a_u} \quad ; \quad S_{\rm X} = \sqrt{1+X^2} \quad ; \quad S_{\rm Y} = \sqrt{1+Y^2} \quad ;$$

the regression line slope estimator, after some algebra, is expressed as [14]:

$$\hat{a}_{\rm B} = \frac{\hat{a}_{\rm Y}\hat{a}_{\rm X} - a_u^2 + \sqrt{a_u^2 + (\hat{a}_{\rm Y})^2}\sqrt{a_u^2 + (\hat{a}_{\rm X})^2}}{\hat{a}_{\rm Y} + \hat{a}_{\rm X}} \quad ; \tag{98}$$

where a_u is the unit slope, $\hat{a}_{\rm Y} = S_{11}/S_{20}$, $\hat{a}_{\rm X} = S_{02}/S_{11}$, and the regression line intercept estimator reads [14]:

$$\hat{b}_{\rm B} = \overline{Y} - \hat{a}_{\rm B}\overline{X} \quad ; \tag{99}$$

where the index, B, denotes bisector regression.

The bisector regression may be considered as a special case of oblique regression where the variance ratio, c^2, is deduced from the combination of Eqs. (54) and (98), requiring $a_{\rm C} = a_{\rm B}$. After a lot of algebra involving the roots of a second-degree equation, the result is:

$$c^2 = (a_{\rm B})^2 \frac{a_{\rm X} \mp a_{\rm B}}{a_{\rm B}} \left(\frac{a_{\rm B} \mp a_{\rm Y}}{a_{\rm Y}}\right)^{-1} \quad ; \tag{100}$$

where the parasite solution must be disregarded. Accordingly, Eqs. (51) and (52) also hold.

For normal residuals and homoscedastic data, the regression line slope and intercept variance estimators may be directly inferred from Eqs. (56) and (57) in the limit, $\hat{a}_{\rm C} \to \hat{a}_{\rm B}$. The result is:

$$[(\hat{\sigma}_{\hat{a}_{\rm B}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm B})^2}{n-2}\left[\frac{(n-2)R_{\rm B}}{\hat{a}_{\rm B}S_{11}} + \Theta(\hat{a}_{\rm B},\hat{a}_{\rm Y},\hat{a}_{\rm X})\right]$$
$$= \frac{(\hat{a}_{\rm B})^2}{n-2}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm B}}{\hat{a}_{\rm B}} + \frac{\hat{a}_{\rm B}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm B},\hat{a}_{\rm Y},\hat{a}_{\rm X})\right] \quad ; \tag{101}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm B}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm B}}\frac{S_{11}}{S_{00}} + (\overline{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm B}})_{\rm N}]^2 - \frac{\hat{a}_{\rm B}}{n-2}\frac{S_{11}}{S_{00}}\Theta(\hat{a}_{\rm B},\hat{a}_{\rm Y},\hat{a}_{\rm X}) \quad ; \tag{102}$$

and for non normal residuals the application of the δ-method yields [14]:

$$(\hat{\sigma}_{\hat{a}_{\rm B}})^2 = \frac{(\hat{a}_{\rm B})^2}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})^2}\left[\frac{a_u^2+(\hat{a}_{\rm X})^2}{a_u^2+(\hat{a}_{\rm Y})^2}(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + \frac{a_u^2+(\hat{a}_{\rm Y})^2}{a_u^2+(\hat{a}_{\rm X})^2}(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right] ; \tag{103}$$

$$(\hat{\sigma}_{\hat{b}_{\rm B}})^2 = \frac{\hat{a}_{\rm B}}{n}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm B}}{\hat{a}_{\rm B}} + \frac{\hat{a}_{\rm B}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2(\hat{\sigma}_{\hat{a}_{\rm B}})^2$$
$$- \frac{2}{n}\overline{X}(\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm B}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm B}}) \quad ; \tag{104}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm B}} = \frac{\hat{a}_{\rm B}\sqrt{a_u^2+(\hat{a}_{\rm X})^2}}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})\sqrt{a_u^2+(\hat{a}_{\rm Y})^2}}\frac{S_{12}+\hat{a}_{\rm Y}\hat{a}_{\rm B}S_{30}-(\hat{a}_{\rm Y}+\hat{a}_{\rm B})S_{21}}{S_{20}} \quad ; \tag{105}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm B}} = \frac{\hat{a}_{\rm B}\sqrt{a_u^2+(\hat{a}_{\rm Y})^2}}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})\sqrt{a_u^2+(\hat{a}_{\rm X})^2}}\frac{S_{03}+\hat{a}_{\rm X}\hat{a}_{\rm B}S_{21}-(\hat{a}_{\rm X}+\hat{a}_{\rm B})S_{12}}{S_{11}} \quad ; \tag{106}$$

where $\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}$ is defined by Eq. (60) and Eqs. (103), (104), are equivalent to their counterparts expressed in the parent paper [14]. For further details refer to D.

For heteroscedastic data, the combination of Eqs. (67) and (98), requiring $a_{\rm C} = a_{\rm B}$, after a lot of algebra involving the roots of a second-degree equation, yields:

$$c^2 = (a_{\rm B})^2\frac{a_{\rm X} \mp a_{\rm B}}{a_{\rm B}}\left(\frac{a_{\rm B} \mp a'_{\rm Y}}{a'_{\rm Y}}\right)^{-1} \quad ; \tag{107}$$

where $\hat{a}_{\rm X} = (\widetilde{w_x})_{02}/(\widetilde{w_x})_{11}$; $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$; and the parasite solution must be disregarded. Accordingly, Eqs. (51) and (52) also hold.

The extension of the above results to heteroscedastic data via Eqs. (98)-(99) and (101)-(106) reads:

$$\hat{b}_{\rm B} = \widetilde{Y} - \hat{a}_{\rm B}\widetilde{X} \ \ ; \tag{108}$$

$$[(\hat{\sigma}_{\hat{a}_{\rm B}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm B})^2}{n-2}\left[\frac{n-2}{n}\frac{R_{\rm B}}{\hat{a}_{\rm B}}\frac{(\widetilde{w_x})_{00}}{(\widetilde{w_x})_{11}} + \Theta(\hat{a}_{\rm B},\hat{a}'_{\rm Y},\hat{a}_{\rm X})\right]$$
$$= \frac{(\hat{a}_{\rm B})^2}{n-2}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm B}}{\hat{a}_{\rm B}} + \frac{\hat{a}_{\rm B}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}} + \Theta(\hat{a}_{\rm B},\hat{a}'_{\rm Y},\hat{a}_{\rm X})\right] \ \ ; \tag{109}$$

$$[(\hat{\sigma}_{\hat{b}_{\rm B}})_{\rm N}]^2 = \left[\frac{1}{\hat{a}_{\rm B}}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2\right][(\hat{\sigma}_{\hat{a}_{\rm B}})_{\rm N}]^2 - \frac{\hat{a}_{\rm B}}{n-2}\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}}\Theta(\hat{a}_{\rm B},\hat{a}'_{\rm Y},\hat{a}_{\rm X}); \tag{110}$$

$$(\hat{\sigma}_{\hat{a}_{\rm B}})^2 = \frac{(\hat{a}_{\rm B})^2}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})^2}\left[\frac{a_u^2+(\hat{a}_{\rm X})^2}{a_u^2+(\hat{a}_{\rm Y})^2}(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 + \frac{a_u^2+(\hat{a}_{\rm Y})^2}{a_u^2+(\hat{a}_{\rm X})^2}(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right]; \tag{111}$$

$$(\hat{\sigma}_{\hat{b}_{\rm B}})^2 = \frac{\hat{a}_{\rm B}}{n}\left[\frac{\hat{a}_{\rm X}-\hat{a}_{\rm B}}{\hat{a}_{\rm B}} + \frac{\hat{a}_{\rm B}-\hat{a}'_{\rm Y}}{\hat{a}'_{\rm Y}}\right]\frac{(\widetilde{w_x})_{11}}{(\widetilde{w_x})_{00}} + (\widetilde{X})^2(\hat{\sigma}_{\hat{a}'_{\rm B}})^2$$
$$- \frac{2}{n}\widetilde{X}(\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm B}} + \hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm B}}) \ \ ; \tag{112}$$

$$\hat{\sigma}_{\hat{b}_{\rm Y}\hat{a}_{\rm B}} = \frac{\hat{a}_{\rm B}\sqrt{a_u^2+(\hat{a}_{\rm X})^2}}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})\sqrt{a_u^2+(\hat{a}_{\rm Y})^2}}\frac{(\widetilde{w_x})_{12}+\hat{a}_{\rm Y}\hat{a}_{\rm B}(\widetilde{w_x})_{30}-(\hat{a}_{\rm Y}+\hat{a}_{\rm B})(\widetilde{w_x})_{21}}{(\widetilde{w_x})_{20}}; \tag{113}$$

$$\hat{\sigma}_{\hat{b}_{\rm X}\hat{a}_{\rm B}} = \frac{\hat{a}_{\rm B}\sqrt{a_u^2+(\hat{a}_{\rm Y})^2}}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})\sqrt{a_u^2+(\hat{a}_{\rm X})^2}}\frac{(\widetilde{w_x})_{03}+\hat{a}_{\rm X}\hat{a}_{\rm B}(\widetilde{w_x})_{21}-(\hat{a}_{\rm X}+\hat{a}_{\rm B})(\widetilde{w_x})_{12}}{(\widetilde{w_x})_{11}}; \tag{114}$$

and, in addition:

$$(\hat{\sigma}_{\hat{a}'_{\rm B}})^2 = \frac{(\hat{a}_{\rm B})^2}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})^2}\left[\frac{a_u^2+(\hat{a}_{\rm X})^2}{a_u^2+(\hat{a}_{\rm Y})^2}(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2 + \frac{a_u^2+(\hat{a}_{\rm Y})^2}{a_u^2+(\hat{a}_{\rm X})^2}(\hat{\sigma}_{\hat{a}_{\rm X}})^2 + 2\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}\right]; \tag{115}$$

where $\hat{a}_{\rm Y} = (\widetilde{w_y})_{11}/(\widetilde{w_y})_{20}$, $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$, $\hat{a}_{\rm X} = (\widetilde{w_x})_{02}/(\widetilde{w_x})_{11}$, R is defined in A, $\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}}$, $(\hat{\sigma}_{\hat{a}'_{\rm Y}})^2$, are expressed by Eqs. (73), (80), respectively, and Θ is formulated in terms of $n(\widetilde{w_x})_{pq}/(\widetilde{w_x})_{00}$ instead of S_{pq}.

In absence of a rigorous proof, Eqs. (109)-(114) must be considered as approximate results.

2.8 Extension to structural models

A nontrivial question is to what extent the above results, valid for extreme structural models, can be extended to generic structural models. In

general, assumptions related to generic structural models are different from their counterparts related to extreme structural models e.g., [3], [4] Chap. 6, §6.4.5, but, on the other hand, they could coincide for a special subclass.

In any case, whatever different assumptions and models can be made with regard to generic and extreme structural models, results from the former are expected to tend to their counterparts from the latter when the instrumental scatter is negligible with respect to the intrinsic scatter. It is worth noticing that most work on linear regression by astronomers involves the situation where both intrinsic scatter and heteroscedastic data are present e.g., [1, 22, 15, 16].

A special subclass of structural models with normal residuals can be defined where, for a selected regression estimator, the regression line slope and intercept variance estimators are independent of the amount of instrumental and intrinsic scatter, including the limit of null intrinsic scatter (functional models) and null instrumental scatter (extreme structural models). More specifically, the dependence occurs only via the total (instrumental + intrinsic) scatter. In this view, the whole subclass of structural models under consideration could be related to functional modelling [8] Chap. 2, §2.1. For further details refer to the parent paper [7].

3 An example of astronomical application

3.1 Astronomical introduction

Heavy elements are synthesised within stars and (partially or totally) returned to the interstellar medium via supernovae. In an ideal situation where the initial stellar mass function (including binary and multiple systems) is universal and the gas returned after star death is instantaneously and uniformly mixed with the interstellar medium, the abundance ratio of primary elements produced mainly by large-mass ($m \gtrsim 8\mathrm{m}_\odot$, where $\mathrm{m}_\odot$ is the solar mass) stars maintains unchanged, which implies a linear relation. This is why large-mass stars have a short lifetime with respect to the age of the universe, and related ejecta (due to type II supernovae) may be considered as instantaneously returned to the interstellar medium.

A linear relation also holds if low-mass ($m \lesssim 8\mathrm{m}_\odot$) stars are considered, where the stellar lifetime can no longer be neglected with respect to the age

of the universe. Close binary systems including a white dwarf with masses, $m_{\rm WD} + m_{\rm C} > m_{\rm Ch}$, are (type Ia) supernovae progenitors, where $m_{\rm WD}$ is the white dwarf mass, $m_{\rm C}$ is the companion mass, and $m_{\rm Ch} \approx 1.44 {\rm m}_{\odot}$ is the Chandrasekhar upper mass limit for stable white dwarfs. An additional restriction, for a linear relation between two generic primary elements in the interstellar medium, is a constant number ratio of type II to type Ia supernovae at any epoch. For further details refer to F.

Restricting to iron and oxygen, the generic linear relation, Eq. (227), reads:

$$[{\rm O/H}]^{*} = a[{\rm Fe/H}]^{*} + b \quad ; \tag{116}$$

where [O/H], [Fe/H], are logarithmic number abundances normalized to the solar value e.g., [5] and the asterisks denote the ideal situation. More specifically, oxygen and iron abundance determinations performed on ideal stars by use of ideal instruments yield coordinates of points lying on the straight line defined by Eq. (116).

The intrinsic dispersion outside or along the ideal regression line may be owing to several processes, such as fluctuations in the stellar initial mass function (including binary and multiple systems) and inhomogeneous mixing of stellar ejecta with the interstellar medium, at different rates for different elements. Accordingly, ideal points, $\mathsf{P}_i^{*} \equiv ([{\rm Fe/H}]_i^{*}, [{\rm O/H}]_i^{*})$, are shifted towards actual points, $\mathsf{P}_{{\rm S}i} \equiv ([{\rm Fe/H}]_{{\rm S}i}, [{\rm O/H}]_{{\rm S}i})$.

More specifically, coeval ideal stars are represented by a single point on the ideal regression line, while related actual stars correspond to points which, in general, are shifted to a different extent outside or along the ideal regression line. Conversely, stars with different age could be represented by a same actual point, $\mathsf{P}_{{\rm S}i}$. The occurrence of instrumental scatter, related to iron and oxygen abundance determination on a star sample, makes actual points, $\mathsf{P}_{{\rm S}i}$, be shifted towards observed points, $\mathsf{P}_i \equiv ([{\rm Fe/H}]_i, [{\rm O/H}]_i)$.

With regard to the ideal regression line, there is a one-to-one correspondence between the coordinates, [Fe/H] and [O/H], while the contrary holds for actual points and observed points. In the limit of extreme structural models, where instrumental scatter is negligible with respect to intrinsic scatter, observed points are very close to actual points (if otherwise, any linear dependence would be hidden). The latter, to a first extent, may be determined along the following steps.

(1) Estimate a plausible regression line.

(2) Calculate the mean distance of observed points from the estimated regression line, parallel to each coordinate axis.

(3) Subdivide each coordinate axis into bins of width equal to the related mean distance calculated in (2).

(4) Evaluate the intrinsic scatter within each bin, using the method described in an earlier attempt [1].

(5) Minimize the loss function and determine the regression line slope and intercept estimators.

(6) Verify the absolute difference between previous and current regression line slope and intercept estimators is less than a previously assigned tolerance value. If otherwise, return to (2) taking into consideration the current estimated regression line.

In general, the total scatter, $\sigma^2_{[\mathrm{W/H}]_i} = \sigma^2_{[\mathrm{W/H}]_{\mathrm{F}i}} + \sigma^2_{[\mathrm{W/H}]_{\mathrm{S}i}}$, should be used for evaluating the weights, w_{x_i}, w_{y_i}, $1 \leq i \leq n$, appearing in the sum of the squared residuals, expressed by Eq. (8a), which implies the knowledge of the instrumental covariance matrix e.g., [1, 16].

3.2 Statistical results

An astronomical application performed in an earlier attempt [7] with regard to functional models, shall be repeated here for extreme structural models. Accordingly, related samples will be left unchanged but with the additional assumptions: (i) the intrinsic scatter is dominant with respect to the instrumental scatter, and (ii) uncertainties mentioned in the parent papers and reported below are related to the intrinsic scatter.

More specifically, the following samples related to the [O/H]-[Fe/H] relation shall be considered: RB09, $n = 49$, heteroscedastic data [20]; Fa09, $n = 44$, homoscedastic data with three different [O/H] determinations, namely LTE (standard local thermodynamical equilibrium for one-dimensional hydrostatic model atmospheres), SH0 (three-dimensional hydrostatic model atmospheres in absence of LTE with no account taken of the inelastic collisions via neutral H atoms, $S_{\mathrm{H}} = 0$), SH1 (three-dimensional hydrostatic model atmospheres in absence of LTE with due account taken of the inelastic collisions via neutral H atoms, $S_{\mathrm{H}} = 1$) [9]; Sa09, $n = 63$, heteroscedastic data

[21]. For further details refer to the parent paper [6]. In any case, [Fe/H] and [O/H] are determined independently for each sample star.

The [O/H]-[Fe/H] empirical relations are interpolated using the regression models, G, Y, X, O, R, B, for heteroscedastic data (FB09 and Sa09 samples) and Y, X, O, R, B, for homoscedastic data (Fa09 sample, cases LTE, SH0, SH1) and heteroscedastic data where intrinsic scatters are taken equal to the typical uncertainties mentioned in the parent papers (FB09, Sa09), $\sigma_{[\mathrm{Fe/H}]} = 0.15$, $\sigma_{[\mathrm{O/H}]} = 0.15$, for both FB09 and Sa09 samples. Model G relates to a general case where the slope and intercept estimators are determined via Eqs. (22) and (21), respectively. For further details refer to the parent papers [23, 25, 7]. Slope and intercept estimators together with related dispersion estimators are listed in Tables 2, 3, and 4, 5, for heteroscedastic and homoscedastic data, respectively.

Owing to high difficulties intrinsic to the determination of slope and intercept dispersion estimators for G models, related calculations were not performed, leaving only approximate expressions [23] and asymptotic formulae [B, Eq. (160) related to G models]. For the remaining models, the regression line slope and intercept estimators and related dispersion estimators are calculated using Eqs. (23)-(29) and (30)-(36), case Y, homoscedastic and heteroscedastic data, respectively; Eqs. (37)-(43) and (44)-(50), case X, homoscedastic and heteroscedastic data, respectively; Eqs. (54)-(62) and (67)-(75), $c^2 = 1$, case O, homoscedastic and heteroscedastic data, respectively; Eqs. (81)-(88) and (89)-(96), case R, homoscedastic and heteroscedastic data, respectively; Eqs. (98)-(106) and (108)-(114), case B, homoscedastic and heteroscedastic data, respectively.

The regression lines determined by use of the above mentioned methods are plotted in Figs. 2 and 3 for heteroscedastic and homoscedastic data, respectively, where sample denomination and population are indicated on each panel together with model captions. Homoscedastic data are conceived as a special case of heteroscedastic data in Fig. 2 to test the computer code, which is different for heteroscedastic and homoscedastic data. It can be seen that lower panels of Figs. 2 and 3 coincide, and the regression lines related to models G and O in lower panels of Figs. 2 also coincide, as expected. The whole set of regression lines for all methods and all samples is shown in the upper right panel of Figs. 2 and 3.

Regression line slope and intercept estimators have the same expression for both structural and funcional models. Accordingly, Figs. 2 and 3 maintain unchanged with respect to their counteparts shown in an earlier attempt [7]

Table 2: Regression line slope estimators, $\hat{a}$, and related dispersion estimators, $\hat{\sigma}_{\hat{a}}$, for heteroscedastic models, G, Y, X, O, R, B, applied to the [O/H]-[Fe/H] empirical relation deduced from the following samples (from up to down): RB09, Sa09. Dispersion column captions: ENNR - extreme structural models with non normal residuals [14]; ENRR - extreme structural models with normal residuals [10]; FNRR - functional models with normal residuals [23, 25, 7]; YANR - approximate formula for normal residuals [23, 25, 7]; AFNR - asymptotic formula for normal residuals [B, Eq. (160) related to the appropriate model]. For G models, exact expressions of slope estimators were not evaluated in the present attempt. For Y models and normal residuals, different slope dispersion estimators yield coinciding values, as expected.

m	$\hat{a}$			$\hat{\sigma}_{\hat{a}}$			sample
		ENNR	ENRR	FNRR	YANR	AFNR	
G	0.7279				0.0294	0.0288	RB09
Y	0.6714	0.0302	0.0314	0.0314	0.0314	0.0314	
X	0.7305	0.0271	0.0290	0.0290	0.0279	0.0290	
O	0.6964	0.0277	0.0276	0.0278	0.0271	0.0274	
R	0.7050	0.0254	0.0280	0.0282	0.0272	0.0277	
B	0.7005	0.0254	0.0278	0.0280	0.0271	0.0275	
G	0.6383				0.0435	0.0582	Sa09
Y	0.6167	0.0810	0.0398	0.0398	0.0398	0.0398	
X	0.8652	0.0772	0.0833	0.0829	0.0664	0.0829	
O	0.6355	0.0753	0.0609	0.0637	0.0541	0.0580	
R	0.6927	0.0677	0.0664	0.0700	0.0560	0.0626	
B	0.7336	0.0662	0.0704	0.0738	0.0579	0.0666	

Table 3: Regression line intercept estimators, $\hat{b}$, and related dispersion estimators, $\hat{\sigma}_{\hat{b}}$, for heteroscedastic models, G, Y, X, O, R, B, applied to the [O/H]-[Fe/H] empirical relation deduced from the following samples (from up to down): RB09, Sa09. Dispersion column captions: ENNR - extreme structural models with non normal residuals [14]; ENRR - extreme structural models with normal residuals [10]; FNRR - functional models with normal residuals [23, 25, 7]; YANR - approximate formula for normal residuals [23, 25, 7]; AFNR - asymptotic formula for normal residuals [via B, Eq. (160) related to the appropriate model]. For G models, exact expressions of intercept estimators were not evaluated in the present attempt. For Y models and normal residuals, different intercept dispersion estimators yield coinciding values, as expected.

m	$\hat{b}$	$\hat{\sigma}_{\hat{b}}$					sample
		ENNR	ENRR	FNRR	YANR	AFNR	
G	+0.0043				0.0672	0.0660	RB09
Y	−0.1121	0.0608	0.0675	0.0675	0.0675	0.0675	
X	+0.0316	0.0636	0.0736	0.0735	0.0712	0.0735	
O	−0.0512	0.0523	0.0702	0.0707	0.0689	0.0697	
R	−0.0305	0.0521	0.0710	0.0725	0.0693	0.0704	
B	−0.0413	0.0522	0.0706	0.0711	0.0691	0.0700	
G	+0.0619				0.0251	0.0336	Sa09
Y	+0.0439	0.0105	0.0198	0.0198	0.0198	0.0198	
X	+0.3080	0.0712	0.0676	0.0673	0.0575	0.0673	
O	+0.1461	0.0396	0.0509	0.0525	0.0469	0.0491	
R	+0.1864	0.0436	0.0547	0.0549	0.0485	0.0524	
B	+0.2153	0.0455	0.0575	0.0603	0.0501	0.0552	

Table 4: Regression line slope estimators, $\hat{a}$, and related dispersion estimators, $\hat{\sigma}_{\hat{a}}$, for homoscedastic models, Y, X, O, R, B, applied to the [O/H]-[Fe/H] empirical relation deduced from the following samples (from up to down): RB09, Sa09, Fa09, cases LTE, SH0, SH1. Dispersion column captions: ENNR - extreme structural models with non normal residuals [14]; ENRR - extreme structural models with normal residuals [10]; FNRR - functional models with normal residuals [23, 25, 7]; YANR - approximate formula for normal residuals [23, 25, 7]; AFNR - asymptotic formula for normal residuals [B, Eq. (160) related to the appropriate model]. For Y models and normal residuals, different slope dispersion estimators yield coinciding values, as expected.

m	$\hat{a}$	$\hat{\sigma}_{\hat{a}}$					sample
		ENNR	ENRR	FNRR	YANR	AFNR	
Y	0.6917	0.0268	0.0317	0.0317	0.0317	0.0317	RB09
X	0.7600	0.0326	0.0349	0.0348	0.0332	0.0348	
O	0.7143	0.0282	0.0327	0.0331	0.0319	0.0324	
R	0.7251	0.0278	0.0332	0.0336	0.0321	0.0328	
B	0.7253	0.0278	0.0333	0.0336	0.0321	0.0329	
Y	0.5868	0.0596	0.0461	0.0461	0.0461	0.0461	Sa09
X	0.8077	0.0563	0.0637	0.0635	0.0541	0.0635	
O	0.6476	0.0620	0.0509	0.0526	0.0468	0.0491	
R	0.6885	0.0523	0.0541	0.0562	0.0479	0.0519	
B	0.6916	0.0513	0.0544	0.0565	0.0480	0.0521	
Y	0.8961	0.0333	0.0303	0.0303	0.0303	0.0303	Fa09
X	0.9381	0.0294	0.0318	0.0317	0.0310	0.0317	(LTE)
O	0.9150	0.0319	0.0310	0.0311	0.0305	0.0308	
R	0.9168	0.0310	0.0310	0.0312	0.0305	0.0308	
B	0.9169	0.0310	0.0310	0.0312	0.0305	0.0308	
Y	1.2261	0.0459	0.0432	0.0432	0.0432	0.0432	Fa09
X	1.2884	0.0434	0.0454	0.0454	0.0443	0.0454	(SH0)
O	1.2640	0.0449	0.0445	0.0448	0.0436	0.0443	
R	1.2569	0.0441	0.0443	0.0445	0.0435	0.0440	
B	1.2568	0.0441	0.0443	0.0445	0.0435	0.0440	
Y	1.0492	0.0358	0.0341	0.0341	0.0341	0.0341	Fa09
X	1.0946	0.0315	0.0356	0.0356	0.0348	0.0356	(SH1)
O	1.0732	0.0337	0.0349	0.0350	0.0343	0.0347	
R	1.0716	0.0332	0.0348	0.0350	0.0343	0.0346	
B	1.0716	0.0332	0.0348	0.0350	0.0343	0.0346	

Table 5: Regression line intercept estimators, $\hat{b}$, and related dispersion estimators, $\hat{\sigma}_{\hat{b}}$, for homoscedastic models, Y, X, O, R, B, applied to the [O/H]-[Fe/H] empirical relation deduced from the following samples (from up to down): RB09, Sa09, Fa09, cases LTE, SH0, SH1. Dispersion column captions: ENNR - extreme structural models with non normal residuals [14]; ENRR - extreme structural models with normal residuals [10]; FNRR - functional models with normal residuals [23, 25, 7]; YANR - approximate formula for normal residuals [23, 25, 7]; AFNR - asymptotic formula for normal residuals [via B, Eq. (160) related to the appropriate model]. For Y models and normal residuals, different intercept dispersion estimators yield coinciding values, as expected.

m	$\hat{b}$	$\hat{\sigma}_{\hat{b}}$					sample
		ENNR	ENRR	FNRR	YANR	AFNR	
Y	−0.0766	0.0598	0.0737	0.0737	0.0737	0.0737	RB09
X	+0.0742	0.0811	0.0807	0.0806	0.0773	0.0806	
O	−0.0268	0.0658	0.0759	0.0766	0.0741	0.0752	
R	−0.0030	0.0664	0.0770	0.0778	0.0746	0.0762	
B	−0.0025	0.0665	0.0770	0.0778	0.0746	0.0762	
Y	+0.0908	0.0211	0.0338	0.0338	0.0338	0.0338	Sa09
X	+0.2011	0.0423	0.0431	0.0430	0.0397	0.0430	
O	+0.1212	0.0268	0.0357	0.0363	0.0343	0.0351	
R	+0.1416	0.0279	0.0373	0.0381	0.0352	0.0365	
B	+0.1431	0.0282	0.0375	0.0382	0.0352	0.0367	
Y	+0.5476	0.0761	0.0663	0.0663	0.0663	0.0663	Fa09
X	+0.6366	0.0665	0.0693	0.0693	0.0678	0.0693	(LTE)
O	+0.5877	0.0725	0.0676	0.0680	0.0666	0.0673	
R	+0.5916	0.0706	0.0678	0.0681	0.0667	0.0674	
B	+0.5916	0.0706	0.0678	0.0681	0.0667	0.0674	
Y	+0.8717	0.1017	0.0945	0.0945	0.0945	0.0945	Fa09
X	+1.0037	0.1003	0.0992	0.0991	0.0968	0.0991	(SH0)
O	+0.9519	0.1019	0.0973	0.0978	0.0953	0.0967	
R	+0.9369	0.0998	0.0967	0.0973	0.0950	0.0961	
B	+0.9367	0.0998	0.0967	0.0973	0.0950	0.0961	
Y	+0.6518	0.0808	0.0745	0.0745	0.0745	0.0745	Fa09
X	+0.7479	0.0730	0.0777	0.0777	0.0761	0.0777	(SH1)
O	+0.7027	0.0772	0.0762	0.0765	0.0750	0.0758	
R	+0.6993	0.0760	0.0761	0.0764	0.0750	0.0757	
B	+0.6993	0.0760	0.0761	0.0764	0.0749	0.0757	

where, on the other hand, B models were not included.

An inspection of Tables 2-5 and Figs. 2-3 discloses the following.

(1) Either of the inequalities [14]:

$$\hat{a}_{\rm Y} < \hat{a}_{\rm O} < \hat{a}_{\rm R} < \hat{a}_{\rm B} < \hat{a}_{\rm X} \quad ; \qquad \hat{a}_{\rm B} < a_u \quad ; \qquad S_{11} > 0 \quad ; \tag{117a}$$

$$\hat{a}_{\rm Y} < \hat{a}_{\rm B} < \hat{a}_{\rm R} < \hat{a}_{\rm O} < \hat{a}_{\rm X} \quad ; \qquad \hat{a}_{\rm B} > a_u \quad ; \qquad S_{11} > 0 \quad ; \tag{117b}$$

where a_u is the unit slope, is satisfied for homoscedastic data but the contrary holds for heteroscedastic data. In particular, $\hat{a}_{\rm B} < \hat{a}_{\rm R} < a_u$ for RB09 sample, see Table 2. In addition, $\hat{a}_{\rm Y} < \hat{a}_{\rm G} < \hat{a}_{\rm X}$ for heteroscedastic data, but a counterexample is provided in an earlier attempt [23].

(2) Slope and intercept estimators from O, R and B models are in agreement within $\mp\sigma$. The extension of the above result to slope and intercept estimators from Y and X models holds for samples with lower dispersion (Fa09). An increasing dispersion yields marginal (RB09) or no (Sa09) agreement within $\mp\sigma$, for both heteroscedastic and homoscedastic data.

(3) For normal residuals, slope and intercept dispersion estimators related to functional and structural models yield slightly different results, as expected from the fact that related asymptotic formulae coincide [B, Eq. (160) related to the appropriate model]. Asymptotic formulae used in the current attempt make a better fit with respect to earlier approximations [23, 25, 7].

(4) Systematic variations due to different sample data are dominant with respect to the intrinsic scatter.

In conclusion, regression lines deduced from different sample data represent correct (from the standpoint of regression models considered in the current attempt) [O/H]-[Fe/H] relations, but no definitive choice can be made until systematic errors due to different methods and/or spectral lines in determining oxygen abundance, are alleviated.

4 Discussion

For an assigned sample, structural models belonging to a special subclass are indistinguishable from extreme structural models, as outlined in an earlier

attempt [7]. Accordingly, the results of the current paper also apply to structural models of the kind considered. The expression of regression line slope and intercept estimators and related variance estimators in terms of weighted deviation traces, for heteroscedastic and homoscedastic data, makes a second step towards a unified formalism of bivariate least squares linear regression.

Exact expressions of regression line slope and intercept estimators and related variance estimators have been rewritten in a more compact form with respect to an earlier attempt [10] in the limit of oblique regression i.e. $(\sigma_{yy})_i/(\sigma_{xx})_i = c^2$, $1 \leq i \leq n$. It is noteworthy that a constant variance ratio, c^2, for all data points, does not necessarily imply equal variances, $(\sigma_{xx})_i = \sigma_{xx} = \text{const}$, $(\sigma_{yy})_i = \sigma_{yy} = \text{const}$, $1 \leq i \leq n$. While regression line slope and intercept estimators attain a coinciding expression in different attempts [23, 25, 14, 10], the results of the current paper show that the contrary holds for related variance estimators. The same holds for both reduced major-axis and bisector regression.

Approximate expressions provided in earlier attempts for normal residuals [23, 25] make (at least in computed cases) a lower limit to their exact counterparts, as shown in Tables 2-5, YANR vs. ENRR, FNRR. The same holds, to a better extent, for the asymptotic expressions determined in the current paper, as shown in Tables 2-5, AFNR vs. ENRR, FNRR. Related fractional discrepancies for low-dispersion data (RB09, Fa09) do not exceed a few percent, which grows up to about 10% in presence of large-dispersion data (Sa09).

It is well known that the regression line slope and intercept estimators are biased towards zero for Y models e.g., [12] Chap. 1, §1.1.1, [8] Chap. 3, §3.2 [15, 16], [4] Chap. 4, §4.4. Biases can be explicitly expressed in the special case of homoscedastic models with normal residuals. More specifically, the condition $1 - \rho_{20} \ll 1$ ensures bias effects are negligible, where ρ_{20} is the reliability ratio:

$$\rho_{20} = \frac{S_{20}}{S_{20} + (n-1)\sigma_{xx}} \quad ; \tag{118}$$

which implies $0 \leq \rho_{20} \leq 1$. For further details refer to specific monographies e.g., [12] Chap. 1, §1.1.1, [8] Chap. 3, §3.2.1, [4] Chap. 4, §4.4.

Similarly, it can be seen that regression line slope and intercept variance estimators are biased towards infinity for X models. In the special case of homoscedastic models with normal residuals, the condition $1 - \rho_{02} \ll 1$

ensures bias effects are negligible, where ρ_{02} is the reliability ratio:

$$\rho_{02} = \frac{S_{02}}{S_{02} + (n-1)\sigma_{yy}} \quad ; \tag{119}$$

which implies $0 \leq \rho_{02} \leq 1$ e.g., [7].

Accordingly, slopes are understimated in Y models and overstimated in X models by a factor, ρ_{20} and $1/\rho_{02}$, respectively. For C models (oblique regression), O models (orthogonal regression), R models (reduced major-axis regression), B models (bisector regression), the regression line slope estimators lie between their counterparts related to Y and X models, according to Eqs. (118) and (119), which implies bias corrections e.g., [8] Chap. 3, §3.4.2. Though there is skepticism about an indiscriminate use of oblique regression estimators, still it is accepted the method is viable provided both instrumental and intrinsic covariance matrix are known e.g., [8] Chap. 3, §3.4.2, [4] Chap. 4, §4.5.

With regard to heteroscedastic data, an inspection of Tables 2-5 shows that for lower data dispersion (RB09 sample) the values of regression line slope and intercept estimators, deduced for weighted (Tables 2-3) and unweighted (Tables 4-5) data, are systematically smaller in the former case with respect to the latter, but are still in agreement within $\mp\sigma$. For larger dispersion data (Sa09 sample) no systematic trend of the kind considered appears, but the values of regression line slope and intercept estimators are still in agreement within $\mp\sigma$ for O, R, and B models. It may be a general property of the regression models considered in the current attempt or, more realistically, intrinsic to the samples selected for the application performed in subsection 3.2.

The reliability ratios, Eqs. (118) and (119), have been calculated for all sample data and the inequalities, $\rho_{20} > 0.92$, $\rho_{02} > 0.91$, hold in any case except $\rho_{02} > 0.86$ for the Sa09 sample, which implies poorly biased regression line slope and intercept estimators for the samples considered using Y and X models and, a fortiori, using C, O, R, and B models.

Numerical simulations can determine the performance of the regression coefficients in presence of small samples and large scatter, and evaluate whether the approximations made in deriving variances are accurate. According to the results of a classical paper [14], the uncertainties to the slope predicted by O models are, on average, larger than those predicted by Y, R, or B models. For this reason, skepticism is expressed towards O models and,

in any case, caution is urged in interpreting slopes when small samples and large scatter are involved [14].

On the other hand, O models are special cases of C models, which could also include R and B models, and the predicted slopes lie between their counterparts related to the limiting cases of Y and X models. Extended numerical simulations should be used for searching a relation between the family of C models, $c^2 = c^2_{\min}$, with the lowest uncertainty to the slope, and values of population variances and covariance, namely $c^2_{\min} = f(\sigma_{XX}, \sigma_{YY}, \sigma_{XY})$. In this view, it should be recommended use of C models where $c^2 = c^2_{\min}$ for assigned sample variances and covariance, which estimate their counterparts related to the parent population.

Concerning samples listed in Tables 2-5 and represented in Figs. 2-3, the slope uncertainty predicted by O models is slightly larger than the slope predicted by R and B models for non normal residuals (ENNR), while the reverse occurs for normal residuals (ENNR). In addition, the slope uncertainty predicted by G models (the general case), when estimated, is close to the slope uncertainty predicted by O, R, and B models.

5 Conclusion

From the standpoint of a unified analytic formalism of bivariate least squares linear regression, extreme structural models have been conceived as a limiting case where the instrumental scatter is negligible (ideally null) with respect to the intrinsic scatter.

Within the framework of a variant of the classical additive error model e.g., [8] Chap. 1, §1.2, Chap. 3, §3.2.1, [4] Chap. 4, §4.3, [16], the classical results presented in earlier papers [14, 10] have been rewritten in a more compact form using a new formalism in terms of weighted deviation traces which, for homoscedastic data, reduce to usual quantities, leaving aside an unessential (but dimensional) multiplicative factor.

Regression line slope and intercept estimators, and related variance estimators, have been expressed in the special case of uncorrelated errors in X and in Y for the following models: (Y) errors in X negligible (ideally null) with respect to errors in Y; (X) errors in Y negligible (ideally null) with respect to errors in X; (C) oblique regression; (O) orthogonal regression; (R) reduced major-axis regression; (B) bisector regression. Related variance estimators have been expressed for both non normal and normal residuals and

compared to their counterparts determined for functional models [7].

Under the assumption that regression line slope and intercept variance estimators for homoscedastic and heteroscedastic data are connected to a similar extent in functional and structural models, the above mentioned results have been extended from homoscedastic to heteroscedastic data. In absence of a rigorous proof, related expressions have been considered as approximate results.

An example of astronomical application has been considered, concerning the [O/H]-[Fe/H] empirical relations deduced from five samples related to different populations and/or different methods of oxygen abundance determination. For low-dispersion samples and assigned methods, different regression models have been found to yield results which are in agreement within the errors ($\mp\sigma$) for both heteroscedastic and homoscedastic data, while the contrary has been shown to hold for large-dispersion samples. In any case, samples related to different methods have been found to produce discrepant results, due to the presence of (still undetected) systematic errors, which implies no definitive statement can be made at present.

Asymptotic expressions have been found to approximate regression line slope and intercept variance estimators, for normal residuals, to a better extent with respect to earlier attempts [23, 25]. Related fractional discrepancies have been shown to be not exceeding a few percent for low-dispersion data, which has grown up to about 10% in presence of large-dispersion data.

An extension of the formalism to generic structural models has been left to further investigation.

Acknowledgements

Thanks are due to G.J. Babu, E.D. Feigelson, M.A. Bershady, I. Lavagnini, S.J. Schmidt for fruitful e-mail correspondance on their quoted papers [10, 1, 17, 21], respectively. The author is indebted to G.J. Babu and E.D. Feigelson for having kindly provided the erratum of their quoted paper [10] before publication [11].

References

[1] M.G. Akritas, M.A. Bershady, Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter, The Astrophysical Journal 470 (1996) 706-714.

[2] P.J. Brown, Mesaurement, Regression and Calibration, Oxford Statistical Science Series 12, Oxford Science Publications, 1993.

[3] J.P. Buonaccorsi, Estimation in two-stage models with heteroscedasticity, International Statistical Review 74 (2006) 403-415.

[4] J.P. Buonaccorsi, Mesaurement Error: Models, Methods and Applications, Chapman & Hall/CRC, London, 2010.

[5] R. Caimmi, The G-dwarf problem in the Galactic spheroid, New Astronomy 12 (2007) 289-321.

[6] R. Caimmi, A Simple Multistage Closed-(Box+Reservoir) Model of Chemical Evolution, Serbian Astronomical Journal 183 (2011) 37-61.

[7] R. Caimmi, Bivariate least squares linear regression: Towards a unified analytic formalism. I. Functional models, New Astronomy 16 (2011) 337-356.

[8] R.J. Carroll, T.D. Ruppert, L.A. Stefanski,C.M. Crainiceanu, Mesaurement Error in Nonlinear Models, Monographs on Statistics and Applied Probability 105, Chapman & Hall/CRC, London, 2006.

[9] D. Fabbian, P.E. Nisseu, M. Asplund, et al., The C/O ratio at low metallicity: constraints on early chemical evolution from observations of Galactic halo stars, Astronomy & Astrophysics 500 (2009) 1143-1155.

[10] E.D. Feigelson, G.J. Babu, Linear regression in astronomy II, The Astrophysical Journal 397 (1992) 55-67.

[11] E.D. Feigelson, G.J. Babu, Linear regression in astronomy II erratum, The Astrophysical Journal 728 (2011) 72-72.

[12] W.A. Fuller, Mesaurement Error Models, J. Wiley & Sons, New York, 1987.

[13] J.S. Garden, D.G. Mitchell, D.G., Mills, Nonconstant variance regression techniques for calibration-curve-based analysis, Analytical Chemistry 52 (1980) 2310-2315.

[14] T. Isobe, E.D. Feigelson, M.G. Akritas, G.J. Babu, Linear regression in astronomy, The Astrophysical Journal 364 (1990) 104-113.

[15] B.C. Kelly, Some Aspects of Measurement Error in Linear Regression of Astronomical Data, The Astrophysical Journal 665 (2007) 1489-1506.

[16] B.C Kelly, Measurement Error Models in Astronomy, arxiv:1112.1745 (2011) 1-15.

[17] I. Lavagnini, F. Magno, A statistical overview on univariate calibration, inverse regression and detection limits. Application to gas chromatography/mass spectrometry technique, Mass Spectrometry Reviews 26 (2007) 1-18.

[18] R.P. Miller, Simultaneous Statistical Inference, McGrew-Hill, New York, 1966.

[19] C. Osborne, Statistical calibration: a review, International Statistical Review 59 (1991) 309-336.

[20] J.A.Rich, A.M. Boesgaard, Beryllium, Oxygen, and Iron Abundances in Extremely Metal-Deficient Stars. The Astrophysical Journal 701 (2009) 1519-1533.

[21] S.J. Schmidt, G. Wallerstein, V.M. Woolf, J.L. Bean, Cool Star Oxygen Abundances from Spectral Synthesis of TiO Bands, Publications of the Astronomical Society of the Pacific 121 (2009) 1083-1089.

[22] S. Tremaine, K. Ghebhardt, R. Bender, et al., The Slope of the Black Hole Mass versus Velocity Dispersion Correlation, The Astrophysical Journal 574 (2002) 740-753.

[23] D. York, Least squares fitting of a straight line, Canadian Journal of Physics, 44 (1966) 1079-1086.

[24] D. York, The best isochron, Earth and Planetary Science Letters 2 (1967) 479-482.

25] D. York, Least squares fitting of a straight line with correlated errors, Earth and Planetary Science Letters 5 (1969) 320-324.

A Euclidean and statistical squared residual sum

For homoscedastic data, the sum of squared (dimensional) Euclidean distances between observed points, $\mathsf{P}_i(X_i, Y_i)$, and adjusted points on the estimated regression line, $\hat{\mathsf{P}}_i(x_i, y_i)$, $y_i = \hat{a}x_i + \hat{b}$, is expressed as e.g., [12] Chap. 1, §1.3.3, [10, 7]:

$$(n-2)R = \sum_{i=1}^{n} \left[(Y_i - \overline{Y}) - \hat{a}(X_i - \overline{X}) \right]^2 = S_{02} + (\hat{a})^2 S_{20} - 2\hat{a}S_{11} \quad ; \quad (120)$$

where R is denoted as s_{vv} in the earlier quotation [12].

The sum of squared (dimensionless) statistical distances e.g., [12] Chap. 1, §1.3.3, between the above mentioned points, $\mathsf{P}_i(X_i, Y_i)$ and $\hat{\mathsf{P}}_i(x_i, y_i)$, reads [7]:

$$T_{\overline{R}} = W \left[S_{02} + (\hat{a})^2 S_{20} - 2\hat{a}S_{11} \right] \quad ; \quad (121)$$

which, for heteroscedastic data, takes the general expression [7]:

$$T_{\widetilde{R}} = \widetilde{W}_{02} + (\hat{a})^2 \widetilde{W}_{20} - 2\hat{a}\widetilde{W}_{11} \quad ; \quad (122)$$

accordingly, the extension of Eq. (120) to heteroscedastic data reads:

$$(n-2)R = \frac{n[\widetilde{W}_{02} + (\hat{a})^2 \widetilde{W}_{20} - 2\hat{a}\widetilde{W}_{11}]}{\widetilde{W}_{00}} \quad ; \quad (123)$$

which, in the limit of homoscedastic data, $W_i = W = \overline{W}$, $1 \leq i \leq n$, $\widetilde{W}_{00} = n\overline{W} = nW$, $\widetilde{W}_{pq} = WS_{pq}$, via Eqs. (9), (10), reduces to Eq. (120), as expected.

B Equivalence between earlier and current formulation

Let oblique regression models be taken into consideration under the following restrictive assumptions: (1) homoscedastic data; (2) uncorrelated

errors in Y and in X; (3) normal residuals. Accordingly, the regression line slope variance estimator is expressed by Eq. (56) where the function, $\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})$, may be different for different methods and/or models, as shown in Table 1. Aiming to a formal demonstration, some preliminary relations are needed.

In terms of dimensionless ratios, using Eqs. (23) and (37), Eq. (120) translates into:

$$\frac{(n-2)R}{\hat{a}S_{11}} = \frac{\hat{a}_{\rm X}}{\hat{a}} + \frac{\hat{a}}{\hat{a}_{\rm Y}} - 2 = \frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}} + \frac{\hat{a} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{124}$$

where the following identities:

$$\frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}} + \frac{\hat{a} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} = \frac{\hat{a}_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} - \frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}}\frac{\hat{a} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{125}$$

$$\frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}}\frac{\hat{a} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} = \frac{\hat{a}_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} - \frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}} - \frac{\hat{a} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} = \frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}_{\rm Y}} - \frac{\hat{a}_{\rm X} - \hat{a}}{\hat{a}}; \tag{126}$$

may easily be verified.

In the case under discussion of oblique regression models, $\hat{a} = \hat{a}_{\rm C}$, the following inequalities hold [14]:

$$\hat{a}_{\rm X} \geq \hat{a}_{\rm C} \geq \hat{a}_{\rm Y} \quad ; \qquad S_{11} > 0 \quad ; \tag{127}$$

$$\hat{a}_{\rm X} \leq \hat{a}_{\rm C} \leq \hat{a}_{\rm Y} \quad ; \qquad S_{11} < 0 \quad ; \tag{128}$$

which makes the left-hand side of Eq. (124) always positive provided $S_{11} \neq 0$.

Using the method of partial differentiation, the regression line slope variance estimator in the case under discussion is [7]:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = \frac{(\hat{a}_{\rm C})^2}{n-2}\left[2\frac{S_{02}S_{20} - (S_{11})^2}{(S_{11})^2} + 2 - \frac{S_{02} + (\hat{a}_{\rm C})^2 S_{20}}{\hat{a}_{\rm C} S_{11}}\right] \quad ; \tag{129}$$

and the substitution of Eqs. (23) and (37) into (129), using (125) and (126) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = \frac{(\hat{a}_{\rm C})^2}{n-2}\left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}} + \frac{\hat{a}_{\rm C} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + 2\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}}\frac{\hat{a}_{\rm C} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}}\right] \quad ; \tag{130}$$

from which the following is inferred by comparison with Eq. (56):

$$\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) = 2\frac{\hat{a}_{\rm X} - \hat{a}_{\rm C}}{\hat{a}_{\rm C}}\frac{\hat{a}_{\rm C} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{131}$$

as listed in Table 1.

Using the method of moments estimators, the elements sample covariance matrix are:

$$m_{XX} = \frac{S_{20}}{n-1} \quad ; \qquad m_{YY} = \frac{S_{02}}{n-1} \quad ; \qquad m_{XY} = m_{YX} = \frac{S_{11}}{n-1} \quad ; \quad (132)$$

which, in terms of the variance estimators, $(\hat{\sigma}_{xx})_{\mathrm{S}}$ (intrinsic x error distribution), $(\hat{\sigma}_{xx})_{\mathrm{F}}$ (instrumantal x error distribution), $(\hat{\sigma}_{yy})_{\mathrm{F}}$ (instrumental y error distribution) via $c_{\mathrm{F}}^2 = (\sigma_{yy})_{\mathrm{F}}/(\sigma_{xx})_{\mathrm{F}}$, and regression line slope estimator, $\hat{a}_{\mathrm{C}}$, are expressed as:

$$m_{XX} = (\hat{\sigma}_{xx})_{\mathrm{S}} + (\hat{\sigma}_{xx})_{\mathrm{F}} \quad ; \tag{133a}$$

$$m_{YY} = (\hat{a}_{\mathrm{C}})^2(\hat{\sigma}_{xx})_{\mathrm{S}} + (c_{\mathrm{F}})^2(\hat{\sigma}_{xx})_{\mathrm{F}} \quad ; \tag{133b}$$

$$m_{XY} = m_{YX} = \hat{a}_{\mathrm{C}}(\hat{\sigma}_{xx})_{\mathrm{S}} \quad ; \tag{133c}$$

for further details and specification of the model refer to the parent paper [12] Chap. 1, §1.3.2.

The substitution of Eqs. (132) and (133) into (120) yields:

$$(n-2)R_{\mathrm{C}} = (n-1)[(\hat{a}_{\mathrm{C}})^2 + (c_{\mathrm{F}})^2](\hat{\sigma}_{xx})_{\mathrm{F}} \quad ; \tag{134}$$

where the variance ratio, $(c_{\mathrm{F}})^2$, may explicitly be expressed using Eqs. (132) and (133). The result is:

$$(c_{\mathrm{F}})^2 = \frac{\hat{a}_{\mathrm{C}}(S_{02} - \hat{a}_{\mathrm{C}}S_{11})}{\hat{a}_{\mathrm{C}}S_{20} - S_{11}} \quad ; \tag{135}$$

which, using Eq. (120) and performing some algebra, takes the equivalent form:

$$(\hat{a}_{\mathrm{C}})^2 + (c_{\mathrm{F}})^2 = \frac{\hat{a}_{\mathrm{C}}(n-2)R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}S_{20} - S_{11}} \quad ; \tag{136}$$

finally, the substitution of Eq. (136) into (134) yields:

$$(\hat{\sigma}_{xx})_{\mathrm{F}} = \frac{\hat{a}_{\mathrm{C}}S_{20} - S_{11}}{(n-1)\hat{a}_{\mathrm{C}}} \quad ; \tag{137}$$

where the dependence on the variance ratio, $(c_{\mathrm{F}})^2$, has been eliminated.

In the limit of large samples ($n \gg 1$, ideally $n \to +\infty$) where, in addition, $S_{11} \neq 0$, the regression line slope variance estimator is [12] Chap. 1, §1.3.2:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{1}{n-1}\frac{1}{[(\hat{\sigma}_{xx})_{\mathrm{S}}]^2}\left\{[(\hat{\sigma}_{xx})_{\mathrm{S}} + (\hat{\sigma}_{xx})_{\mathrm{F}}]R_{\mathrm{C}} - (\hat{a}_{\mathrm{C}})^2[(\hat{\sigma}_{xx})_{\mathrm{F}}]^2\right\} \quad ; \quad (138)$$

and the substitution of Eqs. (124), (132), (133), (137), into (138) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2} \times \left[\frac{\hat{a}_{\mathrm{X}} - \hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}} + \frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{\hat{a}_{\mathrm{X}} - \hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2\right] \quad ; \quad (139)$$

from which the following is inferred by comparison with Eq. (56):

$$\Theta(\hat{a}_{\mathrm{C}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}}) = \frac{\hat{a}_{\mathrm{X}} - \hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}} - \hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2 \quad ; \quad (140)$$

as listed in Table 1.

On the other hand, the regression line slope variance estimator reported in an earlier attempt [10] Eq. (4) therein, reads:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2}\left[\frac{(n-2)R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}S_{11}} + \frac{\hat{a}_{\mathrm{C}}(S_{02} - \hat{a}_{\mathrm{C}}S_{11})}{(c_{\mathrm{S}})^2}\frac{(n-2)R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}(S_{11})^2} - \frac{n-2}{n-1}\frac{(\hat{a}_{\mathrm{C}})^2}{(S_{11})^2}\left(\frac{S_{02} - \hat{a}_{\mathrm{C}}S_{11}}{(c_{\mathrm{S}})^2}\right)^2\right] \quad ; \quad (141)$$

where $c_{\mathrm{S}} = c$ and the counterpart of Eq. (135) holds [7]:

$$c^2 = c_{\mathrm{S}}^2 = \frac{\hat{a}_{\mathrm{C}}(S_{02} - \hat{a}_{\mathrm{C}}S_{11})}{\hat{a}_{\mathrm{C}}S_{20} - S_{11}} \quad ; \quad (142)$$

and the substitution of Eqs. (124), (142), into (141), after some algebra yields Eq. (139). Then the regression line slope variance estimator, expressed by Eq. (141), coincides with its counterpart deduced by use of the method of moment estimators, expressed by Eq. (139).

Using the method of least squares estimation, under the assumption that the entire instrumental covariance matrix is known, the regression line slope variance estimator reads [12] Chap. 1, §1.3.3:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{1}{n-1}\frac{1}{(\hat{m}_{XX})^2}\left\{\hat{m}_{XX}\hat{\sigma}_{vv} + (\sigma_{xx})_{\mathrm{F}}\hat{\sigma}_{vv} - (\hat{\sigma}_{xv})^2\right\} \quad ; \tag{143}$$

$$\hat{\sigma}_{vv} = (\sigma_{yy})_{\mathrm{F}} + (\hat{a}_{\mathrm{C}})^2(\sigma_{xx})_{\mathrm{F}} - 2\hat{a}_{\mathrm{C}}(\sigma_{xy})_{\mathrm{F}} \quad ; \tag{144}$$

$$\hat{\sigma}_{xv} = (\sigma_{xy})_{\mathrm{F}} - \hat{a}_{\mathrm{C}}(\sigma_{xx})_{\mathrm{F}} \quad ; \tag{145}$$

where $\hat{m}_{XX}$ is the maximum likelihood estimator for $(\sigma_{xx})_{\mathrm{S}}$, $\hat{m}_{XX} = (\hat{\sigma}_{xx})_{\mathrm{S}}$.

In the special case under consideration, $(\sigma_{yy})_{\mathrm{F}} = (c_{\mathrm{F}})^2(\sigma_{xx})_{\mathrm{F}}$, $(\sigma_{xy})_{\mathrm{F}} = 0$, Eqs. (143), (144), (145), reduce to:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{1}{n-1}\frac{1}{(\hat{m}_{XX})^2}\left\{[\hat{m}_{XX} + (\sigma_{xx})_{\mathrm{F}}]\,\hat{\sigma}_{vv} - (\hat{a}_{\mathrm{C}})^2[(\sigma_{xx})_{\mathrm{F}}]^2\right\} \quad ; \tag{146}$$

$$\hat{\sigma}_{vv} = [(\hat{a}_{\mathrm{C}})^2 + (c_{\mathrm{F}})^2](\sigma_{xx})_{\mathrm{F}} \quad ; \tag{147}$$

$$\hat{\sigma}_{xv} = -\hat{a}_{\mathrm{C}}(\sigma_{xx})_{\mathrm{F}} \quad ; \tag{148}$$

if, in addition, least squares estimators are proportional to corresponding moments estimators, the following relations hold:

$$[(\hat{\sigma}_{xx})_{\mathrm{U}}]_{\mathrm{lsc}} = C_{X_{\mathrm{U}}}[(\hat{\sigma}_{xx})_{\mathrm{U}}]_{\mathrm{mme}} \quad ; \qquad \mathrm{U} = \mathrm{F}, \mathrm{S} \quad ; \tag{149a}$$

$$(\hat{\sigma}_{vv})_{\mathrm{lsc}} = C_v(\hat{\sigma}_{vv})_{\mathrm{mme}} \quad ; \tag{149b}$$

where $C_{X_{\mathrm{U}}}$, C_v, are constants and the indices, lsc, mme, mean least squares estimators and methods of moments estimators, respectively.

The substitution of Eq. (149) into (146) yields:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{1}{n-1}\frac{1}{(C_{X_{\mathrm{S}}})^2[(\hat{\sigma}_{xx})_{\mathrm{S}}]^2}\left\{[C_{X_{\mathrm{S}}}(\hat{\sigma}_{xx})_{\mathrm{S}} + C_{X_{\mathrm{F}}}(\hat{\sigma}_{xx})_{\mathrm{F}}]\,C_v[(\hat{a}_{\mathrm{C}})^2 + (c_{\mathrm{F}})^2]\right.$$
$$\left. \times\ (\hat{\sigma}_{xx})_{\mathrm{F}} - (\hat{a}_{\mathrm{C}})^2[C_{X_{\mathrm{F}}}(\hat{\sigma}_{xx})_{\mathrm{F}}]^2\right\}; \tag{150}$$

where the index, mme, has been omitted for simplifying the notation.

The substitution of Eq. (134) into (150) produces:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{1}{n-1}\frac{1}{[(\hat{\sigma}_{xx})_{\mathrm{S}}]^2}$$
$$\times\left\{\frac{C_v}{C_{X_{\mathrm{S}}}}\left[(\hat{\sigma}_{xx})_{\mathrm{S}} + \frac{C_{X_{\mathrm{F}}}}{C_{X_{\mathrm{S}}}}(\hat{\sigma}_{xx})_{\mathrm{F}}\right]\frac{n-2}{n-1}R_{\mathrm{C}} - (\hat{a}_{\mathrm{C}})^2\frac{(C_{X_{\mathrm{F}}})^2}{(C_{X_{\mathrm{S}}})^2}[(\hat{\sigma}_{xx})_{\mathrm{F}}]^2\right\}; \tag{151}$$

where the estimators, $(\hat{\sigma}_{xx})_{\mathrm{F}}$ and $(\hat{\sigma}_{xx})_{\mathrm{S}}$, are expressed by Eqs. (132), (133), (135), (137). Accordingly, the explicit expression of Eq. (151) after some algebra reads:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{(S_{11})^2}\left\{\frac{C_v}{C_{X_{\mathrm{S}}}}\left[\frac{S_{11}}{\hat{a}_{\mathrm{C}}} + \frac{C_{X_{\mathrm{F}}}}{C_{X_{\mathrm{S}}}}\frac{S_{02}-\hat{a}_{\mathrm{C}}S_{11}}{(c_{\mathrm{F}})^2}\right]\frac{n-2}{n-1}R_{\mathrm{C}} - \frac{(C_{X_{\mathrm{F}}})^2}{(C_{X_{\mathrm{S}}})^2}\frac{(\hat{a}_{\mathrm{C}})^2}{n-1}\left[\frac{S_{02}-\hat{a}_{\mathrm{C}}S_{11}}{(c_{\mathrm{F}})^2}\right]^2\right\}; \quad (152)$$

where the restrictive assumptions:

$$\frac{C_{X_{\mathrm{F}}}}{C_{X_{\mathrm{S}}}} = 1 \;\; ; \qquad \frac{C_v}{C_{X_{\mathrm{S}}}} = \frac{n-1}{n-2} \;\; ; \quad (153)$$

make Eq. (152) reduce to:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{(S_{11})^2}\left\{\left[\frac{S_{11}}{\hat{a}_{\mathrm{C}}} + \frac{S_{02}-\hat{a}_{\mathrm{C}}S_{11}}{(c_{\mathrm{F}})^2}\right]R_{\mathrm{C}} - \frac{(\hat{a}_{\mathrm{C}})^2}{n-1}\left[\frac{S_{02}-\hat{a}_{\mathrm{C}}S_{11}}{(c_{\mathrm{F}})^2}\right]^2\right\}; \quad (154)$$

which formally coincides with the result of an earlier attempt where $c_{\mathrm{S}} = c$ appears instead of c_{F} [10] Eq. (4) therein.

Finally, the substitution of Eqs. (124) and (135) into (154) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2} \times \left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}} + \frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2\right] \;\; ; \quad (155)$$

from which the following is inferred by comparison with Eq. (56):

$$\Theta(\hat{a}_{\mathrm{C}}, \hat{a}_{\mathrm{Y}}, \hat{a}_{\mathrm{X}}) = \frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}} + \frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2 \;\; ; \quad (156)$$

as listed in Table 1.

The revised version of the regression line variance estimator reported in an earlier attempt [10] Eq. (4) therein, [11] reads:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2}\left\{\frac{(n-2)R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}S_{11}} + \frac{1}{n-1}\frac{(\hat{a}_{\mathrm{C}})^2}{(\hat{a}_{\mathrm{C}})^2+c^2}\left[1-\frac{n-2}{n-1}\frac{(\hat{a}_{\mathrm{C}})^2}{(\hat{a}_{\mathrm{C}})^2+c^2}\right]\left[\frac{(n-2)R_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}S_{11}}\right]^2\right\} \;\; ; \quad (157)$$

and the substitution of Eqs. (124), (142), into (157), after a lot of algebra yields:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2}\left\{\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}+\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right.$$
$$\left.+\frac{1}{n-1}\left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}+\frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2\right]\right\} \quad ; \qquad (158)$$

from which the following is inferred by comparison with Eq. (56):

$$\Theta(\hat{a}_{\mathrm{C}},\hat{a}_{\mathrm{Y}},\hat{a}_{\mathrm{X}}) = \frac{1}{n-1}\left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}+\frac{1}{n-1}\left(\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right)^2\right] \quad ; \quad (159)$$

as listed in Table 1.

The asymptotic expression ($n \to +\infty$) of Eq. (158) is obtained neglecting the terms of higher order with respect to $1/n$. The result is:

$$(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})^2 = \frac{(\hat{a}_{\mathrm{C}})^2}{n-2}\left\{\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{C}}}{\hat{a}_{\mathrm{C}}}+\frac{\hat{a}_{\mathrm{C}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right\} \quad ; \qquad (160)$$

which implies $\Theta(\hat{a}_{\mathrm{C}},\hat{a}_{\mathrm{Y}},\hat{a}_{\mathrm{X}}) = 0$, as listed in Table 1. The asymptotic formula, Eq. (160), coincides with an approximation reported in earlier attempts [23, 25] for Y models and makes a better approximation for X, C, O, R, and B models.

C Data-independent residuals

Let u_{A}, u_{B}, be independent random variables, $f_{\mathrm{A}}(u_{\mathrm{A}})\,\mathrm{d}u_{\mathrm{A}}$, $f_{\mathrm{B}}(u_{\mathrm{B}})\,\mathrm{d}u_{\mathrm{B}}$, related distributions, u^*_{A}, u^*_{B}, related expectation values, and $\hat{u}_{\mathrm{A}}$, $\hat{u}_{\mathrm{B}}$, related estimators. The random variable, $u = u_{\mathrm{A}}u_{\mathrm{B}}$, obeys the distribution, $f(u)\,\mathrm{d}u = \int_{\mathrm{U}} f_{\mathrm{A}}(u_{\mathrm{A}})f_{\mathrm{B}}(u_{\mathrm{B}})\,\mathrm{d}u_{\mathrm{A}}\,\mathrm{d}u_{\mathrm{B}}$, where U is the domain for which the product, $u_{\mathrm{A}}u_{\mathrm{B}}$, equals a fixed u. According to a theorem of statistics, the expectation value is $u^* = (u_{\mathrm{A}}u_{\mathrm{B}})^* = u^*_{\mathrm{A}}u^*_{\mathrm{B}}$ and the related estimator is $\hat{u} = \widehat{u_{\mathrm{A}}u_{\mathrm{B}}} \approx \hat{u}_{\mathrm{A}}\hat{u}_{\mathrm{B}}$.

The special case of the arithmetic mean reads $\overline{u} = \overline{u_{\mathrm{A}}u_{\mathrm{B}}} \approx \overline{u_{\mathrm{A}}}\ \overline{u_{\mathrm{B}}}$ or:

$$\frac{1}{n}\sum_{i=1}^{n}(u_{\mathrm{A}})_i(u_{\mathrm{B}})_i \approx \frac{1}{n}\sum_{i=1}^{n}(u_{\mathrm{A}})_i\ \frac{1}{n}\sum_{i=1}^{n}(u_{\mathrm{B}})_i \quad ; \qquad (161)$$

with regard to $u_{\rm A}$ and $u_{\rm B}$ samples with population equal to n.

With these general results in mind, let Eqs. (27), (41), (60), be rewritten into the explicit form [14] Eqs. (A4)-(A6) therein[2]:

$$(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 = \frac{1}{(S_{20})^2}\sum_{i=1}^{n}\Big\{(X_i-\overline{X})^2[(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})]^2\Big\} \quad ; \tag{162}$$

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{1}{(S_{11})^2}\sum_{i=1}^{n}\Big\{(Y_i-\overline{Y})^2[(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})]^2\Big\} \quad ; \tag{163}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{1}{S_{20}S_{11}}\sum_{i=1}^{n}\Big\{(X_i-\overline{X})(Y_i-\overline{Y})[(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})] \times [(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})]\Big\} \quad ; \tag{164}$$

where (dimensional) residuals related to Y and X models are enclosed in square brackets via Eqs. (24) and (38), respectively.

If residuals are independent of coordinates of observed points, $\mathsf{P}_i \equiv (X_i, Y_i)$, $1 \le i \le n$, then the particularization of Eq. (161) to $u_{\rm A} = (X_i-\overline{X})^2$, $(Y_i-\overline{Y})^2$, $(X_i-\overline{X})(Y_i-\overline{Y})$; $u_{\rm B} = [(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})]^2$, $[(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})]^2$, $[(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})][(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})]$; respectively, makes Eqs. (162)-(164) reduce to:

$$(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 = \frac{1}{n}\frac{1}{(S_{20})^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}[(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})]^2 \quad ; \tag{165}$$

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{1}{n}\frac{1}{(S_{11})^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\sum_{i=1}^{n}[(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})]^2 \quad ; \tag{166}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{1}{n}\frac{1}{S_{20}S_{11}}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\sum_{i=1}^{n}[(Y_i-\overline{Y})-\hat{a}_{\rm Y}(X_i-\overline{X})] \times [(Y_i-\overline{Y})-\hat{a}_{\rm X}(X_i-\overline{X})] \quad ; \tag{167}$$

as outlined in an earlier attempt [14].

[2]With regard to the above quoted Eqs. (A4)-(A6), it is worth noticing $a_{\rm Y}$, $a_{\rm X}$, are denoted as β_1, β_2, respectively, and β_1 has to be replaced by $(\beta_1)^{-1}$ in Eq. (A6) to get the right dimensions and to be consistent with the expression of the covariance term [14] note to Table 1 therein.

Using Eqs. (15), (23), (37), while performing some algebra, Eqs. (165)-(167) may be cast into the form:

$$(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 = \frac{(\hat{a}_{\rm Y})^2}{n}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{168}$$

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{(\hat{a}_{\rm X})^2}{n}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{169}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{(\hat{a}_{\rm Y})^2}{n}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{170}$$

which provide correct asymptotic ($n \to +\infty$) formulae but understimate the true regression coefficient uncertainty in samples with low ($n \lesssim 50$) or weakly correlated population [10].

An inspection of Table 1 shows Eq. (25) and the asymptotic ($n \to +\infty$) expression of Eq. (39) match Eqs. (168) and (169), respectively, provided n therein is replaced by $(n-2)$. Accordingly, Eqs. (168)-(170) translate into:

$$(\hat{\sigma}_{\hat{a}_{\rm Y}})^2 = \frac{(\hat{a}_{\rm Y})^2}{n-2}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{171}$$

$$(\hat{\sigma}_{\hat{a}_{\rm X}})^2 = \frac{(\hat{a}_{\rm X})^2}{n-2}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{172}$$

$$\hat{\sigma}_{\hat{a}_{\rm Y}\hat{a}_{\rm X}} = \frac{(\hat{a}_{\rm Y})^2}{n-2}\frac{\hat{a}_{\rm X}-\hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} \quad ; \tag{173}$$

which are expected to yield improved values for samples with low or weakly correlated population.

With regard to oblique regression models, the substitution of Eqs. (171)-(173) into (63) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = \frac{(\hat{a}_{\rm C})^2}{n-2} \times \left\{ A_{\rm XY} + \frac{2(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2(A_{\rm XY})^2 A_{\rm XC}A_{\rm CY}}{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2 A_{\rm XC}A_{\rm CY} + [\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY} - (\hat{a}_{\rm C})^2 A_{\rm XC}]^2} \right\} ; \tag{174}$$

where the identity:

$$A_{\rm XY} = A_{\rm XU} + A_{\rm UY} + A_{\rm XU}A_{\rm UY} \quad ; \tag{175}$$

may easily be verified, being U = C in the case under discussion. Accordingly, Eq. (174) may be cast under the form:

$$(\hat{\sigma}_{\hat{a}_{\rm C}})^2 = \frac{(\hat{a}_{\rm C})^2}{n-2}\left[A_{\rm XC} + A_{\rm CY} + \Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})\right] \quad ; \tag{176}$$

$$\Theta(\hat{a}_{\rm C}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) = A_{\rm XC}A_{\rm CY} \times \left\{1 + \frac{2(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2(A_{\rm XY})^2}{4(\hat{a}_{\rm Y})^2(\hat{a}_{\rm C})^2 A_{\rm XC}A_{\rm CY} + [\hat{a}_{\rm Y}\hat{a}_{\rm X}A_{\rm CY} - (\hat{a}_{\rm C})^2 A_{\rm XC}]^2}\right\} \tag{177}$$

where Eqs. (176) and (56) coincide in the limit, $\Theta \to 0$.

With regard to reduced major axis regression models, the substitution of Eqs. (171)-(173) into (85) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = \frac{(\hat{a}_{\rm R})^2}{n-2}\frac{A_{\rm XY}}{2}\left(1 + \frac{\hat{a}_{\rm Y}}{\hat{a}_{\rm X}}\right) \quad ; \tag{178}$$

where the identity:

$$\frac{\hat{a}_{\rm Y}}{\hat{a}_{\rm X}} = \frac{1}{A_{\rm XY}+1} \quad ; \tag{179}$$

may easily be verified. Accordingly, Eq. (178) via (175) may be cast into the form:

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = \frac{(\hat{a}_{\rm R})^2}{n-2}\left[A_{\rm XR} + A_{\rm RY} + \Theta(\hat{a}_{\rm R}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})\right] \quad ; \tag{180}$$

$$\Theta(\hat{a}_{\rm R}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) = A_{\rm XR}A_{\rm RY} - \frac{1}{2}\frac{(A_{\rm XY})^2}{A_{\rm XY}+1} \quad ; \tag{181}$$

where Eqs. (180) and (83) coincide in the limit, $\Theta \to 0$.

With regard to bisector regression models, the substitution of Eqs. (171)-(173) into (103) yields after some algebra:

$$(\hat{\sigma}_{\hat{a}_{\rm B}})^2 = \frac{(\hat{a}_{\rm B})^2}{n-2}\frac{A_{\rm XY}(\hat{a}_{\rm Y})^2}{(\hat{a}_{\rm Y}+\hat{a}_{\rm X})^2}\left[\frac{a_u^2 + (\hat{a}_{\rm X})^2}{a_u^2 + (\hat{a}_{\rm Y})^2} + \frac{a_u^2 + (\hat{a}_{\rm Y})^2}{a_u^2 + (\hat{a}_{\rm X})^2}\frac{(\hat{a}_{\rm X})^2}{(\hat{a}_{\rm Y})^2} + 2\right] \quad ; \tag{182}$$

which, using Eqs. (175) and (179), may be cast under the form:

$$(\hat{\sigma}_{\hat{a}_{\rm B}})^2 = \frac{(\hat{a}_{\rm B})^2}{n-2}\left[A_{\rm XB} + A_{\rm BY} + \Theta(\hat{a}_{\rm B}, \hat{a}_{\rm Y}, \hat{a}_{\rm X})\right] \quad ; \tag{183}$$

$$\Theta(\hat{a}_{\rm B}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) = \frac{A_{\rm XY}}{(A_{\rm XY}+2)^2}\left[\frac{a_u^2 + (\hat{a}_{\rm X})^2}{a_u^2 + (\hat{a}_{\rm Y})^2} + \frac{a_u^2 + (\hat{a}_{\rm Y})^2}{a_u^2 + (\hat{a}_{\rm X})^2}\frac{(\hat{a}_{\rm X})^2}{(\hat{a}_{\rm Y})^2} + 2\right] - A_{\rm XY} + A_{\rm XB}A_{\rm BY} \quad ; \tag{184}$$

where Eqs. (183) and (101) coincide in the limit, $\Theta \to 0$.

D Special cases of oblique regression

With regard to homoscedastic data, special cases of oblique regression may be considered starting from the expression of regression line slope and intercept estimators, Eqs. (54) and (55), and related variance estimators, Eqs. (56) and (57) for normal residuals or (58) and (59) for non normal residuals. As outlined in the parent paper [10], the special cases, $c \to +\infty$, $c \to 0$, $c \to 1$, correspond to errors in X negligible with respect to errors in Y, errors in Y negligible with respect to errors in X, and orthogonal regression, respectively. In addition, the limiting case, $c \to c_{\rm rma} = \sqrt{a_{\rm X} a_{\rm Y}}$, corresponds to reduced major-axis regression e.g., [14, 7]. An exhaustive discussion related to regression line slope and intercept estimators, can be found in an earlier attempt [7]. Finally, the limiting case, $c \to c_{\rm bis}$, where $c_{\rm bis}$ is expressed by Eq. (100), corresponds to bisector regression. The result is:

$$\lim_{c\to+\infty} \hat{a}_{\rm C} = \hat{a}_{\rm Y} \ ; \quad \lim_{c\to 0} \hat{a}_{\rm C} = \hat{a}_{\rm X} \ ; \quad \lim_{c\to 1} \hat{a}_{\rm C} = \hat{a}_{\rm O} \ ; \quad \lim_{c\to c_{\rm rma}} \hat{a}_{\rm C} = \hat{a}_{\rm R} \ ; \tag{185a}$$

$$\lim_{c\to c_{\rm bis}} \hat{a}_{\rm C} = \hat{a}_{\rm B} \ ; \tag{185b}$$

where related models are denoted by the indices, Y, X, O, R, B, respectively.

Concerning regression line slope variance estimators for normal residuals, the following relations can be inferred from Eq. (56):

$$\lim_{c\to+\infty} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm Y})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm Y}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \right] \ ; \tag{186}$$

$$\lim_{c\to 0} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm X})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm X}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \right] \ ; \tag{187}$$

$$\lim_{c\to 1} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm O})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm O}}{\hat{a}_{\rm O}} + \frac{\hat{a}_{\rm O} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm O}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \right] \ ; \tag{188}$$

$$\lim_{c\to c_{\rm rma}} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm R})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm R}}{\hat{a}_{\rm R}} + \frac{\hat{a}_{\rm R} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm R}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \right] \ ; \tag{189}$$

$$\lim_{c\to c_{\rm bis}} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = \frac{(\hat{a}_{\rm B})^2}{n-2} \left[\frac{\hat{a}_{\rm X} - \hat{a}_{\rm B}}{\hat{a}_{\rm B}} + \frac{\hat{a}_{\rm B} - \hat{a}_{\rm Y}}{\hat{a}_{\rm Y}} + \Theta(\hat{a}_{\rm B}, \hat{a}_{\rm Y}, \hat{a}_{\rm X}) \right] \ ; \tag{190}$$

where the function, Θ, is listed in Table 1 for different methods and/or models.

A comparison between Eqs. (25), (39), and (186), (187), respectively, yields:

$$\lim_{c\to+\infty} [(\hat{\sigma}_{\hat{a}_{\rm C}})_{\rm N}]^2 = [(\hat{\sigma}_{\hat{a}_{\rm Y}})_{\rm N}]^2 \ ; \tag{191}$$

$$\lim_{c\to 0}[(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{a}_{\mathrm{X}}})_{\mathrm{N}}]^2 \quad ; \tag{192}$$

and, on the other hand:

$$\lim_{c\to 1}[(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{a}_{\mathrm{O}}})_{\mathrm{N}}]^2 \quad ; \tag{193}$$

$$\lim_{c\to c_{\mathrm{rma}}}[(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{a}_{\mathrm{R}}})_{\mathrm{N}}]^2 \quad ; \tag{194}$$

$$\lim_{c\to c_{\mathrm{bis}}}[(\hat{\sigma}_{\hat{a}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{a}_{\mathrm{B}}})_{\mathrm{N}}]^2 \quad ; \tag{195}$$

by definition of orthogonal regression e.g., [8] Chap. 3, §4.4.2, reduced major-axis regression e.g., [14, 7], and bisector regression e.g., [14].

Concerning regression line intercept variance estimators for normal residuals, the following relations can be inferred from Eq. (57):

$$\lim_{c\to +\infty}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = \frac{\hat{a}_{\mathrm{Y}}}{n-2}\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\frac{S_{11}}{S_{00}} + (\overline{X})^2[(\hat{\sigma}_{\hat{a}_{\mathrm{Y}}})_{\mathrm{N}}]^2 \quad ; \tag{196}$$

$$\lim_{c\to 0}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = \frac{\hat{a}_{\mathrm{X}}}{n-2}\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\frac{S_{11}}{S_{00}} + (\overline{X})^2[(\hat{\sigma}_{\hat{a}_{\mathrm{X}}})_{\mathrm{N}}]^2 \quad ; \tag{197}$$

$$\lim_{c\to 1}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = \frac{\hat{a}_{\mathrm{O}}}{n-2}\left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{O}}}{\hat{a}_{\mathrm{O}}} + \frac{\hat{a}_{\mathrm{O}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2[(\hat{\sigma}_{\hat{a}_{\mathrm{O}}})_{\mathrm{N}}]^2 \quad ; \tag{198}$$

$$\lim_{c\to c_{\mathrm{rma}}}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = \frac{\hat{a}_{\mathrm{R}}}{n-2}\left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{R}}}{\hat{a}_{\mathrm{R}}} + \frac{\hat{a}_{\mathrm{R}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2[(\hat{\sigma}_{\hat{a}_{\mathrm{R}}})_{\mathrm{N}}]^2 \quad ; \tag{199}$$

$$\lim_{c\to c_{\mathrm{bis}}}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = \frac{\hat{a}_{\mathrm{B}}}{n-2}\left[\frac{\hat{a}_{\mathrm{X}}-\hat{a}_{\mathrm{B}}}{\hat{a}_{\mathrm{B}}} + \frac{\hat{a}_{\mathrm{B}}-\hat{a}_{\mathrm{Y}}}{\hat{a}_{\mathrm{Y}}}\right]\frac{S_{11}}{S_{00}} + (\overline{X})^2[(\hat{\sigma}_{\hat{a}_{\mathrm{B}}})_{\mathrm{N}}]^2 \quad ; \tag{200}$$

due to Eqs. (185) and (191)-(195).

A comparison between Eqs. (26), (40), and (196), (197), respectively, yields:

$$\lim_{c\to +\infty}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{b}_{\mathrm{Y}}})_{\mathrm{N}}]^2 \quad ; \tag{201}$$

$$\lim_{c\to 0}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{b}_{\mathrm{X}}})_{\mathrm{N}}]^2 \quad ; \tag{202}$$

and, on the other hand:

$$\lim_{c\to 1}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{b}_{\mathrm{O}}})_{\mathrm{N}}]^2 \quad ; \tag{203}$$

$$\lim_{c\to c_{\mathrm{rma}}}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{b}_{\mathrm{R}}})_{\mathrm{N}}]^2 \quad ; \tag{204}$$

$$\lim_{c\to c_{\mathrm{bis}}}[(\hat{\sigma}_{\hat{b}_{\mathrm{C}}})_{\mathrm{N}}]^2 = [(\hat{\sigma}_{\hat{b}_{\mathrm{B}}})_{\mathrm{N}}]^2 \quad ; \tag{205}$$

by definition of orthogonal regression e.g., [8] Chap. 3, §4.4.2, reduced major-axis regression e.g., [14, 7], and bisector regression e.g., [14].

Concerning regression line slope variance estimators for non normal residuals, the following relations can be inferred from Eq. (58):

$$\lim_{c\to+\infty}(\hat{\sigma}_{\hat{a}_\mathrm{C}})^2 = (\hat{\sigma}_{\hat{a}_\mathrm{Y}})^2 \quad ; \tag{206}$$

$$\lim_{c\to 0}(\hat{\sigma}_{\hat{a}_\mathrm{C}})^2 = (\hat{\sigma}_{\hat{a}_\mathrm{X}})^2 \quad ; \tag{207}$$

$$\lim_{c\to 1}(\hat{\sigma}_{\hat{a}_\mathrm{C}})^2 = (\hat{a}_\mathrm{O})^2\frac{a_u^4(\hat{\sigma}_{\hat{a}_\mathrm{Y}})^2 + (\hat{a}_\mathrm{Y})^4(\hat{\sigma}_{\hat{a}_\mathrm{X}})^2 + 2(\hat{a}_\mathrm{Y})^2a_u^2\hat{\sigma}_{\hat{a}_\mathrm{Y}\hat{a}_\mathrm{X}}}{(\hat{a}_\mathrm{Y})^2[4(\hat{a}_\mathrm{Y})^2a_u^2 + (\hat{a}_\mathrm{Y}\hat{a}_\mathrm{X} - a_u^2)^2]} \quad ; \tag{208}$$

where $a_u = 1$ is the (dimensional) unit slope, according to their counterparts expressed in the parent paper [14] provided $|\hat{a}_\mathrm{Y}|$ is replaced by $\hat{a}_\mathrm{Y}$ therein.

On the other hand, the following relation holds:

$$\lim_{c\to 1}(\hat{\sigma}_{\hat{a}_\mathrm{C}})^2 = (\hat{\sigma}_{\hat{a}_\mathrm{O}})^2 \quad ; \tag{209}$$

by definition of orthogonal regression e.g., [8] Chap. 3, §4.4.2.

The intercept variance estimators for special cases of oblique regression, are expressed by a single formula characterized by different dimensionless coefficients, γ_{1k}, γ_{2k}, where $k = 1, 2, 4$, for Y, X, O, models, respectively [14]. The extended expressions for oblique regression, where $k = 6$ for C models, read:

$$\gamma_{16} = \frac{\hat{a}_\mathrm{C}c^2}{\hat{a}_\mathrm{Y}[4(\hat{a}_\mathrm{Y})^2c^2 + (\hat{a}_\mathrm{Y}\hat{a}_\mathrm{X} - c^2)^2]^{1/2}} \quad ; \tag{210}$$

$$\gamma_{26} = \frac{\hat{a}_\mathrm{C}\hat{a}_\mathrm{Y}}{[4(\hat{a}_\mathrm{Y})^2c^2 + (\hat{a}_\mathrm{Y}\hat{a}_\mathrm{X} - c^2)^2]^{1/2}} \quad ; \tag{211}$$

which, for the above mentioned special cases, reduce to:

$$\gamma_{11} = \lim_{c\to+\infty}\gamma_{16} = 1 \quad ; \tag{212}$$

$$\gamma_{21} = \lim_{c\to+\infty}\gamma_{26} = 0 \quad ; \tag{213}$$

$$\gamma_{12} = \lim_{c\to 0}\gamma_{16} = 0 \quad ; \tag{214}$$

$$\gamma_{22} = \lim_{c\to 0}\gamma_{26} = 1 \quad ; \tag{215}$$

according to their counterparts expressed in the parent paper [14] and, in addition:

$$\gamma_{14} = \lim_{c\to 1}\gamma_{16} = \frac{\hat{a}_\mathrm{O}a_u^2}{\hat{a}_\mathrm{Y}[4(\hat{a}_\mathrm{Y})^2a_u^2 + (\hat{a}_\mathrm{Y}\hat{a}_\mathrm{X} - a_u^2)^2]^{1/2}} \quad ; \tag{216}$$

$$\gamma_{24} = \lim_{c\to 1}\gamma_{26} = \frac{\hat{a}_{\rm O}\hat{a}_{\rm Y}}{[4(\hat{a}_{\rm Y})^2 a_u^2 + (\hat{a}_{\rm Y}\hat{a}_{\rm X} - a_u^2)^2]^{1/2}} \quad ; \tag{217}$$

where $a_u = 1$ is the (dimensional) unit slope, according to their counterparts expressed in the parent paper [14] provided $|\hat{a}_{\rm Y}|$ is replaced by $\hat{a}_{\rm Y}$ therein.

The validity of Eqs. (212)-(217) implies the validity of the following relations:

$$\lim_{c\to+\infty}(\hat{\sigma}_{\hat{b}_{\rm C}})^2 = (\hat{\sigma}_{\hat{b}_{\rm Y}})^2 \quad ; \tag{218}$$

$$\lim_{c\to 0}(\hat{\sigma}_{\hat{b}_{\rm C}})^2 = (\hat{\sigma}_{\hat{b}_{\rm X}})^2 \quad ; \tag{219}$$

$$\lim_{c\to 1}(\hat{\sigma}_{\hat{b}_{\rm C}})^2 = (\hat{\sigma}_{\hat{b}_{\rm O}})^2 \quad ; \tag{220}$$

in the general case of non normal residuals.

The above results cannot be extended to R and B models i.e. $k = 5, 3$, respectively, due to use of the δ-method for determining variance estimators [14], which implies $\lim_{u_{\rm C}\to u_{\rm U}}(\hat{\sigma}_{\hat{u}_{\rm C}})^2 \neq (\hat{\sigma}_{\hat{u}_{\rm U}})^2$; $u = a, b$; U=R, B.

With regard to heteroscedastic data, the above results can be extended starting from the expression of regression line slope and intercept estimators, Eqs. (69) and (70) for normal residuals, which yields counterparts of Eqs. (196)-(199) where $n(\widetilde{w_x})_{pq}/(\widetilde{w_x})_{00}$ appears in place of S_{pq} and $\hat{a}'_{\rm Y} = (\widetilde{w_x})_{11}/(\widetilde{w_x})_{20}$ in place of $\hat{a}_{\rm Y} = (\widetilde{w_y})_{11}/(\widetilde{w_y})_{20}$. A similar procedure can be used for non normal residuals, starting from Eqs. (71) and (72).

E C11 erratum

Due to the occurrence of printing errors, Eqs. (147) and (152) in an earlier attempt [7] were lacking of a dimensionless factor and must be corrected as follows:

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = \frac{2}{n-2}\frac{(\widetilde{w_x})_{02}}{(\widetilde{w_x})_{20}}\left\{\frac{1}{(\lambda_{w_x})^2} - {\rm sgn}[(\widetilde{w_x})_{11}]\frac{1}{\lambda_{w_x}}\right\} \quad ; \tag{147}$$

$$(\hat{\sigma}_{\hat{a}_{\rm R}})^2 = \frac{2}{n-2}\frac{S_{02}}{S_{20}}\left[\frac{1}{(\lambda_S)^2} - {\rm sgn}(S_{11})\frac{1}{\lambda_S}\right] \quad ; \tag{152}$$

which are equivalent to their alternative expressions, Eqs. (149) and (154) therein, respectively.

Sample FB09 listed in Table 2 therein has to be read as RB09.

F A linear relation between primary elements from stellar nucleosynthesis

The composition of the interstellar medium, from which a star generation was born, remains locked in stellar atmospheres. Attention shall be restricted to primary elements i.e. those synthesized in stellar cores starting from hydrogen and helium regardless of the initial composition. Conversely, secondary elements can be synthesized only in presence of heavier (with respect to hydrogen and helium) nuclides, which are called metals in astrophysics.

The ideal situation, where a linear relation holds between primary elements in stellar atmospheres, is defined by the following assumptions.

(i) The initial stellar mass function is universal, which implies the star distribution in mass (including binary and multiple systems), normalized to unity, maintains unchanged regardless of the formation place and the formation epoch.

(ii) Gas returned after star death is instantaneously and uniformly mixed with the interstellar medium.

(iii) The yield of primary elements synthesized within a star depends only on the mass regardless of the initial composition.

(iv) Supernovae may occur as either type II ($m > 8\mathrm{m}_\odot$) or type Ia ($m \leq 8\mathrm{m}_\odot$).

Accordingly, the composition of the interstellar medium is due to the accretion of newly synthesized material via supernovae. With regard to a sufficiently short time step, Δt, let n_{I} and n_{II} be the number of type Ia and type II supernovae, respectively; in addition, let $\delta m_{\mathrm{W\,I}}$ and $\delta m_{\mathrm{W\,II}}$ be the mean mass in the primary element, W, newly synthesized and returned to the interstellar medium via type Ia and type II supernovae, respectively.

Within a time range equal to the first ℓ steps, $t_\ell - t_0 = \ell \Delta t$, the interstellar medium has been enriched by a mass in the primary element, W, as:

$$\begin{aligned} m_{\mathrm{W}} &= \sum_{k=1}^{\ell} [(n_{\mathrm{II}})_k \delta m_{\mathrm{W\,II}} + (n_{\mathrm{I}})_k \delta m_{\mathrm{W\,I}}] = \\ &= \sum_{k=1}^{\ell} (n_{\mathrm{II}})_k \delta m_{\mathrm{W\,II}} \left[1 + \frac{(n_{\mathrm{I}})_k}{(n_{\mathrm{II}})_k} \frac{\delta m_{\mathrm{W\,I}}}{\delta m_{\mathrm{W\,II}}} \right] \quad ; \end{aligned} \tag{221}$$

where $\delta m_{\mathrm{W\,I}}$ and $\delta m_{\mathrm{W\,II}}$ may be considered, to a good extent, as time independent.

The further assumption of time independent number ratios:

$$\frac{(n_{\mathrm{I}})_k}{(n_{\mathrm{II}})_k} = \frac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \quad ; \tag{222}$$

makes Eq. (221) reduce to:

$$m_{\mathrm{W}} = \sum_{k=1}^{\ell} (n_{\mathrm{II}})_k \delta m_{\mathrm{W\,II}} \left[1 + \frac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \frac{\delta m_{\mathrm{W\,I}}}{\delta m_{\mathrm{W\,II}}} \right] \quad ; \tag{223}$$

which implies an abundance ratio of two generic primary elements, A and B, in stellar atmospheres i.e. interstellar medium, as:

$$\begin{aligned} \frac{\exp_{10}[\mathrm{A/H}]}{\exp_{10}[\mathrm{B/H}]} \approx \frac{\phi_{\mathrm{A}}}{\phi_{\mathrm{B}}} &= \frac{Z_{\mathrm{A}}/(Z_{\mathrm{A}})_{\odot}}{Z_{\mathrm{B}}/(Z_{\mathrm{B}})_{\odot}} = \frac{m_{\mathrm{A}}}{m_{\mathrm{B}}} \frac{(Z_{\mathrm{B}})_{\odot}}{(Z_{\mathrm{A}})_{\odot}} \\ &= \frac{\delta m_{\mathrm{A\,II}}}{\delta m_{\mathrm{B\,II}}} \frac{1 + \dfrac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \dfrac{\delta m_{\mathrm{A\,I}}}{\delta m_{\mathrm{A\,II}}}}{1 + \dfrac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \dfrac{\delta m_{\mathrm{B\,I}}}{\delta m_{\mathrm{B\,II}}}} \frac{(Z_{\mathrm{B}})_{\odot}}{(Z_{\mathrm{A}})_{\odot}} \quad ; \end{aligned} \tag{224}$$

where [A/H], [B/H], are logarithmic number abundances normalized to the solar value; ϕ_{A}, ϕ_{B}, are mass abundances normalized to the solar value; Z_{A}, Z_{B}, are mass abundances; values are related to the interstellar medium from which the star considered was born, with the exception of $(Z_{\mathrm{A}})_{\odot}$, $(Z_{\mathrm{B}})_{\odot}$, denoting solar abundances. For further details refer to the parent paper [5].

In terms of logarithmic number abundances, Eq. (224) may be cast under the form:

$$[\mathrm{A/H}] = [\mathrm{B/H}] + b \quad ; \tag{225}$$

$$b = \log \left[\frac{\delta m_{\mathrm{A\,II}}}{\delta m_{\mathrm{B\,II}}} \frac{1 + \dfrac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \dfrac{\delta m_{\mathrm{A\,I}}}{\delta m_{\mathrm{A\,II}}}}{1 + \dfrac{n_{\mathrm{I}}}{n_{\mathrm{II}}} \dfrac{\delta m_{\mathrm{B\,I}}}{\delta m_{\mathrm{B\,II}}}} \frac{(Z_{\mathrm{B}})_{\odot}}{(Z_{\mathrm{A}})_{\odot}} \right] \quad ; \tag{226}$$

which is a linear relation with unit slope. The general case:

$$[\mathrm{A/H}] = a[\mathrm{B/H}] + b \quad ; \tag{227}$$

could arise under different assumptions.

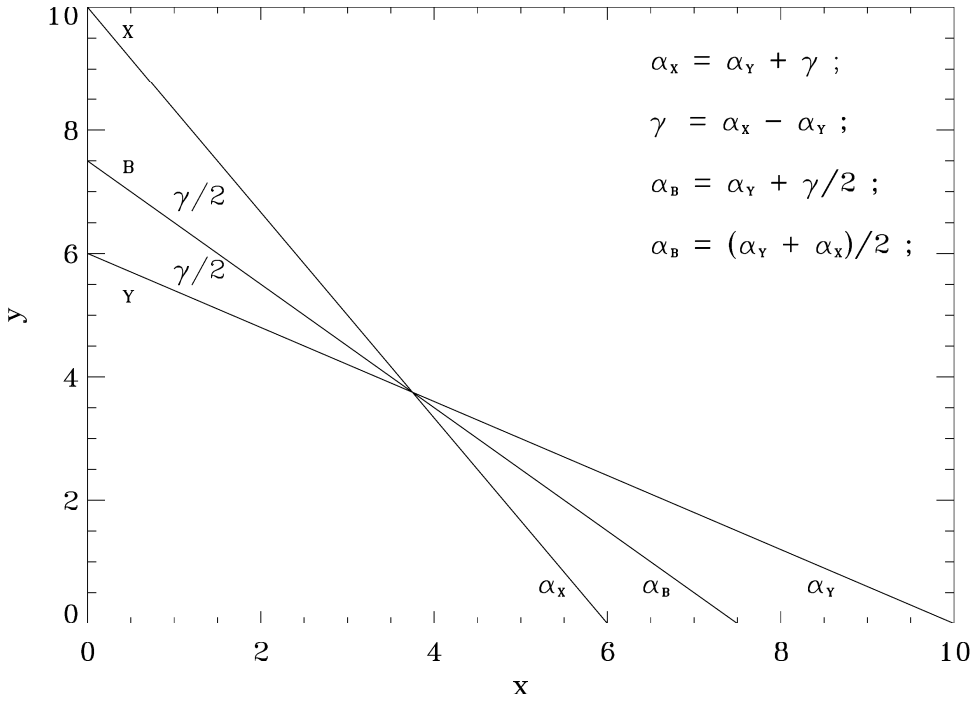


Figure 1: Regression lines related to Y, X, and B models, for an assigned sample. By definition, the B line bisects the angle, γ, formed between Y and X lines. The angle, $\alpha_{\rm B}$, formed between B line and x axis, is the arithmetic mean of the angles, $\alpha_{\rm Y}$ and $\alpha_{\rm X}$, formed between Y line and x axis and between X line and x axis, respectively.

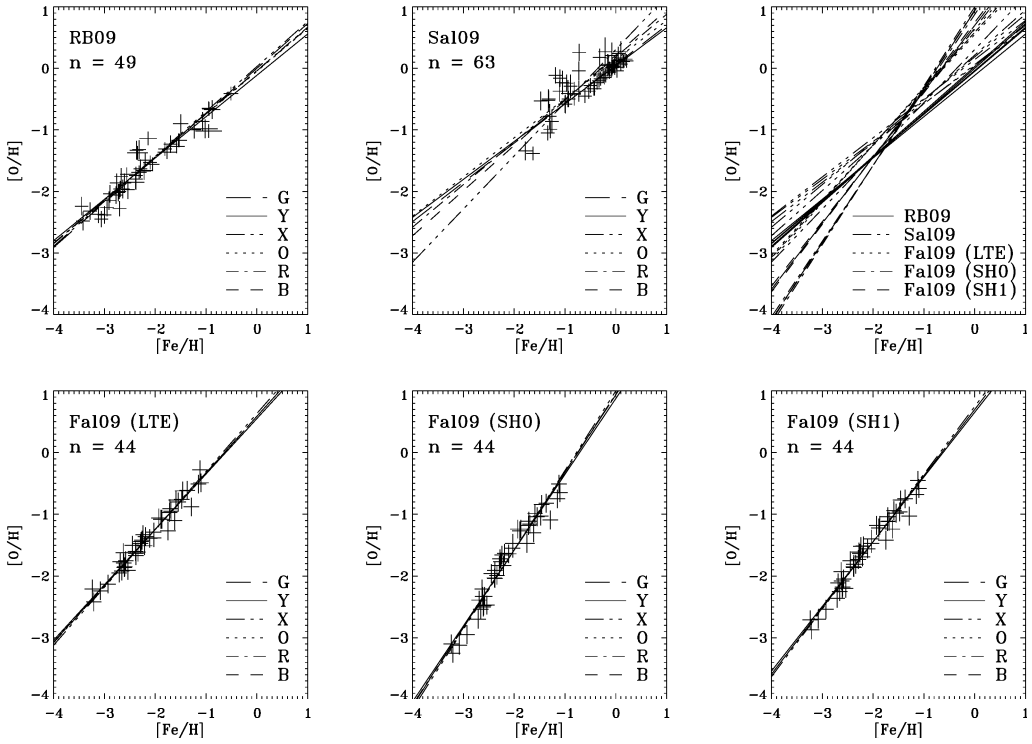


Figure 2: Regression lines related to [O/H]-[Fe/H] empirical relations deduced from two samples with heteroscedastic data, RB09 and Sa09, and three samples with homoscedastic data (using the computer code for heteroscedastic data), Fa09, cases LTE, SH0, and SH1, indicated on each panel together with related population and model captions. The regression lines related to six different methods are shown for each sample on the top right panel. For further details refer to the text.

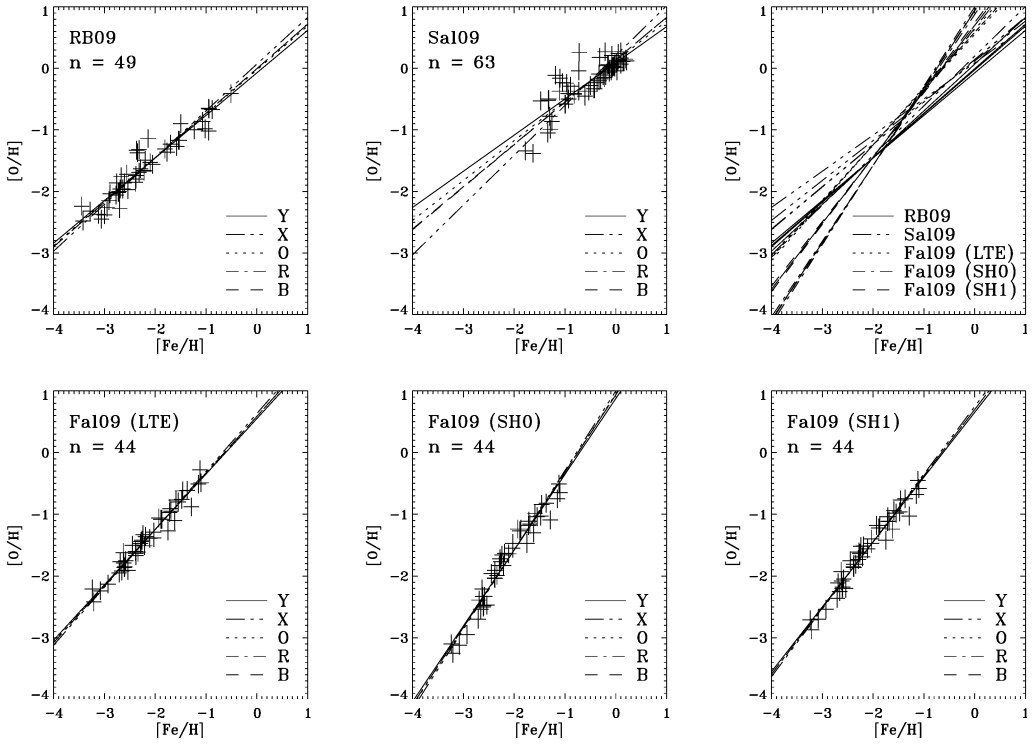


Figure 3: Regression lines related to [O/H]-[Fe/H] empirical relations deduced from two samples with heteroscedastic data (with instrumental scatters taken equal to related typical values), RB09 and Sa09, and three samples with homoscedastic data, Fa09, cases LTE, SH0, and SH1, indicated on each panel together with related population and model captions. The regression lines related to five different methods are shown for each sample on the top right panel. For further details refer to the text.

A Simple But Effective Canonical Dual Theory Unified Algorithm for Global Optimization

Jiapu Zhang

Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Victoria, Australia

Emails: j.zhang@ballarat.edu.au, jiapu zhang@hotmail.com, Phone: 61-423 487 360

Abstract

Numerical global optimization methods are often very time consuming and could not be applied for high-dimensional nonconvex/nonsmooth optimization problems. Due to the nonconvexity/nonsmoothness, directly solving the primal problems sometimes is very difficult. This paper presents a very simple but very effective canonical duality theory (CDT) unified global optimization algorithm. This algorithm has convergence is proved in this paper. More important, for this CDT-unified algorithm, numerous numerical computational results show that it is very powerful not only for solving low-dimensional but also for solving high-dimensional nonconvex/nonsmooth optimization problems, and the global optimal solutions can be easily and elegantly got with zero dual gap.

1 Introduction

In recent years large-scale global optimization (GO) problems have drawn considerable attention. These problems have many applications, in particular in data mining, com-putational biology/chemistry. Numerical methods for GO are often very time consum-ing and could not be applied for high-dimensional nonconvex/nonsmooth optimization problems. Due to the nonconvexity/nonsmoothness, directly solving the primal prob-lems sometimes is very difficult; however, in this paper, their prime-dual Gao-Strang

complementary problems enable us to elegantly and easily not only get the global optimal solutions of primal problems but also of canonical dual problems. The canonical duality theory (CDT) was originally developed for nonconvex/nonsmooth mechanics [7]. It is now realized that this potentially powerful theory can be used for solving a large class of nonconvex/nonsmooth GO problems [3, 9, 10, 12, 13, 14, 15] with applications to data mining clustering, sensor network problems, and molecular distance geometry problem [20], etc.. In this paper, a CDT-unified GO algorithm is presented. According to the CDT, the proof of the convergent theorem for the algorithm is very easy and clear. High-dimensional nonconvex/nonsmooth GO examples (such as the minimization problem of Rosenbrock function) are tested by the new algorithm and numerical results pleasantly show that the new algorithm has a great promise for solving some GO problems.

2 The Algorithm and its Convergence

The algorithm is established from CDT [7]. In this paper we will solve the following nonconvex GO problem [18, 16]:

$$(\mathcal{P}) \qquad \min_x \{P(x) = V(\Lambda(x)) - F(x) | x \in \mathcal{X}_a\}, \tag{1}$$

where $\mathcal{X}_a \subset \mathbb{R}^n$ is a feasible space, $V(\Lambda(x))$ is not necessarily convex with respect to x and it is a so-called canonical function satisfying

$$V^*(\varsigma) = sta\{\langle \xi; \varsigma \rangle - V(\xi) | \xi \in \mathcal{V}\} : \mathcal{V}^* \to \mathbb{R}, \tag{2}$$

$$\varsigma = \nabla V(\xi) \Leftrightarrow \xi \in \nabla V^*(\varsigma) \Leftrightarrow V(\xi) + V^*(\varsigma) = \langle \xi; \varsigma \rangle, \tag{3}$$

Λ is a geometrically admissible (objective) mapping from the feasible space $\mathcal{X}_a$ into a canonical measure space $\mathcal{V}$, $F(x) = \langle x, f \rangle - \frac{1}{2}\langle x, Ax \rangle$, $\langle *, * \rangle$ denotes a bilinear form in $\mathbb{R}^n \times \mathbb{R}^n, f \in \mathbb{R}^n, A = A^T \in \mathbb{R}^{n \times n}$ are given, and $\langle *; * \rangle$ represents a bilinear form which puts $\mathcal{V}$ and $\mathcal{V}^*$ in duality. By (2)-(3), the nonconvex function $P(x)$ in (1) can be rewritten as

$$\Xi(x, \varsigma) = \langle \Lambda(x); \varsigma \rangle - V^*(\varsigma) - F(x). \tag{4}$$

In real applications, the objective operator Λ is usually a quadratic measure over a given a field [7, 18] and in finite space this quadratic measure can be written as [8, 18, 16]:

$$\Lambda(x) = \left\{ \frac{1}{2} x^T B_k x + b_k^T x + c_k = \xi_k, k = 1, 2, \ldots, m \right\} = \xi : \mathbb{R}^n \to \mathbb{R}^m, \tag{5}$$

where $B_k \in \mathbb{R}^{n \times n}$ is a symmetrical matrix, $b_k \in \mathbb{R}^n$, $c_k \in \mathbb{R}^1$ for each $k = 1, 2, \ldots, m$. Denote

$$G(\varsigma) = A + \sum_{k=1}^{m} \varsigma_k B_k : \mathbb{R}^m \to \mathbb{R}^{n \times n}, \tag{6}$$

$$d(\varsigma) = f - \sum_{k=1}^{m} \varsigma_k b_k : \mathbb{R}^m \to \mathbb{R}^n, \tag{7}$$

and define

$$S_a = \{\varsigma \in \mathbb{R}^m | d(\varsigma) \in Col(G(\varsigma))\}, \tag{8}$$

$$S_a^+ = \{\varsigma \in S_a | G(\varsigma) \succeq 0\}, \tag{9}$$

where $Col(G(\varsigma))$ is the space spanned by the column of $G(\varsigma)$, and $G(\varsigma) \succeq 0$ stands for $G(\varsigma)$ is a semi-positive definite matrix (if $detG(\varsigma) = 0, G(\varsigma)^{-1}$ should be understood as the generalized inverse [8, 18, 16]).

The following algorithm is designed:

Algorithm 1 - *A canonical dual theory algorithm.*

Step 1. Call the subroutine Algorithm 2, or simply the *Matlab's fsolve* program if dimension of problems are less than a few thousands, to solve differential equations $\Xi(x,\varsigma)' = 0$.

Step 2. Output the roots $\bar{\varsigma}, \bar{x}$.

Step 3. Check whether $\bar{\varsigma} \in S_a$ are in S_a^+; if so, then

Step 4. Pick up the $\bar{x}$s that are satisfying $G(\bar{\varsigma})\bar{x} = d(\bar{\varsigma})$.

Theorem 1 *(Convergence of the Algorithm) $\bar{x}$ generated by Algorithm 1 is the global optimal solution of $(\mathcal{P})$.*

Proof. By Theorem 3 of [18] and its reference [6] we know that $\bar{x}$ generated by Algorithm 1 is a global optimal solution of $(\mathcal{P})$. ■

The powerful of this simple algorithm can be easily demonstrated by easily and effectively solving the following benchmark test problems of GO, even calculated by hands on paper.

Example 1. (Two-dimensional Rosenbrock function) $P(x) = (x_1 - 1)^2 + 100(x_2 - x_1^2)^2$.
Solution. $\Xi(x,\varsigma) = (x_1 - 1)^2 + (x_1^2 - x_2)\varsigma - \frac{1}{400}\varsigma^2$ and $S_a^+ = \{\varsigma \in \mathbb{R}^1 | \varsigma > -1\}$ are got. Let $\Xi(x,\varsigma)' = 0$ a critical point $(\bar{x}_1, \bar{x}_2, \bar{\varsigma}) = (1, 1, 0)$ is got. We easily know $\bar{\varsigma} = 0 \in S_a^+$ and satisfying $G(\bar{\varsigma})\bar{x} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} = f = d(\bar{\varsigma})$. Thus, $(\bar{x}_1, \bar{x}_2) = (1, 1)$ is a global minimum of Rosenbrock function.

Example 2. (Two-dimensional De Jong function) $P(x) = 100(x_2 - x_1^2)^2 + (x_1 - 1)^2 + (x_2 - 1)^2$.
Solution. $\Xi(x,\varsigma) = (x_1 - 1)^2 + (x_2 - 1)^2 + (x_1^2 - x_2)\varsigma - \frac{1}{400}\varsigma^2$ and $S_a^+ = \{\varsigma \in \mathbb{R}^1 | \varsigma > -1\}$ are got. Let $\Xi(x,\varsigma)' = 0$ a critical point $(\bar{x}_1, \bar{x}_2, \bar{\varsigma}) = (1, 1, 0)$ is got. We easily know $\bar{\varsigma} = 0 \in S_a^+$ and satisfying $G(\bar{\varsigma})\bar{x} = f = d(\bar{\varsigma})$. Thus, $(\bar{x}_1, \bar{x}_2) = (1, 1)$ is a global minimal

solution.

Example 3. (Colville function) $P(x) = 100(x_2 - x_1^2)^2 + 90(x_3^2 - x_4)^2 + (x_1 - 1)^2 + (x_3 - 1)^2 + 10.1((x_2 - 1)^2 + (x_4 - 1)^2) + 19.8(x_2 - 1)(x_4 - 1)$.
Solution. Let $\Xi(x, \varsigma)' = 0$ and a unique critical point $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{\varsigma}_1, \bar{\varsigma}_2) = (1, 1, 1, 1, 0, 0)$ is got such that $(\bar{\varsigma}_1, \bar{\varsigma}_2) \in S_a^+$ and $G(\bar{\varsigma})\bar{x} = f = d(\bar{\varsigma})$. Thus, $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4) = (1, 1, 1, 1)$ is a global minimal solution.

Example 4. (Two-dimensional Zakharov function) $P(x) = x_1^2 + x_2^2 + (0.5x_1 + x_2)^2 + (0.5x_1 + x_2)^4$.
Solution. Let $\Xi(x, \varsigma)' = 0$ and a unique critical point $(\bar{x}_1, \bar{x}_2, \bar{\varsigma}) = (0, 0, 0)$ is got such that $\bar{\varsigma} \in S_a^+$, satisfying the formula $G(\bar{\varsigma})\bar{x} = \begin{pmatrix} 5/2 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = f = d(\bar{\varsigma})$. Thus, (0,0) is the global optimal solution.

Example 5. (Four-dimensional Powell function) $P(x) = (x_1 - 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - x_3)^4 + 10(x_1 - x_4)^4$.
Solution. Let $\Xi(x, \varsigma)' = 0$ and a unique critical point $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{\varsigma}_1, \bar{\varsigma}_2) = (0, 0, 0, 0, 0, 0)$ is got such that $(\bar{\varsigma}_1, \bar{\varsigma}_2) \in S_a^+$ and $G(\bar{\varsigma})\bar{x} = f = d(\bar{\varsigma})$. Thus, $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4) = (0, 0, 0, 0)$ is a global minimal solution.

Example 6. (Double Well function) $P(x) = \frac{1}{2}(\frac{1}{2}x^2 - 2)^2 - \frac{1}{2}x$.
Solution. $\Xi(x, \varsigma) = (\frac{1}{2}x^2 - 2)\varsigma - \frac{1}{2}\varsigma^2 - \frac{1}{2}x$ and $S_a^+ = \{\varsigma \in \mathbb{R}^1 | \varsigma > 0\}$ are got. Let $\Xi(x, \varsigma)' = 0$ and three critical points of $\Xi(x, \varsigma)$: $(\bar{x}^1, \bar{\varsigma}^1) = (2.11491, 0.236417), (\bar{x}^2, \bar{\varsigma}^2) = (-1.86081, -0.268701), (\bar{x}^3, \bar{\varsigma}^3) = (-0.254102, -1.96772)$ are got, and a unique point $\bar{\varsigma}^1$ in S_a^+ is got. Thus, $\bar{x}^1 = 2.11491$ is the global minimal solution of the Double Well function. This can also be illuminated in Figure 1.

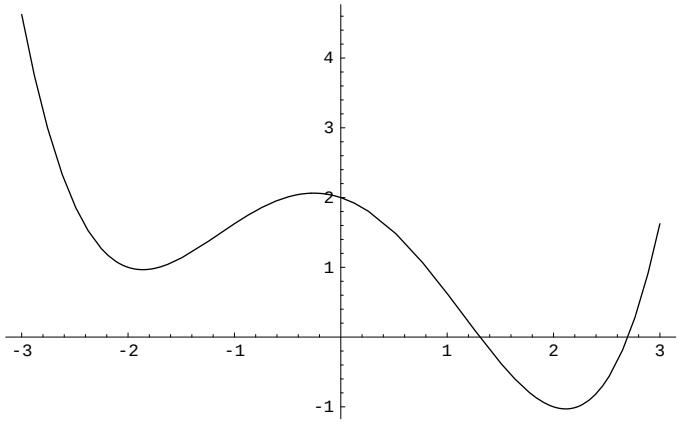


Figure 1: The prime double-well function.

The canonical dual problem of Double Well function is:

$$\max_{\varsigma > 0} -\frac{1}{8\varsigma} - \frac{1}{2}\varsigma^2 - 2\varsigma, \tag{10}$$

which is equal to

$$\min_{\varsigma>0} \frac{1}{8\varsigma} + \frac{1}{2}\varsigma^2 + 2\varsigma \tag{11}$$

that can be transferred into (but not equal to (because of the constant of ς_0)) (this idea is communicated from Wu C.Z.)

$$\min_{t,\varsigma} t \quad subject \quad to \begin{pmatrix} \varsigma & 0.5/\sqrt{2} \\ 0.5/\sqrt{2} & t - \frac{1}{2}\varsigma_0^T\varsigma - 2\varsigma \end{pmatrix} \succeq 0 \tag{12}$$

and solved by Semi-Definite Programming (SDP) package SeDuMi 1_3 [24] (this idea is communicated from Wu C.Z. In this whole paper, this is the single idea of other researchers). However, after 6 circles in all 33 iterations SeDuMi 1_3 died at the solution 0.2585, which is still very far from the real global optimal solution 0.236417 of the dual problem (10). This means the popular SDP package SeDuMi 1_3 has a poor performance for Double Well function minimization problem, compared with our Algorithm 1. But other strategies to solve the following dual problem of (1) are sought:

$$\min_{G(\varsigma)\succeq 0} P^d(\varsigma) \quad = \quad \frac{1}{2}d(\varsigma)^T G(\varsigma)^{-1} d(\varsigma) + \sum_{k=1}^{m}\left(\frac{1}{2}\alpha_k^{-1}\varsigma_k^2 - c_k\varsigma_k\right); \tag{13}$$

for example, one strategy is to replace (13) by the following quadratic semidefinite programming problem (QSDP)

$$\min_{t,\varsigma} \quad t + \frac{1}{2}\alpha^{-1}\varsigma^T\varsigma + +\varsigma^T c \quad subject \quad to \begin{pmatrix} G(\varsigma) & \frac{1}{\sqrt{2}}d(\varsigma) \\ \frac{1}{\sqrt{2}}d(\varsigma)^T & t \end{pmatrix} \succeq 0, \tag{14}$$

where α, c, ς are vectors, and some known QSDP packages can solve (14) for low-dimensional problems and for high-dimensional problems we can design QSDP algorithms by ourselves. This is one direction of algorithm design for CDT. Another direction is to efficiently and effectively get the roots of $\Xi(x,\varsigma)' = 0$ for Algorithm 1, i.e. to solve the $m+n$ quadratic equations (15) given as follows.

For high-dimensional nonconvex GO problems, the finite element method (FEM)-based [25] subroutine of Algorithm 1 is well designed as follows. *Step 1* of Algorithm 1 is to find the roots of $\Xi(x,\varsigma)' = 0$, i.e. the following $m+n$ quadratic equations:

$$\begin{cases} \frac{1}{2}x^T B_k x + b_k^T x + c_k = \alpha_k^{-1}\varsigma_k, k = 1,2,\ldots,m & (by \quad \Xi(x,\varsigma)'_\varsigma = 0), \\ G(\varsigma)x = d(\varsigma), \quad i.e. \quad (A + \sum_{k=1}^m \varsigma_k B_k)x = f - \sum_{k=1}^m \varsigma_k b_k & (by \quad \Xi(x,\varsigma)'_x = 0), \end{cases} \tag{15}$$

where $\alpha_k, k = 1,2,\ldots,m$ are the coefficients in $V(\Lambda(x)) = \sum_{k=1}^m \frac{1}{2}\alpha_k(\frac{1}{2}x^T B_k x + b_k^T x + c_k)^2$.

Algorithm 2 - *A subroutine finding roots of (15). E.g. the finite element discretized $\Xi(x,\varsigma)'$ method if the dimension of the problem is large than a few thousands.*

Firstly the minimization problem for Rosenbrock function for (15) is tested. In global optimization, the nonconvex minimization problem of Rosenbrock function is a benchmark test problem that is extensively used to test the performance of optimization algorithms and approaches. The global minimum is inside a long, deep, narrow, parabolic/banana shaped flat valley. The shallow global minimum is inside a deeply curved valley. To find the valley and to converge to the global minimum is difficult. Detailed introduction to the difficulty of the problem and the excellence as the testing problem of GO can be referred to [19].

Minimizing the Rosenbrock function:

$$\min P(x) = \sum_{k=1}^{n-1}[100(x_k^2 - x_{k+1})^2 + (x_k - 1)^2] : x \in \mathbb{R}^n \tag{16}$$

$$= n - 1 + \sum_{k=1}^{n-1}\frac{1}{2}\alpha_k(x_k^2 - x_{k+1})^2 + \sum_{k=1}^{n-1}x_k^2 - \sum_{k=1}^{n-1}2x_k \tag{17}$$

$$= n - 1 + \sum_{k=1}^{n-1}\frac{1}{2}\alpha_k\left(\frac{1}{2}x^T B_k x + b_k^T x + c_k\right)^2 + \frac{1}{2}x^T Ax - x^T f, \tag{18}$$

where $\alpha_k = 200$, $B_k = \begin{pmatrix} 0 & & & & & & \\ & . & & & & & \\ & & . & & & & \\ & & & 0 & & & \\ & & & & 2 & & \\ & & & & & 0 & \\ & & & & & & . \\ & & & & & & & . \\ & & & & & & & & 0 \end{pmatrix}_{n\times n}$, $b_k = \begin{pmatrix} 0 \\ . \\ . \\ . \\ 0 \\ -1 \\ 0 \\ . \\ 0 \end{pmatrix}_{n\times 1}$, $c_k = 0, k = 1, \ldots, n-1$, $A = \begin{pmatrix} 2 & & & & & & \\ & . & & & & & \\ & & . & & & & \\ & & & . & & & \\ & & & & . & & \\ & & & & & . & \\ & & & & & & 2 \\ & & & & & & & 0 \end{pmatrix}_{n\times n}$ $f = \begin{pmatrix} 2 \\ . \\ . \\ . \\ . \\ . \\ . \\ 2 \\ 0 \end{pmatrix}_{n\times 1}$, $m = n - 1$, and $S_a^+ = \{\varsigma_k > -1, k = 1, 2, \ldots, n-1\}$. For Rosenbrock function, (15) is written as:

$$\begin{cases} x_k^2 - x_{k+1} = 0.005\varsigma_k, k = 1, 2, \ldots, n-1, \\ 2(1 + \varsigma_{k+1})x_{k+1} = 2 + \varsigma_k, k = 1, 2, \ldots, n-3, \\ \varsigma_{n-1} = 0, \\ (1 + \varsigma_1)x_1 = 1, \\ 2x_{n-1} = 2 + \varsigma_{n-2}, \end{cases} \tag{19}$$

(19) is solved by Matlab's *fsolve* on the Intel(R) Celeron(R) CPU 900@2.20GHz Windows Vista$^{\rm TM}$ Home Basic personal notebook computer. The initial solution for $(x;\varsigma)$ is set as $(3,\ldots,3;2,\ldots,2)$ and the numerical computational results are shown in Table 1. We may see in Table 1 that the global minimal solution $(1,\ldots,1)$ for (1) can be easily and accurately got by Algorithm 1, at the same time the global maximal solution $(0,\ldots,0)$ for the (13) over S_a^+ can be easily and accurately got by Algorithm 1 directly too. The smart Matlab's *fsolve* does not output any other solution in S/S_a^+.

Table 1: Results of numerical computations

Dimension n	Iterations	Calls of Equations (19)	prime and dual opt sln		CPU time
3	6	42	$(1,\ldots,1)$	$(0,\ldots,0)$	0.0468
4	6	56	$(1,\ldots,1)$	$(0,\ldots,0)$	0.0468
5	6	70	$(1,\ldots,1)$	$(0,\ldots,0)$	0.0624
6	7	96	$(1,\ldots,1)$	$(0,\ldots,0)$	0.0780
7	7	112	$(1,\ldots,1)$	$(0,\ldots,0)$	0.1248
8	7	128	$(1,\ldots,1)$	$(0,\ldots,0)$	0.1248
9	7	144	$(1,\ldots,1)$	$(0,\ldots,0)$	0.0936
10	7	160	$(1,\ldots,1)$	$(0,\ldots,0)$	0.1404
20	7	320	$(1,\ldots,1)$	$(0,\ldots,0)$	0.2808
30	8	540	$(1,\ldots,1)$	$(0,\ldots,0)$	0.4368
40	8	720	$(1,\ldots,1)$	$(0,\ldots,0)$	0.6084
50	8	900	$(1,\ldots,1)$	$(0,\ldots,0)$	0.8736
60	8	1,080	$(1,\ldots,1)$	$(0,\ldots,0)$	1.0920
70	8	1,260	$(1,\ldots,1)$	$(0,\ldots,0)$	1.2948
80	8	1,440	$(1,\ldots,1)$	$(0,\ldots,0)$	1.6536
90	8	1,620	$(1,\ldots,1)$	$(0,\ldots,0)$	1.9344
100	8	1,800	$(1,\ldots,1)$	$(0,\ldots,0)$	2.2464
200	9	4,000	$(1,\ldots,1)$	$(0,\ldots,0)$	7.7844
300	9	6,000	$(1,\ldots,1)$	$(0,\ldots,0)$	15.8497
400	9	8,000	$(1,\ldots,1)$	$(0,\ldots,0)$	27.4250
500	9	10,000	$(1,\ldots,1)$	$(0,\ldots,0)$	42.4947
600	9	12,000	$(1,\ldots,1)$	$(0,\ldots,0)$	61.2616
700	9	14,000	$(1,\ldots,1)$	$(0,\ldots,0)$	88.2342
800	9	16,000	$(1,\ldots,1)$	$(0,\ldots,0)$	118.3736
900	10	19,800	$(1,\ldots,1)$	$(0,\ldots,0)$	170.8367
1,000	10	22,000	$(1,\ldots,1)$	$(0,\ldots,0)$	221.8334
2,000	10	44,000	$(1,\ldots,1)$	$(0,\ldots,0)$	1491.9

Chemical database clustering problem. Algorithm 1 is applied to solve the real chemical database clustering problem (34)-(35) of [27], i.e.

$$\min P(Y_1,Y_2,\ldots,Y_n) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{2}\alpha_{ij}\left(||Y_i-Y_j||^2-d_{ij}^2\right)^2 \quad (20)$$

$$= \sum_{i,j=1,i<j}^{n}\frac{1}{2}\alpha_{ij}\left((y_{i1}-y_{j1})^2+(y_{i2}-y_{j2})^2-d_{ij}^2\right)^2, \quad (21)$$

where $\alpha_{ij} = \frac{1}{2d_{ij}^4}$ if $d_{ij}^4 \geq 10^{-12}$, $\alpha_{ij} = \frac{1}{2}$ if $d_{ij}^4 < 10^{-12}$, for all $i, j = 1, 2, \ldots, n$ (d_{ij} is calculated from the original n by 9 dataset [27]) and $Y_i \in \mathbb{R}^2, i = 1, 2, \ldots, n$. The *Gao-Strang generalized complementary function* for (21) is:

$$\Xi(Y, \varsigma) = \sum_{i,j=1,i<j}^{n} \left((||Y_i - Y_j||^2 - d_{ij}^2)\varsigma_{ij} - \frac{1}{2\alpha_{ij}}\varsigma_{ij}^2 \right) \tag{22}$$

$$= \sum_{i,j=1,i<j}^{n} \left(\left((y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2 - d_{ij}^2\right) \varsigma_{ij} - \frac{1}{2\alpha_{ij}}\varsigma_{ij}^2 \right). \tag{23}$$

Choose the SVD-reduced initial solution $Y_i^0, i = 1, 2, \ldots, n$ [27, 29] and calculate $\varsigma_{ij}^0 = \alpha_{ij}(||Y_i^0 - Y_j^0||^2 - d_{ij}^2)$ as the initial solution for ς_{ij}^0 in solving the following quadratic nonlinear equations of $\Xi(Y, \varsigma)'$=0:

$$\begin{cases} 2(y_{i1} - y_{j1})\varsigma_{ij} = 0, i = 1, 2, \ldots, n-1, j = i+1, \ldots, n, \\ -2(y_{i1} - y_{j1})\varsigma_{ij} = 0, i = 1, 2, \ldots, n-1, j = i+1, \ldots, n, \\ 2(y_{i2} - y_{j2})\varsigma_{ij} = 0, i = 1, 2, \ldots, n-1, j = i+1, \ldots, n, \\ -2(y_{i2} - y_{j2})\varsigma_{ij} = 0, i = 1, 2, \ldots, n-1, j = i+1, \ldots, n, \\ (y_{i1} - y_{j1})^2 + (y_{i2} - y_{j2})^2 - d_{ij}^2 - \frac{1}{\alpha_{ij}}\varsigma_{ij} = 0, i = 1, 2, \ldots, n-1, j = i+1, \ldots, n. \end{cases} \tag{24}$$

The global optimal prime or dual solution is got and the comparison of Algorithm 1 with the SD (steepest descent) method, CG (conjugate gradient) method, BFGS (approximated Newton) method, NR (Newton-Raphson) method, and TN (truncated Newton) method T-IHN (truncated incomplete Hessian Newton) method, etc can be made. The equations (24) are illuminated to make readers to easily understand the equations. When $n = 2$, i.e. $i = 1, j = 2$ there are the following 5 quadratic equations with 5 variables (with initial value $(y_{11}^0, y_{12}^0, y_{21}^0, y_{22}^0; \varsigma_{12}^0) = (331.5590, -188.4908, 364.6889, -158.1010; 0)$):

$$\begin{cases} 2(y_{11} - y_{21})\varsigma_{12} = 0, \\ -2(y_{11} - y_{21})\varsigma_{12} = 0, \\ 2(y_{12} - y_{22})\varsigma_{12} = 0, \\ -2(y_{12} - y_{22})\varsigma_{12} = 0, \\ (y_{11} - y_{21})^2 + (y_{12} - y_{22})^2 - 15701874.4205486\varsigma_{12} - 2801.95239257813 = 0, \end{cases} \tag{25}$$

In (24), different database of [29] has different n. Numerical computational results of solving (24) will be updated.

For CDT, there are two research directions for its algorithm design. One is to design the CDT algorithm to solve (13); for example, one strategy is to design the quadratic semidefinite programming (QSDP) algorithm to solve (14). Another research direction is to design the CDT algorithm to solve the special m + n quadratic (nonlinear) equations, with m+n variables, (15); for example, Newton-type iteration algorithms, gradient-type iteration algorithms, trust-region-type iteration algorithms, nonlinear finite-element-type algorithms, Hamiltonian system symplectic-type algorithms, etc are good strategy to solve (15). Researchers may design a powerful CDT algorithm along these two directions to solve (13) and (15) respectively.

3 Advantages of Solving the Dual Problem Than Solving the Prime Problem

In this section, we still use the benchmark GO test problem, minimizing the Rosenbrock function, to illuminate some advantages of solving the canonical dual problem (13) compared with directly solving the prime problem (1). By the CDT [7, 16], the canonical dual problem of (16) is:

$$\max_{\varsigma>-1} P^d(\varsigma) = n - 1 - \sum_{k=1}^{n-1} \left[\frac{(\varsigma_{k-1}+2)^2}{4(\varsigma_k+1)} + \frac{1}{400}\varsigma_k^2 \right], \tag{26}$$

where $\varsigma_0 = 0$. (26) is solved by the Discrete Gradient (DG) method [1], a local search optimization solver for nonconvex and/or nonsmooth optimization problems, and the numerical computational results are listed in Tables 2-3 (where seed1 is the initial solution $(x^0;\varsigma^0) = (3,3,\ldots,3;-2/3,-2/3,\ldots,-2/3,0)$ and seed2 is the initial solution $(x^0;\varsigma^0) = (100,100,\ldots,100;100,100,\ldots,100,0)$).

In Tables 2-3, we may see that the dual problem (26) can be elegantly, easily, quickly and accurately solved to get its objective function value 0.00000000, compared with the prime problem (16). We may also see that (26) is convenient for MPI (Message Passing Interface) parallel computation. The successfully tested MPI code is followed:

```
broadcast n − 1
    call MPI_BCAST (n − 1,1,MPI_INTEGER, 0, MPI_COMM_WORLD , ierr)
check for quit signal
    if ( n − 1 .le. 0 ) goto 30
calculate every partials
    sum = 0.0d0
    do 20 i = myid+1, n − 1, numprocs
        if (i − 1 .eq. 0) then ς(0)=0
        sum = sum +(ς(i − 1) + 2.0)/(4 ∗ (ς(i) + 1.0)) + (1.0/400.0) ∗ ς(i) ∗ ∗2
    20 continue
    f = sum
collect all the partial sums
    call MPI_REDUCE (f,objf,1,MPI_DOUBLE_PRECISION, MPI_SUM, 0,
    & MPI_COMM_WORLD, ierr )
30 node 0 (i.e. myid = 0) prints the sums = objf
```

Table 2: Results of numerical experiments for seed1

Dimension n	Iterations		Function calls		Objective function value	
	Prime	Dual	Prime	Dual	Prime	Dual
2	120	24	2843	28	0.00001073	0.00000000
3	422	26	8996	137	0.00401438	0.00000000
4	3737	35	48352	202	0.00615273	0.00000000
5*	335	34	10179	399	3.96077434	0.00000000
6*	2375	44	43770	868	4.00635895	0.00000000
7*	1223	53	28009	1625	4.09419146	0.00000000
8	2160	55	46792	2100	0.01246714	0.00000000
9	2692	51	61017	2526	0.01397307	0.00000000
10	4444	63	91470	3979	0.01055630	0.00000000
20	3042	55	140924	10084	0.00940077	0.00000000
30	2321	58	133980	20515	0.01075478	0.00000000
40	1659	60	173795	26818	0.01227866	0.00000000
50	2032	57	219233	36459	0.01264147	0.00000000
60	1966	61	260701	50495	0.01048188	0.00000000
70	1876	56	272919	52545	0.01531147	0.00000000
80	1405	61	195156	59684	0.01594730	0.00000000
90	2142	61	371963	71320	0.01055831	0.00000000
100	2676	60	510722	70208	0.01125514	0.00000000
200	1395	61	653604	188589	0.01115318	0.00000000
300	1368	60	882760	235163	0.01574873	0.00000000
400	2085	66	1869675	301805	0.00928066	0.00000000
500	1155	59	1394240	358938	0.01168440	0.00000000
600	1226	63	1808285	451817	0.00918730	0.00000000
700	1557	60	2134359	559378	0.01257100	0.00000000
800	1398	61	2098062	522726	0.01442714	0.00000000
900	716	65	1904187	763449	0.01074534	0.00000000
1000	1825	61	3598608	681509	0.00897202	0.00000000
2000	257	62	2087277	1455472	0.00937219	0.00000000
3000	3221	60	20642543	2714296	0.01250373	0.00000000
4000*	679	60	7581502	3659292	4.11193171	0.00000000

This shows the advantages of solving the dual problem (13) than solving the prime problem (1). We may use maximizing the canonical dual function of Colville function (see Example 3) to illuminate this point again.

It is known that directly solving the minimizing of Colville function is difficult. But, it is very easy to solving the following canonical dual problem:

$$\max_{\varsigma>-1} P^d(\varsigma) = 42 - \frac{1}{400}\varsigma_1^2 - \frac{1}{360}\varsigma_2^2$$
$$-\frac{1}{2}\begin{pmatrix} 2 \\ 40+\varsigma_1 \\ 2 \\ 40+\varsigma_2 \end{pmatrix}^T \begin{pmatrix} 2+2\varsigma_1 & 0 & 0 & 0 \\ 0 & 20.2 & 0 & 19.8 \\ 0 & 0 & 2+2\varsigma_2 & 0 \\ 0 & 19.8 & 0 & 20.2 \end{pmatrix}^+ \begin{pmatrix} 2 \\ 40+\varsigma_1 \\ 2 \\ 40+\varsigma_2 \end{pmatrix},$$

Table 3: Results of numerical experiments for seed2

Dimension n	Iterations		Function calls		Objective function value	
	Prime	Dual	Prime	Dual	Prime	Dual
2	10013	24	227521	28	47.23824896	0.00000000
3	144	32	4869	235	96.49814330	0.00000000
4	144	81	5279	938	82.46230602	0.00000000
5	148	137	5682	1768	94.19254867	0.00000000
6	154	166*	6238	2590	88.84382963	0.00000000
7	159	179*	7097	3288	237.63078399	0.00000000
8	165	202*	7502	4300	238.41126013	0.00000000
9	153	206*	7137	5083	84.54205412	0.00000000
10	162	216*	7491	5920	83.23094398	0.00000000
20	225	285*	19111	17458	83.94779152	0.00000000
30	216	301*	20939	28543*	156.95838274	0.00000000
40	163	291*	19775	40444*	83.30960344	0.00000000
50	158	298*	33269	51888*	85.93091895	0.00000000
60	158	312*	34094	61767*	89.07412094	0.00000000
70	162	284*	35436	69865*	92.45725362	0.00000000
80	209	297*	35607	89127*	157.69955825	0.00000000
90	227	294*	60398	98748*	82.44035053	0.00000000
100	202	290*	57792	102796*	81.94595276	0.00000000
200	1826	262	436413	189293	83.77165551	0.00000000
300	195	259*	169238	261320*	152.95671738	0.00000000
400	195	278*	212104	375816*	82.49253919	0.00000000
500	190	297*	331637	522695*	82.40170647	0.00000000
600	292	303*	431092	559068*	150.15456693	0.00000000
700	189	275*	383735	758631*	89.14575473	0.00000000
800	198	270*	429674	701053*	84.50538257	0.00000000
900	198	280*	416150	867398*	85.32757049	0.00000000
1000	193	283*	445326	930761*	89.48369379	0.00000000
2000	232	310*	1123240	2030104*	84.26810981	0.00000000

which is a simple problem of maximizing a concave function over a convex set: by watching Figure 2, $\bar{\varsigma} = (0, 0)$ is easily known as the global optimal solution.

Thus, the following optimization algorithm designed to solve the canonical dual problem (14) is very necessary.

Algorithm 3 - *An optimization algorithm to solve the Quadratic Semidefinite Programming (14).*

In Example 6, (11) is equal to

$$\begin{array}{ll} \min\limits_{t,\varsigma\in\mathbb{R}} & \frac{1}{2}\begin{pmatrix} t \\ \varsigma \end{pmatrix}^T \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} t \\ \varsigma \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}^T \begin{pmatrix} t \\ \varsigma \end{pmatrix} \\ s.t. & \begin{pmatrix} \varsigma & 0.5/\sqrt{2} \\ 0.5/\sqrt{2} & t \end{pmatrix} \succeq 0, \end{array}$$

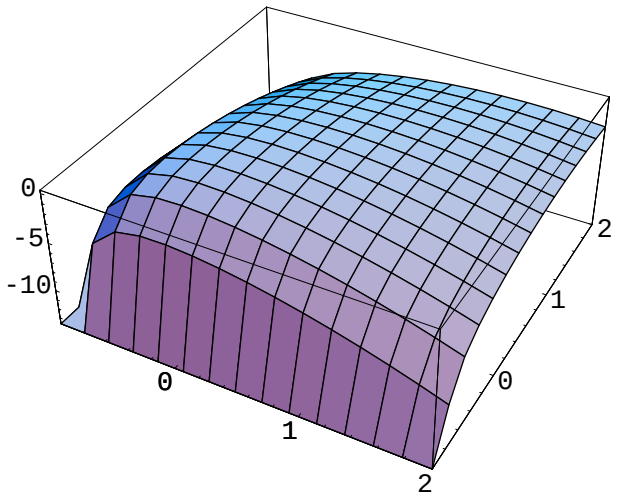


Figure 2: Canonical dual Colville function on $S_a^+ = \{\varsigma > -1\}$

which can solved by Algorithm 3.

References

[1] Bagirov, A.M., Karasozen, B., Sezer, M.: Discrete gradient method: a derivative free method for nonsmooth optimization. J. Opt. Theor. Appl. 137: 317-334 (2008)

[2] Bagirov, A.M., Rubinov, A.M., Zhang, J.P.: Local optimization method with global multidimensional search. J. Glob. Optim. 32: 161–179 (2005)

[3] Fang, S.C., Gao, D.Y., Sheu, R.L., Wu, S.Y.: Canonical dual approach to solving 0-1 quadratic programming problems. J. Ind. Manag. Opt. 4 (1): 125–142 (2008)

[4] Feng, K.: Numerical Computational Algorithms. National Defence Industrial Press (1978) (in Chinese)

[5] Feng, K., Qin M.Z.: Hamiltonian Symplectic System. Zhejiang Science and Technology Press (2003) (in Chinese)

[6] Gao, D.Y.: Dual extremum principles in finite deformation theory with applications to post-buckling analysis of extended nonlienar beam theory. Appl. Mech. Rev. 50 (11), S64-S71 (1997)

[7] Gao, D.Y.: Duality Principles in Nonconvex Systems: Theory, Methods and Applications. Kluwer Academic Publishers, Dordrecht/Boston/London (2000)

[8] Gao, D.Y.: Canonical dual transformation method and generalized triality theory in nonsmooth global optimization. J. Glob. Optim. 17 (1), 127-160 (2000)

[9] Gao, D.Y.: Perfect duality theory and complete set of solutions to a class of global optimization problems. Opt. 52 (4–5): 467–493 (2003)

[10] Gao, D.Y.: Canonical duality theory and solutions to constrained nonconvex quadratic programming. J. Glob. Opt. 29: 377–399, (2004)

[11] Gao, D.Y.: Complementary principle, algorithm, and complete solution to phase transitions in solids governed by Landau-Ginzburg equation. Math. Mech. Solids 9: 61-79 (2004)

[12] Gao, D.Y.: Sufficient conditions and perfect duality in nonconvexs minimization with inequality. J. Ind. Manag. Opt. 1 (1): 53-63 (2005)

[13] Gao, D.Y.: Complete solutions and extremality criteria to polynomial optimization problems, J. Glob. Opt. 35 : 131-143 (2006)

[14] Gao, D.Y.: Solutions and optimality criteria to box constrained nonconvex minimization problems, J. Ind. Manag. Opt. 3(2), 293-304 (2007)

[15] Gao, D.Y.: Advances in canonical duality theory with applications to global optimization. Proceedings Foundations of Computer-Aided Process Operations (FOCAPO 2008), June 29July 2, 2008, Cambridge, MA (2008)

[16] Gao, D.Y., Ruan, N., Pardalos, P.: Canonical dual solutions to sum of fourth-order polynomials minimization problems with applications to sensor network localization (2010)

[17] Gao, D.Y., Ruan, N., Zhang, J.: Canonical dual solutions to a global optimization problem in database analysis (2011)

[18] Gao, D.Y., Wu, C.Z.: On the triality theory in global optimization, J. Glob. Opt. (2011) accepted, arXiv:1104.2970v1

[19] Gao, D.Y., Zhang, J.P.: Canonical duality theory for solving minimization problem of Rosenbrock function, J. Glob. Opt. (2011) in revision

[20] Gao, D.Y., Zhang, J.P.: A novel canonical dual computational approach for prion AGAAAAGA amyloid fibril molecular modeling. J. Theor. Biol.: in revision (2011)

[21] Ghosh, R., Rubinov, A.M., Zhang, J.P.: Optimization approach for clustering datasets with weights. Optim. Method Softw. 20 (2) 335–351 (2005)

[22] Li, Q.Y., Mo, Z.Z., Qi, L.Q.: Numerical Solutions for the Systems of Nonlinear Equations Beijing: Seience Press (1987) (in Chinese)

[23] Lu, C., Wang, Z.B., Xing, W.X., Fang, S.C.: Extended canonical duality and conic programming for solving 0-1 quadratic programming problem, J. Ind. Manag. Optim. 6 (4): 779-793 (2010)

[24] Lu, W.S.: Use SeDuMi to Solve LP, SDP and SCOP Problems: Remarks and Examples (December 2, 2009 Version):
http://www.ece.uvic.ca/∼wslu/Talk/SeDuMi-Remarks.pdf

[25] Santosa, H.A.F.A, Gao, D.Y.: Canonical dual finite element method for solving post-buckling problems of a large deformation elastic beam. Int. J. Nonlinear Mech. (to appear) (2011)

[26] Sun, K., Tu, S.K., Gao, D.Y., Xu, L.: Canonical dual approach to binary factor analysis. The Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation 2009, Paraty, Brazil, March 15–18, 2009.

[27] Xie, D.X. and Ni, Q.: An incomplete Hessian Newton minimization method and its application in a chemical database problem. Comput. Optim. Appl. 44: 467–485 (2009)

[28] Xie, D.X., Singh, S.B., Fluder, E.M. and Schlick, T.: Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases. Math. Program. 95: 161–185 (2003)

[29] Xie, D.X., Tropsha, A. and Schlick, T.: An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-newton minimization. J. Chem. Inf. Comput. Sci. 40: 167–177 (2000)

Canonical Duality Theory for Solving Minimization Problem of Rosenbrock Function

David Yang Gao · Jiapu Zhang

Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Victoria, Australia

Abstract This paper presents a *canonical duality theory* for solving nonconvex minimization problem of Rosenbrock function. Extensive numerical results show that this benchmark test problem can be solved precisely and efficiently to obtain global optimal solutions.

Keywords global optimization · canonical duality · NP-hard problems · triality

1 Introduction

Nonconvex minimization problem of Rosenbrock function, introduced in [15], is a benchmark test problem in global optimization that has been used extensively to test performance of optimization algorithms and numerical approaches. The global minimizer of this function is located in a long, deep, narrow, parabolic/banana shaped flat valley (Figure 1).

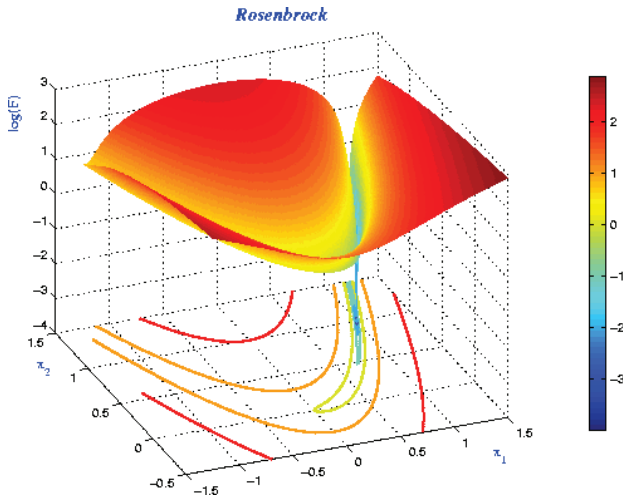


Figure 1: 2-dimensional Rosenbrock function (www2.imm.dtu.dk/courses/02610/)

Although to find this valley is trivial for most cases, to accurately locate the global optimal solution is very difficult by almost all gradient-type methods and some derivative-free methods. Due to the nonconvexity, it can be easily tested that if the initial point is chosen to be $(3, 3, \ldots, 3)$, direct algorithms are always trapped into a local minimizer for problems with dimensions $n = 5 \sim 7$ as well as $n \geq 4000$; if the initial point is chosen at $(100, 100, \ldots, 100)$, iterations will be stopped at a local min with the objective function value > 47.23824896 even for a two-dimensional problem. This paper will show that by the canonical duality theory, this well-known benchmark problem can be solved efficiently in an elegant way.

The *canonical duality theory* was originally developed in nonconvex/nonsmooth mechanics [9]. It is now realized that this potentially powerful theory can be used for solving a large class of nonconvex/nonsmooth/discrete problems [10, 12]. In this short research note, we will first show the nonconvex minimization problem of Rosenbrock function can be reformulated as a canonical dual problem (with zero duality gap) and the critical point of the Rosenbrock function can be analytically expressed in terms of its canonical dual solutions. Both global and local extremal solutions can be identified by the triality theorem. Extensive numerical examples and discussion are presented in the last section.

2 Primal Problem and Its Canonical Dual

The primal problem is

$$(\mathcal{P}): \quad \min \left\{ P(\mathbf{x}) = \sum_{i=1}^{n-1} \left[(x_i - 1)^2 + \frac{1}{2}\alpha (x_{i+1} - x_i^2)^2 \right] \;\; | \;\; \mathbf{x} \in \mathcal{X} \right\}, \tag{1}$$

where $\mathbf{x} = \{x_i\} \in \mathcal{X} = \mathbb{R}^n$ is a real unknown vector, $\alpha = 2N$ and N is a given real number. Clearly, this is a nonconvex minimization problem which could have multiple local minimizers.

In order to use the canonical duality theory for solving this nonconvex problem, we need to define a *geometrically admissible* canonical measure

$$\boldsymbol{\xi} = \{\xi_i\} = \{x_i^2 - x_{i+1}\} \in \mathcal{E}_a \subset \mathbb{R}^{n-1}. \tag{2}$$

The canonical function $V : \mathcal{E}_a \to \mathbb{R}$ can be defined by

$$V(\boldsymbol{\xi}) = \sum_{j=1}^{n-1} \frac{1}{2}\alpha\xi_j^2, \tag{3}$$

which is a convex function. The canonical dual variable $\boldsymbol{\varsigma} = \boldsymbol{\xi}^*$ can be defined uniquely by

$$\boldsymbol{\varsigma} = \{\varsigma_j\} = \nabla V(\boldsymbol{\xi}) = \{\alpha\xi_j\}. \tag{4}$$

Therefore, by the Legendre transformation, the conjugate function $V^* : \mathcal{S} = \mathbb{R}^{n-1} \to \mathbb{R}$ is obtained as

$$V^*(\boldsymbol{\varsigma}) = \text{sta}\{\boldsymbol{\xi}^T\boldsymbol{\varsigma} - V(\boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \mathcal{E}_a\} = \sum_{j=1}^{n-1} \frac{1}{2}\alpha^{-1}\varsigma_j^2. \tag{5}$$

Replacing $\sum_{i=1}^{n-1} \frac{1}{2}\alpha(x_{i+1} - x_i^2)^2$ by the Legendre equality $V(\Lambda(\mathbf{x})) = \Lambda(\mathbf{x})^T\boldsymbol{\varsigma} - V^*(\boldsymbol{\varsigma})$, the total complementary function $\Xi : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ is given by

$$\Xi(\mathbf{x}, \boldsymbol{\varsigma}) = \sum_{i=1}^{n-1} \left[(x_i - 1)^2 + \varsigma_i(x_i^2 - x_{i+1}) - \frac{1}{2}\alpha^{-1}\varsigma_i^2\right]. \tag{6}$$

Let $\delta^\sharp$ and $\delta^\flat$ be shifting operators such that $\delta^\sharp\varsigma_i = \varsigma_{i+1}$ and $\delta^\flat\varsigma_i = \varsigma_{i-1}$. We define $\delta^\flat\varsigma_1 = 0$. Then on the canonical dual feasible space

$$\mathcal{S}_a = \{\boldsymbol{\varsigma} \in \mathcal{S} |\ \ \varsigma_i + 1 \neq 0\ \ \forall i = 1, \ldots, n-2,\ \varsigma_{n-1} = 0\}, \tag{7}$$

the canonical dual can be obtained by

$$P^d(\boldsymbol{\varsigma}) = \text{sta}\{\Xi(\mathbf{x}, \boldsymbol{\varsigma}) |\ \mathbf{x} \in \mathcal{X}\} = n - 1 - \sum_{i=1}^{n-1} \left[\frac{(\delta^\flat\varsigma_i + 2)^2}{4(\varsigma_i + 1)} + \frac{1}{2}\alpha^{-1}\varsigma_i^2\right]. \tag{8}$$

Based on the *complementary-dual principle* proposed in the canonical duality theory (see [**?**]), we have the following result.

Theorem 1 *If $\bar{\boldsymbol{\varsigma}}$ is a critical point of $P^d(\boldsymbol{\varsigma})$, then the vector $\bar{\mathbf{x}} = \{\bar{x}_i\}$ defined by*

$$\bar{x}_i = \frac{\delta^\flat\bar{\varsigma}_i + 2}{2(\bar{\varsigma}_i + 1)},\ \ i = 1, \cdots, n-1,\ \ \bar{x}_n = \bar{x}_{n-1}^2 \tag{9}$$

is a critical point of $P(\mathbf{x})$ and

$$P(\bar{\mathbf{x}}) = \Xi(\bar{\mathbf{x}}, \bar{\boldsymbol{\varsigma}}) = P^d(\bar{\boldsymbol{\varsigma}}). \tag{10}$$

This theorem presents actually an "analytic" solution form to the Rosenbrock function, i.e. the critical point of the Rosenbrock function must be in the form of (9) for each dual solution $\bar{\varsigma}$. The first version of this analytical solution form was presented in nonconvex variational problems in phase transitions and finite deformation mechanics [5, 6, 7]. The extremality of the analytical solution is governed by the so-called *triality theory*. Let

$$\mathcal{S}_a^+ = \{\boldsymbol{\varsigma} \in \mathcal{S}_a | \ \varsigma_i + 1 > 0 \ \ \forall i = 1, \dots, n-1\}, \tag{11}$$

we have the following theorem:

Theorem 2 *Suppose that $\bar{\boldsymbol{\varsigma}}$ is a critical point of $P^d(\boldsymbol{\varsigma})$ and the vector $\bar{\mathbf{x}} = \{\bar{x}_i\}$ is defined by Theorem 1.*

If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^+$, then $\bar{\boldsymbol{\varsigma}}$ is a global maximal solution to the canonical dual problem on $\mathcal{S}_a^+$, i.e.,

$$(\mathcal{P}_+^d): \ \ \max\{P^d(\boldsymbol{\varsigma}) \mid \boldsymbol{\varsigma} \in \mathcal{S}_a^+\}, \tag{12}$$

the vector $\bar{\mathbf{x}}$ is a global minimal to the primal problem, and

$$P(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}) = \max_{\boldsymbol{\varsigma} \in \mathcal{S}_a^+} P^d(\boldsymbol{\varsigma}) = P^d(\bar{\boldsymbol{\varsigma}}). \tag{13}$$

Theorem 2 shows that the canonical dual problem $(\mathcal{P}_+^d)$ provides a global optimal solution to the nonconvex primal problem. Since $(\mathcal{P}_+^d)$ is a concave maximization problem over a convex space which can be solved easily. This theorem is actually a special application of Gao and Strang's general result on global minimizer in in nonconvex analysis [13].

By introducing

$$\mathcal{S}_a^- = \mathcal{S}/\mathcal{S}_a^+ = \{\boldsymbol{\varsigma} \in \mathbb{R}^{n-1} | \ \varsigma_i + 1 < 0 \ \ \forall i = 1, \dots, n-1\}, \tag{14}$$

recently the triality theory (see [14]) leads to the following theorem.

Theorem 3 *Suppose that $\bar{\boldsymbol{\varsigma}}$ is a critical point of $P^d(\boldsymbol{\varsigma})$ and the vector $\bar{\mathbf{x}} = \{\bar{x}_i\}$ is defined by Theorem 1.*

If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$, then on a neighborhood $\mathcal{X}_o \times \mathcal{S}_o \subset \mathcal{X} \times \mathcal{S}_a^-$ of $(\bar{\mathbf{x}}, \bar{\boldsymbol{\varsigma}})$, we have either

$$P(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in \mathcal{X}_o} P(\mathbf{x}) = \min_{\boldsymbol{\varsigma} \in \mathcal{S}_o} P^d(\boldsymbol{\varsigma}) = P^d(\bar{\boldsymbol{\varsigma}}), \tag{15}$$

or

$$P(\bar{\mathbf{x}}) = \max_{\mathbf{x} \in \mathcal{X}_o} P(\mathbf{x}) = \max_{\boldsymbol{\varsigma} \in \mathcal{S}_o} P^d(\boldsymbol{\varsigma}) = P^d(\bar{\boldsymbol{\varsigma}}), \tag{16}$$

The proof of this Theorem can be derived from the recent paper by Gao and Wu [14]. By the fact that the canonical dual function is a d.c. function (difference of convex functions) on $\mathcal{S}_a^-$, the double-min duality (15) can be used for finding the biggest local minimizer of the Rosenbrock function $P(\mathbf{x})$, while the double-max duality (16) can be used for finding the biggest local maximizer of $P(\mathbf{x})$. In physics and material sciences, this pair of biggest local extremal points play important roles in phase transitions.

Because $\varsigma_{n-1} = 0$, we may know that $\mathcal{S}_a^-$ is an empty set. Thus, by Theorem 3 in this paper we cannot find a local maximizer or minimizer on $\mathcal{S}_a^-$ or its subset for $P^d(\bar{\boldsymbol{\varsigma}})$.

3 Numerical Examples and Discussion

$(\mathcal{P})$ and $(\mathcal{P}_+^d)$ will be solved by the discrete gradient (DG) method ([2]), which is a local search optimization solver for nonconvex and/or nonsmooth optimization problems. In two dimensional space, Rosenbrock function has a long ravine with very steep walls and flat bottom; "because of the curved flat valley the optimization is zig-zagging slowly with small stepsizes towards the minimum" (en.wikipedia.org/wiki/Gradient_descent). This means any gradient method may fail to minimize the Rosenbrock function even from 2 dimensions. The DG method is a derivative-free method which can be applied for miminizing/maximizing Rosenbrock function and its dual. Numerical experiments have been carried out in Intel(R) Celeron(R) CPU 900@2.20GHz Windows Vista$^{\text{TM}}$ Home Basic personal notebook computer.

We try N=100 (when N=10 we find the numerical results are similar to $N = 100$), with the dimensions n=2∼10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000. We first set $(3, 3, \ldots, 3)$ (called seed1) as the initial solution for $(\mathcal{P})$ (usually the feasible solution space is a box constrained between -2.048 and 2.048 [1, 16, 17]). Numerical results (Table 1) show that to solve the primal problem $(\mathcal{P})$, the DG method can easily and quickly get approximate global minimum solution to $\bar{\mathbf{x}} = (1, 1, \ldots, 1)$ with the approximate global optimal values at $P(\bar{\mathbf{x}}) = 0$, except for n=5∼7, 4000, where the DG method can only get a local minimum solution $\bar{\mathbf{x}} = (-1, 1, \ldots, 1)$ with $P(\bar{\mathbf{x}}) = 4$. Then we let $\mathbf{x}_0 = (100, 100, \ldots, 100)$ (called seed2) be the initial solution for $(\mathcal{P})$, searched in the intervals $-500 \leq x_i \leq 500, i = 1, 2, \ldots, n$. We find that the DG method was trapped into local optimal solutions but not getting any global minimum at all, even from a 2 dimensional problem (see Table 2), its objec-

tive function value is 47.23824896. However, from Table 2 we can see that by the same DG method, the global maximum of the dual problem can be obtained very elegantly.

For $(\mathcal{P}_+^d)$, the corresponding dimensions are 1∼9, 19, 29, 39, 49, 59, 69, 79, 89, 99, 199, 299, 399, 499, 599, 699, 799, 899, 999, 1999, 2999, 3999. The initial solution is set as $\boldsymbol{\varsigma}_0 = (-2/3, -2/3, \ldots, -2/3, 0)$ (called seed1), the constraints $\boldsymbol{\varsigma}_i + 1 > 0, i = 1, 2, \ldots, n-1$ were penalized into the objective function; by $\Xi(\mathbf{x}, \boldsymbol{\varsigma})'_{x_n} = 0$ of formula (6), we can set the values of the last variable ς_{n-1} always being 0 (> -1). With these numerical computation settings, the DG method can easily and quickly solve all these test problems to accurately get a global maximizer $\bar{\boldsymbol{\varsigma}} = (0, 0, ..., 0)$ with the optimal value $P^d(\bar{\boldsymbol{\varsigma}}) = 0$ (Table 1). By the fact that the canonical dual problem $(\mathcal{P}_+^d)$ is a concave maximization over a convex open space, the DG method was not trapped into any local optimal solution. But, for the nonconvex primal problem $(\mathcal{P})$ in dimensions n=5∼7 and 4000, the DG method was trapped into local minimizer $\bar{\mathbf{x}} = (-1, 1, \ldots, 1)$. If we set the initial solution as $\boldsymbol{\varsigma}_0 = (100, 100, \ldots, 100, 0)$ (called seed2) and repeat the calculations, our numerical results (Table 2) show again that the canonical dual problem can be easily and quickly solved by the DG method to accurately get the global maximizer $\bar{\boldsymbol{\varsigma}} = (0, 0, ..., 0)$ with the optimal solution $P^d(\bar{\boldsymbol{\varsigma}}) = 0$ for dimensions $n = 1 \sim 9, 19, 29, 39, 49, 59, 69, 79, 89, 99, 199, 299, 399, 499, 599, 699, 799, 899, 999, 1999$.

The comparisons between $(\mathcal{P})$ and $(\mathcal{P}_+^d)$ in view of total number of iterations and total number of objective function evaluations (i.e. function calls) are listed in Tables 1-2. Compared with $(\mathcal{P}_+^d)$, the approximate global and local optimal solutions and their optimal objective function values of $(\mathcal{P})$ are not accurate, and even cannot be obtained if the initial iteration is set to be $\mathbf{x}_0 = (100, 100, \ldots, 100)$. In Table 1, we can see that the total number of iterations and function calls for $(\mathcal{P})$ are always greater than those for $(\mathcal{P}_+^d)$. This means that $(\mathcal{P}_+^d)$ costs less computer calculations than $(\mathcal{P})$, though $(\mathcal{P}_+^d)$ still can get accurate global optimal solutions and the global optimal objective function value. The initial solutions $\mathbf{x}_0 = (100, 100, \ldots, 100)$ and $\boldsymbol{\varsigma}_0 = (100, 100, \ldots, 100, 0)$ respectively for $(\mathcal{P})$ and $(\mathcal{P}_+^d)$ are not practical for real numerical tests so that the total number of iterations and function calls of $(\mathcal{P})$ are sometimes less than those of $(\mathcal{P}_+^d)$. Regarding the CPU times for solving $(\mathcal{P}_+^d)$ with $n = 4000$, the largest CPU time for seed1 is 206.3581 seconds (i.e. 3.4393 minutes).

Example 1. Let $n = 4$ (four dimensions). The global minimizer is known to be

Table 1: Results of numerical experiments for $(\mathcal{P})$ and $(\mathcal{P}_+^d)$: $N = 100$, seed1

Dimension n	Iterations		Function calls		Objective function value	
	$(\mathcal{P})$	$(\mathcal{P}_+^d)$	$(\mathcal{P})$	$(\mathcal{P}_+^d)$	$(\mathcal{P})$	$(\mathcal{P}_+^d)$
2	120	24	2843	28	0.00001073	0.00000000
3	422	26	8996	137	0.00401438	0.00000000
4	3737	35	48352	202	0.00615273	0.00000000
5*	335	34	10179	399	3.96077434	0.00000000
6*	2375	44	43770	868	4.00635895	0.00000000
7*	1223	53	28009	1625	4.09419146	0.00000000
8	2160	55	46792	2100	0.01246714	0.00000000
9	2692	51	61017	2526	0.01397307	0.00000000
10	4444	63	91470	3979	0.01055630	0.00000000
20	3042	55	140924	10084	0.00940077	0.00000000
30	2321	58	133980	20515	0.01075478	0.00000000
40	1659	60	173795	26818	0.01227866	0.00000000
50	2032	57	219233	36459	0.01264147	0.00000000
60	1966	61	260701	50495	0.01048188	0.00000000
70	1876	56	272919	52545	0.01531147	0.00000000
80	1405	61	195156	59684	0.01594730	0.00000000
90	2142	61	371963	71320	0.01055831	0.00000000
100	2676	60	510722	70208	0.01125514	0.00000000
200	1395	61	653604	188589	0.01115318	0.00000000
300	1368	60	882760	235163	0.01574873	0.00000000
400	2085	66	1869675	301805	0.00928066	0.00000000
500	1155	59	1394240	358938	0.01168440	0.00000000
600	1226	63	1808285	451817	0.00918730	0.00000000
700	1557	60	2134359	559378	0.01257100	0.00000000
800	1398	61	2098062	522726	0.01442714	0.00000000
900	716	65	1904187	763449	0.01074534	0.00000000
1000	1825	61	3598608	681509	0.00897202	0.00000000
2000	257	62	2087277	1455472	0.00937219	0.00000000
3000	3221	60	20642543	2714296	0.01250373	0.00000000
4000*	679	60	7581502	3659292	4.11193171	0.00000000

Table 2: Results of numerical experiments for $(\mathcal{P})$ and $(\mathcal{P}_+^d)$: $N = 100$, seed2

Dimension n	Iterations		Function calls		Objective function value	
	$(\mathcal{P})$	$(\mathcal{P}_+^d)$	$(\mathcal{P})$	$(\mathcal{P}_+^d)$	$(\mathcal{P})$	$(\mathcal{P}_+^d)$
2	10013	24	227521	28	47.23824896	0.00000000
3	144	32	4869	235	96.49814330	0.00000000
4	144	81	5279	938	82.46230602	0.00000000
5	148	137	5682	1768	94.19254867	0.00000000
6	154	166*	6238	2590	88.84382963	0.00000000
7	159	179*	7097	3288	237.63078399	0.00000000
8	165	202*	7502	4300	238.41126013	0.00000000
9	153	206*	7137	5083	84.54205412	0.00000000
10	162	216*	7491	5920	83.23094398	0.00000000
20	225	285*	19111	17458	83.94779152	0.00000000
30	216	301*	20939	28543*	156.95838274	0.00000000
40	163	291*	19775	40444*	83.30960344	0.00000000
50	158	298*	33269	51888*	85.93091895	0.00000000
60	158	312*	34094	61767*	89.07412094	0.00000000
70	162	284*	35436	69865*	92.45725362	0.00000000
80	209	297*	35607	89127*	157.69955825	0.00000000
90	227	294*	60398	98748*	82.44035053	0.00000000
100	202	290*	57792	102796*	81.94595276	0.00000000
200	1826	262	436413	189293	83.77165551	0.00000000
300	195	259*	169238	261320*	152.95671738	0.00000000
400	195	278*	212104	375816*	82.49253919	0.00000000
500	190	297*	331637	522695*	82.40170647	0.00000000
600	292	303*	431092	559068*	150.15456693	0.00000000
700	189	275*	383735	758631*	89.14575473	0.00000000
800	198	270*	429674	701053*	84.50538257	0.00000000
900	198	280*	416150	867398*	85.32757049	0.00000000
1000	193	283*	445326	930761*	89.48369379	0.00000000
2000	232	310*	1123240	2030104*	84.26810981	0.00000000

$\bar{\mathbf{x}} = (1, 1, 1, 1)$ and $P(\bar{\mathbf{x}}) = 0$.

Solution: By using the DG method for both primal problem $(\mathcal{P})$ and its canonical

dual ($\mathcal{P}_+^d$), we have the numerical solutions

$$\bar{\mathbf{x}} = (1.0166873133, 1.0337174892, 1.0687306765, 1.1425101552), \quad P(\bar{\mathbf{x}}) = 0.00615273,$$

$$\bar{\boldsymbol{\varsigma}} = (0.0000000119, 0.0000000000, 0.0000000000), \quad P_+^d(\bar{\boldsymbol{\varsigma}}) = 0.00000000.$$

This shows that the canonical dual problem provides more accurate solution.

Example 2. For dimension $n = 5$, the Rosenbrock function has exactly two minima, one is the global optimal solution $(1, 1, 1, 1, 1)$ with global optimal minimum value 0, and another minimum is a local minimum near $(-1, 1, 1, 1, 1)$ with local optimal minimum value 4.

Solution: By the DG method, the primal solution is

$$\bar{\mathbf{x}} = (-0.9856129203, 0.9814803343, 0.9682775584, 0.9398661046, 0.8840549028)$$

with $P(\bar{\mathbf{x}}) = 3.96077434$. Clearly, this is a local minimizer. While the canonical dual problem produces accurately a global optimal solution

$$\bar{\boldsymbol{\varsigma}} = (0.0000004388, 0.0000006036, 0.0000000000, 0.0000000000), \quad P_+^d(\bar{\boldsymbol{\varsigma}}) = 0.$$

Example 3. For $n = 6$ (six dimensions), the Rosenbrock function has exactly two minima, i.e., the global optimal solution

$$\bar{\mathbf{x}}_1 = (1, 1, 1, 1, 1, 1), \quad P(\bar{\mathbf{x}}_1) = 0,$$

and local minimal solution

$$\bar{\mathbf{x}}_2 = (-1, 1, 1, 1, 1, 1), \quad P(\bar{\mathbf{x}}_2) = 4.$$

Solution: To solve the primal problem directly, the DG method can only provide local solution

$$\bar{\mathbf{x}} = (-0.9970726441, 1.0041582933, 1.0133158817, 1.0292928527, 1.0607123926, 1.1258344785)$$

with $P(\bar{\mathbf{x}}) = 4.00635895$. For the canonical dual problem, the DG method produces

$$\bar{\boldsymbol{\varsigma}} = (0.0000001747, -0.0000000559, 0.0000005919, 0.0000000000, 0.0000000000),$$

$$P_+^d(\bar{\boldsymbol{\varsigma}}) = 0.$$

Example 4. Similarly, if $n = 7$, the test problem has the same global optimal solution

$$\bar{\mathbf{x}}_1 = (1, 1, 1, 1, 1, 1, 1), \;\; P(\bar{\mathbf{x}}_1) = 0$$

and the local minimal solution

$$\bar{\mathbf{x}}_2 = (-1, 1, 1, 1, 1, 1, 1), \;\; P(\bar{\mathbf{x}}_2) = 4.$$

Solution: By the DG method, we have

$$\begin{aligned} \bar{\mathbf{x}} &= (-1.0003403494, 1.0106728675, 1.0264433859, 1.0561180077, \\ &\quad 1.1168007274, 1.2483026410, 1.5594822181), \\ P(\bar{\mathbf{x}}) &= 4.09419146, \end{aligned}$$

$$\begin{aligned} \bar{\boldsymbol{\varsigma}} &= (-0.0000001431, -0.0000011147, -0.0000010643, -0.0000003284, \\ &\quad 0.0000000000, 0.0000000000), \\ P_+^d(\bar{\boldsymbol{\varsigma}}) &= 0. \end{aligned}$$

This shows again that the DG iterations for solving the primal problem is trapped to a local min.

Example 5. Now we let $n = 4000$. The Rosenbrock function has many minima. The global optimal solution is still $\bar{\mathbf{x}}_1 = (1, \ldots, 1)$ with $P(\bar{\mathbf{x}}) = 0$. One of local minima is nearby the point $\bar{\mathbf{x}}_2 = (-1, 1, \ldots, 1)$ with $P(\bar{\mathbf{x}}_2) = 4$.

Solution: Again, by the DG method, the primal iteration is trapped at

$$\bar{\mathbf{x}} = (-0.9932861006, 0.9966510741, \ldots, 1.3122885708, 1.7233744896), \;\; P(\bar{\mathbf{x}}) = 4.11193171.$$

The conical dual solution is

$$\begin{aligned} \bar{\boldsymbol{\varsigma}} &= (-0.0000000314, -0.0000000040, -0.0000000437, \ldots, \\ &\quad -0.0000000281, 0.0000000008, -0.0000000214, 0.0000000000, 0.0000000000), \end{aligned}$$

which produce precisely the optimal value $P_+^d(\bar{\boldsymbol{\varsigma}}) = 0$. Indeed, as long as $n \geq 5$, the DG method is always trapped into the local minimizer $\bar{\mathbf{x}} = (-1, 1, \ldots, 1)$ if the initial solution is set to be $\mathbf{x}_0 = (-1.0005, 1.0005, \ldots, 1.0005)$.

It is worth to note that both $P(\mathbf{x})$ and $P^d(\boldsymbol{\varsigma})$ are the sum of $n-1$ items. This is convenient for MPI (Message Passing Interface) parallel computations. We may broadcast (MPI_Bcast) the sum to $n-1$ processes, each process calculates one item, and at last all the partials are reduced (MPI_Reduce) onto one process to get the sum. Thus on Tambo machines of VLSCI (http://www.vlsci.unimelb.edu.au) we should be able to successfully solve (1) and (12) with at least 3.2767×10^7 variables if setting the maximal variables for the DG method to be 4000 (though the DG method and its parallelization version ([3]) can solve optimization problems with more than 4000 variables). The successfully tested MPI code is followed:

broadcast $n-1$

call MPI_BCAST ($n-1$,1,MPI_INTEGER, 0, MPI_COMM_WORLD , ierr)

check for quit signal

if ($n-1$ .le. 0) goto 30

calculate every partials

sum = 0.0d0

do 20 i = myid+1, $n-1$, numprocs

sum = sum $+(x(i)-1.0)**2+100.0*(x(i)**2-x(i+1))**2$

20 continue (for $P(\mathbf{x})$)

do 20 i = myid+1, $n-1$, numprocs

if ($i-1$ .eq. 0) then $\varsigma(0)$=0

sum = sum $+(\varsigma(i-1)+2.0)/(4*(\varsigma(i)+1.0))+(1.0/400.0)*\varsigma(i)**2$

20 continue (for $P^d(\boldsymbol{\varsigma})$)

f = sum

collect all the partial sums

call MPI_REDUCE (f,objf,1,MPI_DOUBLE_PRECISION, MPI_SUM, 0,
& MPI_COMM_WORLD, ierr)

30 node 0 (i.e. myid = 0) prints the sums = objf

4 Conclusion

This research note demonstrates a powerful application of the *canonical duality theory* for solving the nonconvex minimization problem of Rosenbrock function. Extensive numerical computations show that by using the same DG method, the canonical dual problem can be easily solved to produce global solutions.

Acknowledgments:

This research is supported by US Air Force Office of Scientific Research under the grant AFOSR FA9550-10-1-0487. The authors thank Professor Kok Lay Teo (Curtin University, Australia) for his helpful comments and Associate Professor Adil M. Bagirov (Ballarat University, Australia) for his FORTRAN codes of his discrete gradient method.

References

[1] Ai, Y.S., Liu, P.C., Zheng, T.Y.: Adaptive hybrid global inversion algorithm. Science in China (Series D). **41 (2)**, 137–143 (1998)

[2] Bagirov, A.M.: Continuous subdifferential approximations and their applications. J. Math. Sci. **15**, 2567–2609 (2003)

[3] Beliakov, G., Monsalve Tobon, J.E., Bagirov, A.M.: Parallelization of the discrete gradient method of non-smooth optimization and its applications. Computational Science – ICCS 2003 Lecture Notes in Computer Science. **2659/2003**, 701–711 (2003)

[4] Bouvry, P., Arbab, F., Seredynski, F.: Distributed evolutionary optimization, in Manifold: Rosenbrock's function case study. Inf. Sci. **122**, 141–159 (2000)

[5] Gao, D.Y.: Dual extremum principles in finite deformation theory with applications to post-buckling analysis of extended nonlinear beam theory. Applied Mechanics Reviews Vol. **50 (11)**, S64–S71 (1997)

[6] Gao, D.Y.: General analytic solutions and complementary variational principles for large deformation nonsmooth mechanics. Meccanica **34**, 169–198 (1999)

[7] Gao, D.Y.: Analytic solution and triality theory for nonconvex and nonsmooth variational problems with applications. Nonlinear Analysis **42 (7)**, 1161–1193 (2000)

[8] Gao, D.Y.: Canonical dual transformation method and generalized triality theory in nonsmooth global optimization. J. Glob. Opt. **17**, 127-160 (2000)

[9] Gao, D.Y.: Duality Principles in Nonconvex Systems: Theory, Methods and Applications. Kluwer Academic Publishers, Dordrecht/Boston/London (2000)

[10] Gao, D.Y.: Perfect duality theory and complete set of solutions to a class of global optimization. Opt. **52** (4-5), 467–493 (2003)

[11] Gao, D.Y.: Solutions and optimality criteria to box constrained nonconvex minimization problems. J. Indust. Manag. Opt. **3** (2), 293–304 (2007)

[12] Gao, D.Y., Ruan, N., Pardalos P.: Canonical dual solutions to sum of fourth-order polynomials minimization problems with applications to sensor network localization. (2010)

[13] Gao, D.Y. and Strang, G.: Geometric nonlinearity: Potential energy, complementary energy, and the gap function. Quart. Appl. Math. XLVII(3), 487–504 (1989)

[14] Gao, D.Y. and Wu, C.Z.: On the triality theory in global optimization. J. Global Optimization (2010)

[15] Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. The Computer Journal **3**: 175–184 (1960)

[16] Xin, B., Chen, J., Pan, F.: Problem difficulty analysis for particle swarm optimization: deception and modality. 2009 Proceedings of the 1st ACM/SIGEVO Summit on Genetic and Evolutionary Computation 623–630 (2009)

[17] Yu, H.F., Wang, D.W.: Research on food-chain algorithm and its parameters. Front. Electr. Electron. Eng. China **3(4)**, 394-398 (2008)

Five-Dimensional Tangent Vectors in Space-Time
II. Differential-Geometric Approach

Alexander Krasulin
Institute for Nuclear Research of the Russian Academy of Sciences *
krasulin@post.com

Abstract

In this part of the series five-dimensional tangent vectors are introduced first as equivalence classes of parametrized curves and then as differential-algebraic operators that act on scalar functions. I then examine their basic algebraic properties and their parallel transport in the particular case where space-time possesses a special local symmetry. After that I give definition to five-dimensional tangent vectors associated with dimensional curve parameters and show that they can be identified with the five-vectors introduced formally in part I. In conclusion I speak about differential forms associated with five-vectors.

1. Five-vectors as equivalence classes of parametrized curves

A. *Definition*

Consider a set $\Re$ of all smooth parametrized curves going through a fixed space-time point Q. I will lable these curves with calligraphic capital Roman letters: $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, etc. The parameter of curve $\mathcal{A}$ will be denoted as $\lambda_{\mathcal{A}}$.

If f is a real scalar function defined in the vicinity of Q, one can evaluate its derivative at Q along a given curve $\mathcal{A}$:

$$\left.\frac{df(P(\lambda_{\mathcal{A}}))}{d\lambda_{\mathcal{A}}}\right|_{\lambda_{\mathcal{A}}=\lambda_{\mathcal{A}}(Q)},$$

and I will denote this derivative as $\partial_{\mathcal{A}} f|_Q$.

Let us focus our attention on the behaviour of curves in the infinitesimal vicinity of Q. From that point of view, $\Re$ can be divided into classes of equivalent curves that coincide in direction or in direction and parametrization. One can consider three degrees to which two given curves, $\mathcal{A}$ and $\mathcal{B}$, may coincide:

1. The two curves come out of Q in the same direction. A more precise formulation is the following: there exists a real positive number a such that for any scalar function f

$$\partial_{\mathcal{A}} f|_Q = a \cdot \partial_{\mathcal{B}} f|_Q. \tag{1}$$

2. The two curves come out of Q in the same direction and in the vicinity of Q their parameters change with equal rates. More precisely: for any scalar function f

$$\partial_{\mathcal{A}} f|_Q = \partial_{\mathcal{B}} f|_Q. \tag{2}$$

3. The two curves come out of Q in the same direction; their parameters change with equal rates in the vicinity of Q; and the values of these parameters at Q are the same. This means that

$$\lambda_{\mathcal{A}}(Q) = \lambda_{\mathcal{B}}(Q) \tag{3a}$$

and for any scalar function f

$$\partial_{\mathcal{A}} f|_Q = \partial_{\mathcal{B}} f|_Q. \tag{3b}$$

It is a simple matter to check that relations (1), (2) and (3) are all equivalence relations on $\Re$, and for each of them one can consider the corresponding quotient set—the set whose elements are classes of equivalent curves.

Relation (1) is of no interest to us and I will not consider it any further.

The elements of the quotient set corresponding to relation (2) will be denoted with capital boldface Roman letters: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, etc. According to relation (2), the derivative of any scalar function f at Q is the same for all curves belonging to a given class $\mathbf{A}$, so it makes sense to introduce the notation $\partial_{\mathbf{A}} f|_Q$.

In a natural way, one can define the addition of two equivalence classes $\mathbf{A}$ and $\mathbf{B}$: $\mathbf{A}+\mathbf{B}$ is such an equivalence class that for any scalar function f one has

$$\partial_{\mathbf{A}+\mathbf{B}} f|_Q = \partial_{\mathbf{A}} f|_Q + \partial_{\mathbf{B}} f|_Q.$$

*Former affiliation.

It is easy to prove that such a sum exists for any pair of equivalence classes.

In a similar manner one can give definition to the product of an equivalence class $\mathbf{A}$ and a real number k: $k\mathbf{A}$ is such an equivalnce class that

$$\partial_{k\mathbf{A}} f|_Q = k \cdot \partial_{\mathbf{A}} f|_Q$$

for any scalar function f. Again, one can verify that $k\mathbf{A}$ exists for any $\mathbf{A}$ and any k.

With thus defined addition and multiplication by a real number, the set of all equivalence classes corresponding to relation (2) becomes a vector space. This space is four-dimensional, and I will denote it as V_4. As it will be discussed in section 2, the elements of V_4 can be identified with four-dimensional tangent vectors, so in the following I will refer to them as to *four-vectors*.

Let us now turn to the quotient set associated with relation (3). Its elements will be denoted with lowercase boldface Roman letters: $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, etc. As in the case of four-vectors, one can introduce the notation $\partial_{\mathbf{a}} f|_Q$ for the common value of the derivatives of any scalar function f along all the curves belonging to a given equivalence class $\mathbf{a}$. Similarly, the common value of the parameters of all these curves at Q will be denoted as $\lambda_{\mathbf{a}}(Q)$.

One can now give the following definitions to the sum of two equivalence classes $\mathbf{a}$ and $\mathbf{b}$ and to the product of an equivalence class $\mathbf{a}$ and a real number k: $\mathbf{a}+\mathbf{b}$ and $k\mathbf{a}$ are such equivalence classes that

$$\lambda_{\mathbf{a}+\mathbf{b}}(Q) = \lambda_{\mathbf{a}}(Q) + \lambda_{\mathbf{b}}(Q),$$
$$\lambda_{k\mathbf{a}}(Q) = k \cdot \lambda_{\mathbf{a}}(Q)$$

and for any scalar function f

$$\partial_{\mathbf{a}+\mathbf{b}} f|_Q = \partial_{\mathbf{a}} f|_Q + \partial_{\mathbf{b}} f|_Q,$$
$$\partial_{k\mathbf{a}} f|_Q = k \cdot \partial_{\mathbf{a}} f|_Q.$$

One can easily check that such a sum and such a product exist respectively for any two equivalence classes and for any equivalence class and any real number. These two operations turn the quotient set associated with relation (3) into a vector space, whose dimension is evidently five. Let us denote this space as V_5 and call its elements *five-dimensional tangent vectors* or simply *five-vectors*. In section 2 I will consider another, equivalent representation for these vectors and later on will show that they have all the formal properties of those five-vectors that have been introduced in part I.

B. *Structure of the five-vector space*

As any other vector space, V_5 is completely isotropic with respect to its two composition laws and has no distinguished direction nor any other distinguished subspace of nonzero dimension. However, one *can* distinguish two subspaces in V_5 by associating them with certain classes of parametrized curves.

Let us consider all those curves from $\Re$ for which $\partial f|_Q = 0$ for any scalar function f. It is evident that all these curves belong to the same equivalence class with respect to relation (2) and that this class is the zero vector in V_4. With respect to relation (3), the considered curves belong to equivalence classes that make up a one-dimensional subspace in V_5, which I will denote as $\mathcal{E}$. One can say that $\mathcal{E}$ is made up by all those five-vectors that do not correspond to any direction in the manifold.

Another distinguished subspace in V_5 can be obtained by considering all those curves from $\Re$ for which $\lambda(Q) = 0$. The four-vectors corresponding to these curves are all the vectors of V_4. The corresponding five-vectors make up a four-dimensional subspace in V_5, which I will denote as $\mathcal{Z}$. It is easy to see that $\mathcal{E}$ and $\mathcal{Z}$ have only one common element—the zero vector, and that V_5 is the direct sum of $\mathcal{E}$ and $\mathcal{Z}$. The components of an arbitrary five-vector $\mathbf{u}$ in these two subspaces will be denoted as $\mathbf{u}^{\mathcal{E}}$ and $\mathbf{u}^{\mathcal{Z}}$, respectively.

Other properties of $\mathcal{E}$ and $\mathcal{Z}$ will be discussed below.

C. *Relation between four- and five-vectors*

As it follows from the definition of four- and five-vectors given above, there exists a set-theoretic relation between V_4 and V_5: the former is the quotient set corresponding to the following equivalence relation on V_5:

$$\mathbf{a} \equiv \mathbf{b} \Leftrightarrow \partial_{\mathbf{a}} f|_Q = \partial_{\mathbf{b}} f|_Q \text{ for any scalar function } f.$$

Denoting this relation as R, one has $V_4 = V_5/R$. The fact that $\mathbf{A}$ is the equivalence class of $\mathbf{a}$ will be denoted as $\mathbf{a} \in \mathbf{A}$. From the definition of symbols $\partial_{\mathbf{a}}$ and $\partial_{\mathbf{A}}$ it follows that $\mathbf{a} \in \mathbf{A}$ if and only if $\partial_{\mathbf{a}} = \partial_{\mathbf{A}}$. It is a simple matter to see that R has the following linearity properties: if $\mathbf{a} \equiv \mathbf{b} \,(\mathrm{mod} R)$ and $\mathbf{c} \equiv \mathbf{d} \,(\mathrm{mod} R)$, then $\mathbf{a}+\mathbf{c} \equiv \mathbf{b}+\mathbf{d} \,(\mathrm{mod} R)$ and $k\mathbf{a} \equiv k\mathbf{b} \,(\mathrm{mod} R)$, where k is an arbitrary real number. Thus, as any other equivalence relation with such properties, R can be presented in the following form:

$$\mathbf{a} \equiv \mathbf{b} \,(\mathrm{mod} R) \Leftrightarrow \mathbf{a} - \mathbf{b} \in W,$$

where W is the subspace in V_5 that contains all the five-vectors equivalent to the zero vector. It is easy to see that W coincides with the one-dimensional subspace $\mathcal{E}$ introduced in the previous subsection, so R

can be reformulated as:

$$\mathbf{a} \equiv \mathbf{b} \,(\mathrm{mod} R) \Leftrightarrow \mathbf{a} = \mathbf{b} + \mathbf{e}, \text{ where } \mathbf{e} \in \mathcal{E}.$$

The latter condition is equivalent to **a** and **b** having equal components in the four-dimensional subspace $\mathcal{Z}$ or, for that matter, in any subspace complementary to $\mathcal{E}$. This means that there exists a one-to-one correspondence between the five-vectors from $\mathcal{Z}$ and four-vectors, and this correspondence is evidently a homomorphism.

Let me say a few words about the selection of bases in V_4 and V_5 and their transformation.

A typical five-vector basis will be denoted as $\mathbf{e}_A$, where A (as all capital latin indices) runs 0, 1, 2, 3, and 5. One can choose a basis in V_5 arbitrarily, but it is more convenient to select the fifth basis vector belonging to $\mathcal{E}$. Such bases will be called *standard* and will be used in all calculations.

The basis in V_4 can be chosen arbitrarily and independently of the basis in V_5. It is more convenient though to associate it with the five-vector basis. A natural choice is to take $\mathbf{E}_\alpha$ to be the equivalence classes of the basis five-vectors $\mathbf{e}_\alpha$ (the equivalence class of $\mathbf{e}_5$ is the zero four-vector). I will refer to this basis as to the one *associated* with the basis $\mathbf{e}_A$ in V_5.

If $\mathbf{e}_A$ and $\mathbf{e}'_A$ are two standard bases in V_5 and $\mathbf{e}'_A = \mathbf{e}_B L^B_A$, then L^B_A can be shown to satisfy the condition

$$L^\alpha_5 = 0 \text{ for all } \alpha.$$

The corresponding equivalence classes are related as $\mathbf{E}'_\alpha = \mathbf{E}_\beta L^\beta_\alpha$.

D. *Reminder on the inner product of four-vectors*

Four-vectors inherit their inner product from the Riemannian metric of space-time. The latter is a rule that assigns a certain number, called interval, to each finite continuous line. This number is additive, and for an infinitesimal line connecting two points with coordinates x^α and $x^\alpha + dx^\alpha$ it equals

$$\sqrt{g_{\alpha\beta}(x) dx^\alpha dx^\beta} + \text{ terms of higher order in } dx, \quad (4)$$

where $g_{\alpha\beta}$ is a real nondegenerate 4×4 matrix with the signature $(+,-,-,-)$.

Consider now a parametrized curve coming out of a point Q. According to formula (4), the interval assigned to the part of the curve between Q and a nearby point corresponding to the parameter value $\lambda(Q) + d\lambda$ is

$$\sqrt{g_{\alpha\beta}(Q)(\partial x^\alpha/\partial\lambda)_Q (\partial x^\beta/\partial\lambda)_Q} \cdot d\lambda + \text{ terms of higher order in } d\lambda.$$

Since $(\partial x^\alpha/\partial\lambda)_Q$ is the same for all curves from a given equivalence class associated with relation (2), the expression under the radical sign is a function of the four-vector corresponding to the curve rather than of the curve itself. This enables one to assign a number to each four-vector, which is interpreted as its length squared. More precisely, the inner product g is defined as a real bilinear symmetric function of two four-vectors such that for any four-vector **U**

$$g(\mathbf{U}, \mathbf{U}) = g_{\alpha\beta}(Q)(\partial_{\mathbf{U}} x^\alpha)_Q (\partial_{\mathbf{U}} x^\beta)_Q.$$

The interval is a dimensional quantity. It is measured in centimeters or seconds or in any other units of length or time. Accordingly, the quantity under the radical sign in formula (4) is measured in cm^2 or sec^2 or in some other squared units. Throughout sections 1 and 2 of this paper I will consider only dimensionless coordinates and curve parameters. Then, if the interval is measured, say, in centimeters, the elements of the matrix $g_{\alpha\beta}$ will be measured in cm^2, $g^{\alpha\beta}$ will be measured in cm^{-2}, and the connection coefficients for four- and five-vector fields will be dimensionless.

E. *Symmetries*

The set $\Re$ of all parametrized curves going through an arbitrary point Q has a certain symmetry with respect to the behaviour of curves in the infinitesimal vicinity of Q. Namely, there exist certain maps of $\Re$ onto itself that have the following properties:

1. If $\mathcal{A} \mapsto \mathcal{A}'$, then $\lambda_{\mathcal{A}}(Q) = \lambda_{\mathcal{A}'}(Q)$.
2. If $\mathcal{A} \mapsto \mathcal{A}'$ and $\mathcal{B} \mapsto \mathcal{B}'$, and for any scalar function f one has $\partial_{\mathcal{A}} f|_Q = k \cdot \partial_{\mathcal{B}} f|_Q$, where k is some constant factor, then for any scalar function f one has $\partial_{\mathcal{A}'} f|_Q = k \cdot \partial_{\mathcal{B}'} f|_Q$.
3. If $\mathcal{A} \mapsto \mathcal{A}'$, $\mathcal{B} \mapsto \mathcal{B}'$, and $\mathcal{C} \mapsto \mathcal{C}'$, and for any scalar function f one has $\partial_{\mathcal{A}} f|_Q + \partial_{\mathcal{B}} f|_Q = \partial_{\mathcal{C}} f|_Q$, then for any scalar function f one has $\partial_{\mathcal{A}'} f|_Q + \partial_{\mathcal{B}'} f|_Q = \partial_{\mathcal{C}'} f|_Q$.
4. If $\mathcal{A} \mapsto \mathcal{A}'$, then

$$g_{\alpha\beta}(Q)(\partial_{\mathcal{A}} x^\alpha)_Q (\partial_{\mathcal{A}} x^\beta)_Q = g_{\alpha\beta}(Q)(\partial_{\mathcal{A}'} x^\alpha)_Q (\partial_{\mathcal{A}'} x^\beta)_Q.$$

Property 2 at $k = 1$ means that such transformations of $\Re$ induce maps of V_4 onto itself. Properties 2 and 3 mean that these transformations of V_4 are linear, and property 4 means that they conserve the inner product of two four-vectors.

Property 1 and property 2 at $k = 1$ mean that the considered transformations of $\Re$ also induce maps

of V_5 onto itself. Properties 1, 2, and 3 mean that these maps are linear. Property 1 means that a vector from $\mathcal{Z}$ is transformed into a vector from $\mathcal{Z}$. And properties 1 and 2 mean that vectors from $\mathcal{E}$ are not changed at all.

Let us now find the corresponding transformation matrices for V_4 and V_5.

Let $\mathbf{E}_\alpha$ be an arbitrary orthonormal basis in V_4 and let us take that under the considered transformation these basis vectors are transformed into $\mathbf{E}'_\alpha = \mathbf{E}_\beta \Lambda^\beta_{\ \alpha}$. Since the transformation should conserve the inner product, and the basis $\mathbf{E}_\alpha$ is orthonormal, $\Lambda^\beta_{\ \alpha}$ should be a matrix from O(3,1). As a basis in V_5 let us take a standard basis where $\mathbf{e}_\alpha \in \mathcal{Z}$ and $\mathbf{e}_\alpha \in \mathbf{E}_\alpha$. Let us suppose that $\mathbf{e}_A$ are transformed into $\mathbf{e}'_A = \mathbf{e}_B L^B_{\ A}$, where $L^B_{\ A}$ is some real nondegenerate 5×5 matrix. Since vectors from $\mathcal{E}$ do not change under the considered transformation, one should have $L^5_{\ 5} = 1$ and $L^\alpha_{\ 5} = 0$ for all α. Since vectors from $\mathcal{Z}$ are transformed into vectors from $\mathcal{Z}$, one should have $L^5_{\ \alpha} = 0$ for all α. Finally, owing to the one-to-one correspondence between $\mathcal{Z}$ and V_4, one should have $L^\alpha_{\ \beta} = \Lambda^\alpha_{\ \beta} \in$ O(3,1).

F. *Inner product of five-vectors*

The method used in subsection D to define the inner product g for four-vectors is also applicable in the case of five-vectors. The resulting inner product on V_5, which for the time being I will denote as h', is a real bilinear symmetric function of two five-vectors such that for any five-vector $\mathbf{u}$

$$h'(\mathbf{u},\mathbf{u}) = g_{\alpha\beta}(Q)(\partial_{\mathbf{u}} x^\alpha)_Q (\partial_{\mathbf{u}} x^\beta)_Q$$

(Q is the space-time point where one considers the tangent space of five-vectors). Since the value of the derivative $\partial_{\mathbf{u}}$ is the same for all five-vectors corresponding to the same four-vector, h' will be a degenerate inner product. It is not difficult to see that the subspace of all degenerate five-vectors for h' (of all such five-vectors $\mathbf{u}$ that $h'(\mathbf{u},\mathbf{v}) = 0$ for any $\mathbf{v}$) coincides with $\mathcal{E}$ and that h' is nondegenerate within any subspace complementary to $\mathcal{E}$. It is also apparent that for any $\mathbf{u}$ and $\mathbf{v}$ one has

$$h'(\mathbf{u},\mathbf{v}) = g(\mathbf{U},\mathbf{V}), \tag{5}$$

where $\mathbf{u} \in \mathbf{U}$ and $\mathbf{v} \in \mathbf{V}$.

It is not difficult to construct from h' a nondegenerate inner product on V_5. For that one should consider another natural measure that exists for five-vectors: to each five-vector $\mathbf{u}$ one can put into correspondence the value of the relevant curve parameter, $\lambda_{\mathbf{u}}$. If one then interprets this latter number as the length of vector $\mathbf{u}$, one will obtain another inner product—let us denote it as h''—which will also be degenerate. It is easy to see that $h''(\mathbf{u},\mathbf{v}) = \lambda_{\mathbf{u}} \cdot \lambda_{\mathbf{v}}$. Consequently, the subspace of all degenerate vectors for h'' coincides with $\mathcal{Z}$ and h'' is nondegenerate within any (one-dimensional) subspace complementary to $\mathcal{Z}$.

One should now notice that the subspaces of degenerate vectors for h' and h'' are complementary to each other, which means that the sum of h' and h'' will be a nondegenerate inner product on V_5. The only problem in constructing such a sum is that h' is a dimensional quantity and is measured in the same units as g, whereas h'', being the product of curve parameters, does not have a dimension. We thus see that to construct a nondegenerate inner product on V_5 from h' and h'', one needs a dimensional constant, ξ, which would play a role similar to that of the speed of light: it would establish a relation between different units used to measure the same quantity. The resulting inner product measured in the same units as g will be

$$h(\mathbf{u},\mathbf{v}) = h'(\mathbf{u},\mathbf{v}) + \xi \cdot h''(\mathbf{u},\mathbf{v}). \tag{6}$$

The same result can be obtained from considerations of another kind. For that one should adopt the view-point that four-vectors and five-vectors are subordinate objects, whose algebraic properties are determined by the properties of the manifold with which they are associated. In particular, this means that the structure of V_5 should have a symmetry no less than the symmetry of $\Re$. This, in its turn, means that *any* inner product of five-vectors should be invariant under the transformations discussed in the previous subsection.

Let us consider the same five-vector basis $\mathbf{e}_A$ that has been used in subsection E. It is a simple matter to show that the matrix $h_{AB} \equiv h(\mathbf{e}_A, \mathbf{e}_B)$ of any nondegenerate inner product h satisfying the above symmetry requirement has to be of the form

$$h_{\alpha\beta} = a \cdot \eta_{\alpha\beta},\ h_{\alpha 5} = h_{5\alpha} = 0,\ h_{55} = b, \tag{7}$$

where a and b are some nonzero constants. A direct consequence of these formulae is that any five-vector from $\mathcal{Z}$ is orthogonal to any five-vector from $\mathcal{E}$, so for any $\mathbf{u}$ and $\mathbf{v}$

$$h(\mathbf{u},\mathbf{v}) = h(\mathbf{u}^{\mathcal{Z}},\mathbf{v}^{\mathcal{Z}}) + h(\mathbf{u}^{\mathcal{E}},\mathbf{v}^{\mathcal{E}}). \tag{8}$$

Another consequence of formulae (7) is that the inner product of any two five-vectors from $\mathcal{Z}$ is proportional to the inner product of the corresponding four-vectors. Thus, if the overall normalization of h is selected in such a way that the proportionality factor

between h and g be unity, one will have

$$h(\mathbf{u}^{\mathcal{Z}}, \mathbf{v}^{\mathcal{Z}}) = g(\mathbf{U}, \mathbf{V}) = h'(\mathbf{u}, \mathbf{v}).$$

Finally, one should observe that the $\mathcal{E}$-component of any five-vector $\mathbf{u}$ equals $\lambda_{\mathbf{u}} \cdot \mathbf{i}$, where $\mathbf{i}$ is the vector from $\mathcal{E}$ that corresponds to the unity value of the parameter: $\lambda_{\mathbf{i}} = 1$. Consequently,

$$h(\mathbf{u}^{\mathcal{E}}, \mathbf{v}^{\mathcal{E}}) = \lambda_{\mathbf{u}}\lambda_{\mathbf{v}}\, h(\mathbf{i}, \mathbf{i}) = h(\mathbf{i}, \mathbf{i})\, h''(\mathbf{u}, \mathbf{v}),$$

and formula (8) acquires the form of formula (6) with $\xi = h(\mathbf{i}, \mathbf{i})$. Thus, at an appropriate choice of its overall normalization factor, any nondegenerate inner product on V_5 satisfying the above, quite natural symmetry requirement has the form indicated in formula (6).

It is obvious that constant ξ is not determined by the Riemannian metric of space-time nor by symmetry considerations, and consequently the same is true of the nondegenerate inner product of five-vectors. This is a distinctive feature of five-dimensional tangent vectors (and of similar objects in other manifolds) and is a consequence of that specific way in which five-vectors are associated with space-time.

In the previous subsection I have introduced a five-vector basis where $\mathbf{e}_\alpha \in \mathcal{Z}$. As we have seen above, in terms of the five-vector inner product this means that all $\mathbf{e}_\alpha$ are orthogonal to $\mathbf{e}_5$. This is one of the two conditions satisfied by a *regular* five-vector basis defined in section 3 of part I within the formal theory, the other condition being that $h(\mathbf{e}_5, \mathbf{e}_5) = 1$. When five-vectors are introduced as equivalence classes of parametrized curves, it is more convenient to define the regular basis in a slightly different way, equating to unity not the value of $h(\mathbf{e}_5, \mathbf{e}_5)$ (which depends on the choice of ξ) but the value of $\lambda_{\mathbf{e}_5}$. A regular basis will thus be a standard five-vector basis where all $\mathbf{e}_\alpha \in \mathcal{Z}$ and $\mathbf{e}_5 = \mathbf{i}$.

2. Five-vectors as operators

A. *Another representation for five-vectors*

In modern textbooks on differential geometry, ordinary tangent vectors are usually introduced by identifying their fields with linear differential operators (derivations) that act upon scalar functions from a certain set $\Im$ which determines the topological and differential properties of the manifold. Each derivation is a map

$$\mathbf{U} : \Im \to \Im$$

that satisfies the following requirements:

$$\begin{array}{l} \mathbf{U}[k] = 0 \text{ for any constant function } k \in \Im, \\ \mathbf{U}[f+g] = \mathbf{U}[f] + \mathbf{U}[g] \text{ for any } f, g \in \Im, \\ \mathbf{U}[fg] = \mathbf{U}[f] \cdot g + f \cdot \mathbf{U}[g] \text{ for any } f, g \in \Im. \end{array} \tag{9}$$

One can then prove a theorem that in a local coordinate system each derivation can be presented as the following differential operator:

$$\mathbf{U} = U^\alpha(\partial/\partial x^\alpha), \tag{10}$$

where $\partial/\partial x^\alpha$ are derivatives along coordinate lines and U^α are scalar functions from $\Im$. It is evident that at each point in space-time there exists a natural isomorphism between the equivalence classes of parametrized curves corresponding to relation (2) and operators of the form (10):

$$\mathbf{A} \mapsto \partial_{\mathbf{A}},$$

and basing on this isomorphism one can identify the elements of V_4 with four-dimensional tangent vectors.

Let us now find a similar operator representation for five-vectors. First, one should notice that the two conditions that determine the equivalence relation (3) can be replaced with a single requirement: that for any scalar function f

$$\partial_{\mathcal{A}} f|_Q + \lambda_{\mathcal{A}}(Q) f(Q) = \partial_{\mathcal{B}} f|_Q + \lambda_{\mathcal{B}}(Q) f(Q).$$

This enables one to establish a one-to-one correspondence between the equivalence classes of parametrized curves associated with relation (3) and differential-algebraic operators of the form

$$\mathbf{u} = u^\alpha(\partial/\partial x^\alpha) + u^5 \cdot \mathbf{1}, \tag{11}$$

where $\mathbf{1}$ is the identity operator. The simplest variant of such a correspondence is evidently

$$\mathbf{a} \mapsto \partial_{\mathbf{a}} + \lambda_{\mathbf{a}} \cdot \mathbf{1}. \tag{12}$$

One can then consider five-vector fields and basing on the above correspondence, relate them to such maps $\mathbf{u} : \Im \to \Im$ which in any local coordinate system can be presented in the form (11), where u^A are now scalar functions.

Finally, one can find a set of formal requirements, similar to conditions (9) for derivations, that enable one to introduce the above maps without referring to any coordinates. One possible set of such requirements is the following:

$$\begin{array}{l} \mathbf{u}[k] = \upsilon \cdot k \text{ for any constant } k \in \Im, \\ \quad \text{where } \upsilon \in \Im \text{ is characteristic of } \mathbf{u}, \\ \mathbf{u}[f+g] = \mathbf{u}[f] + \mathbf{u}[g] \text{ for any } f, g \in \Im, \\ \mathbf{u}[fg] = \mathbf{u}[f] \cdot g + f \cdot \mathbf{u}[g] - \mathbf{u}[\mathit{1}]fg \text{ for} \\ \quad \text{any } f, g \in \Im, \text{ where } \mathit{1} \text{ is the constant} \\ \quad \text{unity function.} \end{array} \tag{13}$$

It is evident that any operator of the form (11) satisfies these three requirements. Let us now prove the reverse statement:

In any local coordinate system each map $\mathbf{u}: \Im \to \Im$ satisfying requirements (13) can be presented in the form (11), where u^A are scalar functions from $\Im$.

Proof: Let us consider the operator

$$\mathbf{w} \equiv \mathbf{u} - v \cdot \mathbf{1},$$

where v is the scalar function from $\Im$ defined by the first of the requirements (13). It is a simple matter to check that $\mathbf{w}$ satisfies conditions (9) for derivations and therefore can be presented in any local coordinate system as

$$\mathbf{w} = w^\alpha(\partial/\partial x^\alpha),$$

where $w^\alpha \in \Im$. Consequently, in any such system $\mathbf{u}$ can be presented in the form (11) with $u^\alpha = w^\alpha$ and $u^5 = v$. ∎

One may observe that the operator corresponding to a given four-vector $\mathbf{U}$ is exactly the differential part of the operator that corresponds to any five-vector belonging to $\mathbf{U}$. This coincidence is a manifestation of the fact that V_4 is isomorphic to $\mathcal{Z}$. This does not mean, however, that one can identify four-vectors with $\mathcal{Z}$-components of five-vectors, for as one will see in section 3, the isomorphism between V_4 and $\mathcal{Z}$ is not preserved by parallel transport.

The representation of five-vectors with operators enables one to introduce the former in another way: as maps $\Im \to \Im$ that satisfy requirements (13). In its mathematical qualities, such a definition of five-vectors is superior to the one given in section 1 and enables one to introduce in a natural way the commutator of two five-vector fields. On the other hand, in this case one cannot see as clearly the correspondence between five-vectors and parametrized curves, and this is why in this paper I have first considered the representation of five-vectors in the form of equivalence classes associated with relation (3). It turns out, however, that one should make a distinction between a given equivalence class and the five-vector corresponding to it. In view of this, in the following five-vector fields will always be identified with operators satisfying requirements (13), the set of which will be denoted as $\boldsymbol{\mathcal{F}}$.

As in the case of four-vectors, tangent five-vectors at a given point Q can be defined as equivalence classes of maps from $\boldsymbol{\mathcal{F}}$ with respect to the equivalence relation

$$\mathbf{u} \equiv \mathbf{v} \Leftrightarrow \mathbf{u}[\,f\,](Q) = \mathbf{v}[\,f\,](Q) \text{ for any } f \in \Im.$$

The algebraic properties of the five-vectors defined this way are the same as of those defined as classes of equivalent curves, and their analysis would have been almost an exact repetition of the one made in section 1, except for a few obvious changes in the definitions. Let me only mention that a *regular* five-vector basis can now be defined as a basis where all $\mathbf{e}_\alpha$ are purely differential operators and $\mathbf{e}_5 = \mathbf{1}$.

One should also note that the correspondence between equivalence classes of parametrized curves and operators from $\boldsymbol{\mathcal{F}}$ given by formula (12) is not the only one possible. A more general form of such a correspondence is

$$\mathbf{a} \mapsto a \cdot \partial_{\mathbf{a}} + b \cdot \lambda_{\mathbf{a}} \cdot \mathbf{1}, \tag{14}$$

where a and b are some nonzero coefficients independent of $\mathbf{a}$. Since the overall normalization of the operators representing five-vectors is of no importance, one can always choose it so that $a = 1$. In formula (12) the second coefficient has been selected in the simplest way: $b = 1$. However, as one will see in section 3, to give a consistent definition to the five-vectors associated with curves parametrized by *dimensional* parameters, one has to assign to b a certain dimension, so it will equal unity only at some particular choice of the corresponding measurement units.

B. *Commutator of five-vector fields*

The representation of five-vectors with operators enables one to introduce the commutator of five-vector fields. Namely, if $\mathbf{u} = u^\alpha(\partial/\partial x^\alpha) + u^5 \cdot \mathbf{1}$ and $\mathbf{v} = v^\alpha\ (\partial/\partial x^\alpha) + v^5 \cdot \mathbf{1}$, then by definition,

$$[\mathbf{u}, \mathbf{v}](f) = \mathbf{u}(\mathbf{v}(f)) - \mathbf{v}(\mathbf{u}(f))$$

for any scalar function f, and one can show that $\mathbf{w} \equiv [\mathbf{u}, \mathbf{v}]$ is an operator of the form (11) with components

$$w^A = u^\beta(\partial v^A/\partial x^\beta) - v^\beta(\partial u^A/\partial x^\beta). \tag{15}$$

For an arbitrary five-vector basis $\mathbf{e}_A$ one can define the commutation constants, $C_{AB}{}^D$, as

$$[\mathbf{e}_A, \mathbf{e}_B] = C_{AB}{}^D \mathbf{e}_D,$$

and show that the components of $[\mathbf{u}, \mathbf{v}]$ in this basis are

$$\partial_{\mathbf{u}} v^A - \partial_{\mathbf{v}} u^A + u^B v^D C_{BD}{}^A.$$

This is the analog of the well-known formula for components of the commutator of two four-vector fields $\mathbf{U}$ and $\mathbf{V}$ in an arbitrary basis $\mathbf{E}_\alpha$:

$$\partial_{\mathbf{U}} V^\mu - \partial_{\mathbf{V}} U^\mu + U^\alpha V^\beta C_{\alpha\beta}{}^\mu,$$

where $[\mathbf{E}_\alpha, \mathbf{E}_\beta] = C_{\alpha\beta}{}^\mu \mathbf{E}_\mu$.

If $\mathbf{e}_A$ is a standard basis, one has ${C_{\mu 5}}^{\alpha} = 0$. It is a simple matter to show that if $\mathbf{u} \in \mathbf{U}$ and $\mathbf{v} \in \mathbf{V}$, then $[\mathbf{u}, \mathbf{v}] \in [\mathbf{U}, \mathbf{V}]$. Thus, if $\mathbf{e}_\alpha \in \mathbf{E}_\alpha$, then ${C_{\alpha\beta}}^{\mu}|_{\text{for five-vectors}} = {C_{\alpha\beta}}^{\mu}|_{\text{for four-vectors}}$.

Let us now consider two subsets of five-vector fields from $\boldsymbol{\mathcal{F}}$: (i) the subset $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$ of all purely differential operators, and (ii) the subset $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$ of all purely algebraic operators. It is evident that any element of $\boldsymbol{\mathcal{F}}$ can be uniquely presented as a sum of an operator from $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$ and an operator from $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$, so $\boldsymbol{\mathcal{F}} = \boldsymbol{\mathcal{F}}_{\mathcal{Z}} \oplus \boldsymbol{\mathcal{F}}_{\mathcal{E}}$. The components of an arbitrary five-vector field $\mathbf{u}$ in these two subspaces will be denoted as $\mathbf{u}^{\mathcal{Z}}$ and $\mathbf{u}^{\mathcal{E}}$. It is evident that they correspond to the operators $u^\alpha(\partial/\partial x^\alpha)$ and $u^5 \cdot \mathbf{1}$, respectively.

One can easily see that the commutator of two five-vector fields from $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$ is, again, a field from $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$, so $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$ is a subalgebra: $[\boldsymbol{\mathcal{F}}_{\mathcal{Z}}, \boldsymbol{\mathcal{F}}_{\mathcal{Z}}] \subset \boldsymbol{\mathcal{F}}_{\mathcal{Z}}$. Furthermore, the commutator of a field from $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$ with any other field from $\boldsymbol{\mathcal{F}}$ is an element of $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$, so $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$ is an ideal: $[\boldsymbol{\mathcal{F}}_{\mathcal{E}}, \boldsymbol{\mathcal{F}}] \subset \boldsymbol{\mathcal{F}}_{\mathcal{E}}$.

Commutators of four-vector fields enable one to tell whether or not a given four-vector basis is holonomic. Namely, for a given set of basis fields $\mathbf{E}_\alpha$ there exists a system of local coordinates x^α such that $\mathbf{E}_\alpha$ are tangent vectors to coordinate lines ($\mathbf{E}_\alpha = \partial/\partial x^\alpha$) iff $[\mathbf{E}_\alpha, \mathbf{E}_\beta] = 0$. A similar statement for five-vectors is the following:

> For a given set of standard five-vector basis fields $\mathbf{e}_A$ there exists a system of local coordinates x^α such that $\mathbf{e}_\alpha$ are tangent five-vectors to coordinate lines iff
>
> $$[\mathbf{e}^{\mathcal{Z}}_\alpha, \mathbf{e}^{\mathcal{Z}}_\beta] = 0, \tag{16a}$$
> $$[\mathbf{e}^{\mathcal{Z}}_\alpha, \mathbf{e}^{\mathcal{E}}_\beta] = \delta_{\alpha\beta} \cdot \mathbf{1}. \tag{16b}$$
>
> where $\delta_{\alpha\beta}$ is the Kronecker symbol.[1]

Proof: If $\mathbf{e}_\alpha$ are tangent vectors to coordinate lines x^α, then $\mathbf{e}_\alpha = \partial/\partial x^\alpha + x^\alpha \cdot \mathbf{1}$, and equations (16) are evidently obeyed.

If $\mathbf{e}_\alpha$ satisfy equations (16) and $\mathbf{E}_\alpha$ are such that $\mathbf{e}_\alpha \in \mathbf{E}_\alpha$, then

$$\begin{aligned} 0 = [\mathbf{e}^{\mathcal{Z}}_\alpha, \mathbf{e}^{\mathcal{Z}}_\beta] &= \partial_{\mathbf{e}_\alpha}\partial_{\mathbf{e}_\beta} - \partial_{\mathbf{e}_\beta}\partial_{\mathbf{e}_\alpha} \\ &= \partial_{\mathbf{E}_\alpha}\partial_{\mathbf{E}_\beta} - \partial_{\mathbf{E}_\beta}\partial_{\mathbf{E}_\alpha} = [\mathbf{E}_\alpha, \mathbf{E}_\beta], \end{aligned}$$

and by virtue of the corresponding theorem for four-vectors, there exists a system of local coordinates x^α such that $\partial/\partial x^\alpha = \partial_{\mathbf{E}_\alpha} = \partial_{\mathbf{e}_\alpha}$. In these coordinates each $\lambda_{\mathbf{e}_\alpha}$ is a certain real function, which according to (16b) satisfies the equation

$$\partial\lambda_{\mathbf{e}_\beta}(x)/\partial x^\alpha = \delta_{\alpha\beta}.$$

This is only possible if $\lambda_{\mathbf{e}_\alpha}(x) = x^\alpha + c^\alpha$, where c^α are integration constants. Consequently, one has $\mathbf{e}_\alpha = \partial/\partial y^\alpha + y^\alpha \cdot \mathbf{1}$, where $y^\alpha = x^\alpha + c^\alpha$. ∎

By analogy with four-vectors, a standard five-vector basis satisfying requirements (16) can be called a *coordinate* basis. In certain cases, however, it proves to be more convenient to select the $\mathcal{E}$-components of the first four basis five-vectors in a different way, for example, equal to zero. Since such five-vector bases still correspond to a coordinate four-vector basis, it makes sense to call them coordinate, too.

C. *Five-vector Lie derivative*[2]

The formal definition of the Lie derivative with respect to a four-vector field $\mathbf{U}$ is the following:

- the Lie derivative of a four-vector field $\mathbf{V}$ is

$$\boldsymbol{\pounds}_{\mathbf{U}}\mathbf{V} \equiv [\mathbf{U}, \mathbf{V}]; \tag{17}$$

- the Lie derivative of a scalar function f is

$$\boldsymbol{\pounds}_{\mathbf{U}} f \equiv \mathbf{U}f; \tag{18}$$

- the Lie derivatives of all other four-tensor fields can be found from formulae (17) and (18) by using the Leibniz rule, which in schematic form can be presented as

$$\boldsymbol{\pounds}_{\mathbf{U}}(\mathcal{A} * \mathcal{B}) = \boldsymbol{\pounds}_{\mathbf{U}}\mathcal{A} * \mathcal{B} + \mathcal{A} * \boldsymbol{\pounds}_{\mathbf{U}}\mathcal{B}, \tag{19}$$

where $\mathcal{A}$ and $\mathcal{B}$ are any two four-tensor fields and $*$ denotes contraction or tensor product.

In a similar manner one can give a formal definition to the Lie derivative with respect to a *five*-vector field $\mathbf{u}$. I will denote this latter derivative as $\boldsymbol{\pounds}_{\mathbf{u}}$ and will call it the *five-vector Lie derivative*. The analog of rule (17) is quite apparent:

- the five-vector Lie derivative of a five-vector field $\mathbf{v}$ is

$$\boldsymbol{\pounds}_{\mathbf{u}}\mathbf{v} \equiv [\mathbf{u}, \mathbf{v}]. \tag{20}$$

As the analog of rule (18) it seems reasonable to take the following one:

- the five-vector Lie derivative of a scalar function f is

$$\boldsymbol{\pounds}_{\mathbf{u}} f \equiv \mathbf{u}f. \tag{21}$$

[1] For simplicity, this theorem is formulated and proved for $a = b = 1$ in formula (14).

[2] The contents of this subsection is not necessary for understanding the rest of the material and can be skipped by a reader not familiar enough with or not interested in this particular subject.

It is easy to check that the five-vector Lie derivative of the product of two scalar functions and the five-vector Lie derivative of the product of a scalar function and a five-vector field are expressed in terms of the five-vector Lie derivatives of the factors not according to the Leibniz rule but according to the rule

$$\pounds_{\mathbf{u}}(\mathcal{A}*\mathcal{B}) = \pounds_{\mathbf{u}}\mathcal{A}*\mathcal{B}+\mathcal{A}*\pounds_{\mathbf{u}}\mathcal{B}-\pounds_{\mathbf{u}}1\cdot(\mathcal{A}*\mathcal{B}), \quad (22)$$

where, as before, *1* is the constant unity scalar function. In view of this, it is not clear which of the rules — (19), (22) or some other — should hold for the contraction and tensor product. To answer this question and to gain a better understanding of the five-vector Lie derivative, let us find for the latter an interpretation similar to the one that can be given to the ordinary Lie derivative in terms of the one-parameter local group of diffeomorphisms generated by a four-vector field.

Let us recall that any sufficiently smooth four-vector field $\mathbf{U}$ defines in the neighbourhood of any point Q of the space-time manifold $\mathcal{M}$ a congruence of integral curves, and that there always exist such an open neighbourhood $\mathcal{U}$ of Q and such a real number $\varepsilon > 0$ that the map ϕ_t obtained by taking each point of $\mathcal{U}$ a parametric distance t along the corresponding integral curve, at $|t| < \varepsilon$ is a diffeomorphism of $\mathcal{U}$ into $\mathcal{M}$. At sufficiently small s and t one has $\phi_s \circ \phi_t = \phi_{s+t}$ and $(\phi_t)^{-1} = \phi_{-t}$, so these diffeomorphisms form a one-parameter local group.

At each t map ϕ_t defines a certain transformation, Φ_t, of scalar functions: the image $\Phi_t\{f\}$ of a scalar function f is such that

$$\Phi_t\{f\}|_{\phi_t(P)} = f|_P. \quad (23)$$

This transformation, in its turn, generates a certain transformation of four-vector and other four-tensor fields, which is determined by the following rules:

• the image $\Phi_t\{\mathbf{V}\}$ of a four-vector field $\mathbf{V}$ is such that for any scalar function f

$$\Phi_t\{\mathbf{V}\}\Phi_t\{f\} = \Phi_t\{\mathbf{V}f\}; \quad (24)$$

• the image $\Phi_t\{\widetilde{\mathbf{W}}\}$ of a four-vector 1-form field $\widetilde{\mathbf{W}}$ is such that for any four-vector field $\mathbf{V}$

$$< \Phi_t\{\widetilde{\mathbf{W}}\}, \Phi_t\{\mathbf{V}\} > = \Phi_t\{< \widetilde{\mathbf{W}}, \mathbf{V} >\}; \quad (25)$$

• the image $\Phi_t\{\mathcal{A}\otimes\mathcal{B}\}$ of the tensor product of two four-tensor fields $\mathcal{A}$ and $\mathcal{B}$ is such that

$$\Phi_t\{\mathcal{A}\otimes\mathcal{B}\} = \Phi_t\{\mathcal{A}\}\otimes\Phi_t\{\mathcal{B}\}. \quad (26)$$

Within this approach, the Lie derivative of an arbitrary four-tensor field $\mathcal{S}$ is defined as

$$\pounds_{\mathbf{U}}\mathcal{S} \equiv -(d/dt)\,\Phi_t\{\mathcal{S}\}|_{t=0}. \quad (27)$$

It is easy to see that at small t

$$\Phi_t\{f\} = f - t\cdot\mathbf{U}f + O(t^2), \quad (28)$$

from which, using definition (27), one obtains rule (18). In a similar manner, after rewriting equation (24) as

$$\Phi_t\{\mathbf{V}\}f = \Phi_t\{\mathbf{V}\Phi_t^{-1}\{f\}\}$$

and using definition (27), one obtains rule (17). From equation (25) it follows that the Leibniz rule holds for the contraction of a four-vector field and a four-vector 1-form field and from equation (26) it follows that it also holds for the tensor product of any two four-tensor fields. Thus, the definition of the Lie derivative by means of equations (23)–(27) is equivalent to its formal definition according to equations (17)–(19).

It is now apparent that to obtain the desired interpretation of the five-vector Lie derivative, one should associate with every sufficiently smooth five-vector field a certain one-parameter group of transformations of scalar functions and five-tensor fields. Let us denote the transformations from this group as Ψ_t and define the five-vector Lie derivative of an arbitrary five-tensor field $\mathcal{S}$ as

$$\pounds_{\mathbf{u}}\mathcal{S} \equiv -(d/dt)\,\Psi_t\{\mathcal{S}\}|_{t=0}. \quad (29)$$

Considering what has been said above, it seems reasonable to take that at small t

$$\Psi_t\{f\} = f - t\cdot\mathbf{u}f + O(t^2) \quad (30)$$

for any scalar function f, which together with definition (29) gives us rule (21). If, by analogy with rule (24), one then takes that

$$\Psi_t\{\mathbf{v}\}\Psi_t\{f\} = \Psi_t\{\mathbf{v}f\} \quad (31)$$

for any $\mathbf{v}$ and f, from formulae (29) and (30) one will obtain rule (20). Thus, the infinitesimal transformation (30) produces the desired result. Let us now find the corresponding finite transformation.

It is evident that for any sufficiently smooth five-vector field $\mathbf{u}$, in the vicinity of any point Q one can construct a congruence of integral curves of the corresponding four-vector field $\mathbf{U}$. In this case these curves will be called the integral curves of field $\mathbf{u}$. It is not difficult to prove that at finite t the image $\Psi_t\{f\}$ of any scalar function f of class C^∞ equals

$$\Psi_t\{f\}(\lambda) = \exp\{-\int_{\lambda-t}^{\lambda} u^5(\lambda')\,d\lambda'\}\,f(\lambda - t), \quad (32)$$

where λ is the parameter of the integral curve of field $\mathbf{u}$ and u^5 is the fifth component of the latter in a regular basis. We thus see that transformation Ψ_t consists

in "shifting" every value of the function a parametric distance t along the corresponding integral curve and then multiplying it by a certain exponential factor. It is easy to see that this latter factor equals the corresponding value of $\Psi_t\{1\}$, so for an arbitrary scalar function f one has

$$\Psi_t\{f\} = \Psi_t\{1\}\Phi_t\{f\}. \tag{33}$$

From the latter formula it follows that transformations Φ_t induced by four-vector fields are a particular case of transformations Ψ_t — a case that corresponds to the five-vector fields from $\boldsymbol{\mathcal{F}}_{\mathcal{Z}}$. Another particular case are the transformations Ψ_t induced by five-vector fields from $\boldsymbol{\mathcal{F}}_{\mathcal{E}}$. In this case

$$\Psi_t\{f\}(P) = \exp\{-t \cdot u^5(P)\}f(P).$$

It is evident that to each transformation Ψ_t one can put into correspondence a certain map of $\mathcal{U}$ into $\mathcal{M}$, namely, the map ϕ_t induced by the four-vector field corresponding to $\mathbf{u}$. Thus, both in the case of four-vector fields and in the case of five-vector fields one is actually dealing with *two* maps: (i) a map from $\mathcal{U}$ to $\mathcal{M}$ and (ii) a map from the set of restrictions to $\mathcal{U}$ of all the functions from $\Im$ to the set of restrictions of all these functions to $\phi_t(\mathcal{U})$. In the case of four-vector fields there exists a one-to-one correspondence bewteen these two maps, which enables one to think that the second map is induced by the first one. This is not so in the case of five-vector fields: for example, the identity map from $\mathcal{U}$ to $\mathcal{M}$ may correspond to different *non*identical transformations of scalar functions.

From equation (32) it is not difficult to derive that for any two scalar functions f and g

$$\Psi_t\{fg\} = \Phi_t\{f\}\Psi_t\{g\} = \Psi_t\{f\}\Phi_t\{g\}, \tag{34}$$

so in the general case the image of the product of two scalar functions with respect to Ψ_t is not the product of their images. By substituting $\Psi_t^{-1}\{1\}\Psi_t\{f\}$ for $\Phi_t\{f\}$ in formula (34) and differentiating both sides of the latter with respect to t, one can verify that in this case rule (22) is indeed obeyed.

It is natural to define the action of Ψ_t on a tensor product in the following way:

$$\Psi_t\{\mathcal{A}\otimes\mathcal{B}\} = \Psi_t\{\mathcal{A}\}\otimes\Psi_t\{\mathcal{B}\}, \tag{35}$$

where $\mathcal{A}$ and $\mathcal{B}$ are any two five-tensor fields of nonzero rank. This formula does not work, however, if one of the fields or both of them are of rank zero. In the second case this can be seen from formula (34), if one considers that for scalar functions $f\otimes g = fg$. In the first case, if, for example, $\mathcal{A} = f$ and $\mathcal{B} = \mathbf{v}$, from formula (34) and definition (31) one can easily obtain that

$$\Psi_t\{f\otimes\mathbf{v}\} = \Psi_t\{f\mathbf{v}\} = \Phi_t\{f\}\Psi_t\{\mathbf{v}\}. \tag{36}$$

Difficulties also occur with the definition of the action of Ψ_t on five-vector 1-forms. The direct analog of rule (25) is

$$<\Psi_t\{\widetilde{\mathbf{w}}\},\Psi_t\{\mathbf{v}\}> = \Psi_t\{<\widetilde{\mathbf{w}},\mathbf{v}>\}, \tag{37}$$

which means that the operation of contraction is "correlated" with transformation Ψ_t in the sense that the contraction of the image of a five-vector field $\mathbf{v}$ with the image of a five-vector 1-form field $\widetilde{\mathbf{w}}$ equals the image of the scalar function equal to the contraction of $\mathbf{v}$ with $\widetilde{\mathbf{w}}$. The quantity $<\widetilde{\mathbf{w}},\mathbf{v}>$ can also be regarded as a five-tensor field of rank zero obtained by contracting the field $\widetilde{\mathbf{w}}\otimes\mathbf{v}$ of rank $(1,1)$. A similar operation can be performed on other five-tensor fields, for example, on the field $\widetilde{\mathbf{w}}\otimes\mathbf{v}\otimes\mathbf{s}$. For the contraction of this latter field to be correlated with Ψ_t it is necessary that there would hold not rule (37) but the rule

$$<\Psi_t\{\widetilde{\mathbf{w}}\},\Psi_t\{\mathbf{v}\}> = \Phi_t\{<\widetilde{\mathbf{w}},\mathbf{v}>\}. \tag{38}$$

Therefore, in those cases where Ψ_t does not coincide with Φ_t, the requirements of correlation between the contraction and transformation Ψ_t for five-tensor fields of rank $(1,1)$ and for five-tensor fields of other ranks are conflicting.

It is also useful to look at the components of the five-vector Lie derivatives of five-tensor fields of different ranks, in a regular coordinate basis. Let us write out these components for the case where the rule that determines the action of Ψ_t on 1-form fields is

$$<\Psi_t\{\widetilde{\mathbf{w}}\},\Psi_t\{\mathbf{v}\}> = (\Psi_t\{1\})^k\Psi_t\{<\widetilde{\mathbf{w}},\mathbf{v}>\}. \tag{39}$$

According to equation (21), the five-vector Lie derivative of function f is

$$\boldsymbol{\pounds}_{\mathbf{u}}f = u^\alpha\partial_\alpha f + u^5 f. \tag{40}$$

From equations (20) and (15) one finds that the components of the five-vector Lie derivative of a five-vector field $\mathbf{v}$ are

$$(\boldsymbol{\pounds}_{\mathbf{u}}\mathbf{v})^A = u^B(\partial_B v^A) - v^B(\partial_B u^A), \tag{41}$$

where, for convenience, I have introduced the notation $\partial_A \equiv \partial_{\mathbf{e}_A}$, so $\partial_\alpha = \partial/\partial x^\alpha$ and $\partial_5 = 0$. From equation (39) one can easily derive that in the dual basis of five-vector 1-forms $\widetilde{\mathbf{o}}^A$,

$$\begin{aligned}(\boldsymbol{\pounds}_{\mathbf{u}}\widetilde{\mathbf{w}})_A = u^B(\partial_B w_A) + w_B(\partial_A u^B)\\ + (1+k)\,u^5 w_A\end{aligned} \tag{42}$$

Finally, in the general case of an arbitrary five-tensor field

$$\mathbf{T} = T^{A_1\ldots A_m}_{B_1\ldots B_n}\, \mathbf{e}_{A_1} \otimes \ldots \otimes \mathbf{e}_{A_m} \otimes \widetilde{\mathbf{o}}^{B_1} \otimes \ldots \otimes \widetilde{\mathbf{o}}^{B_n}$$

one has

$$\begin{aligned}
(\boldsymbol{\pounds}_{\mathbf{u}}\mathbf{T})^{A_l\ldots A_m}_{B_l\ldots B_n} &= u^H(\partial_H T^{A_1\ldots A_m}_{B_1\ldots B_n}) \\
&+ n(1+k)\cdot u^5\, T^{A_1\ldots A_m}_{B_1\ldots B_n} \\
&- T^{H\ldots A_m}_{B_1\ldots B_n}(\partial_H u^{A_1}) \\
- \ldots &- T^{A_1\ldots H}_{B_1\ldots B_n}(\partial_H u^{A_m}) \\
&+ T^{A_1\ldots A_m}_{H\ldots B_n}(\partial_{B_1} u^H) \\
+ \ldots &+ T^{A_1\ldots A_m}_{B_1\ldots H}(\partial_{B_n} u^H).
\end{aligned} \tag{43}$$

As one can see from the formulae obtained, there exists a distinguished value of parameter k: $k = -1$, at which the terms proportional to u^5 in equations (42) and (43) vanish, and the five-vector Lie derivative of any five-tensor field that has at least one lower index depends only on the *derivative* of u^5, as does the five-vector Lie derivative of a five-vector field. One can also see that in the case of a five-tensor field of rank zero (at $m = n = 0$) formula (43) disagrees with formula (40) for the five-vector Lie derivative of a scalar function.

All these observations suggest that in the case of transformations Ψ_t induced by five-vector fields, one should make a distinction between scalar functions which are elements of $\Im$ and scalar functions which are five-tensor fields of rank zero. Formally, these two types of objects are of different nature: the former are the functions upon which act the operators of five-vector fields; the latter are elements of a commutative ring, by which one can multiply five-vector fields, obtaining five-vector fields again. To establish order in the theory, one should suppose that these two types of functions are transformed by Ψ_t *differently*: the elements of $\Im$ are transformed according to formula (32), whereas the five-tensor fields of rank zero are transformed according to the formula

$$\Psi_t\{\mathrm{f}\}(\lambda) = \mathrm{f}(\lambda - t), \tag{44}$$

which means that for them transformation Ψ_t coincides with Φ_t. Under this assumption formula (35) for the tensor product will be valid for five-tensor fields of zero rank as well. Moreover, since the contraction of a vector and a 1-form is a tensor of rank zero, formula (37) will coincide with formula (38), and consequently the contraction will be correlated with transformation Ψ_t for tensor fields of any rank for which it makes sense. Among other things, the latter two facts mean that the five-vector Lie derivative of a contraction and of a tensor product is expressed in terms of the five-vector Lie derivatives of the factors according to the Leibniz rule. In formulae (42) and (43) one should now put $k = -1$, and so the derivatives $\boldsymbol{\pounds}_{\mathbf{u}}$ of the corresponding five-tensor fields will depend only on the derivative of u^5. Finally, the five-vector Lie derivative of an arbitrary five-tensor field f of rank zero will be

$$\boldsymbol{\pounds}_{\mathbf{u}}\mathrm{f} = \partial_{\mathbf{u}}\mathrm{f} = u^\alpha \partial_\alpha \mathrm{f}, \tag{45}$$

which agrees with formula (43). Let me emphasize once more that in the case of scalar fields from $\Im$, the image of the product of two such functions with respect to Ψ_t will not equal the product of their images, which is inevitable and has no relation to the definition of Ψ_t for five-tensor fields.

3. Some other properties of five-vectors

A. *Parallel transport of five-vectors*

As for any other type of vector-like objects considered in space-time, one can speak of parallel transport of five-vectors from one space-time point to another. One can then define the covariant derivative of five-vector fields; introduce the connection coefficients corresponding to a given five-vector basis; construct the corresponding curvature tensor; etc. In doing all this one does not have to use in any way the fact that five-vectors are associated with space-time by their definition.

One should expect that the origin of five-vectors manifests itself in that the rules of their parallel transport are related in some way to similar rules for four-vectors and to the Riemannian geometry of space-time. It is obvious that this relation cannot be derived from the algebraic properties of five-vectors, and to obtain it one has to make some new assumptions about five-vectors, which ought to be regarded as part of their definition.

Let us first consider the relation between the rules of parallel transport for four- and five-vectors. The simplest and the most natural form of this relation is obtained by postulating that parallel transport preserves the algebraic relation between four- and five-vectors discussed in subsection 1.C. A more precise formulation of this statement is the following:

> *If four-vector* $\mathbf{U}$ *is the equivalence class of five-vector* $\mathbf{u}$*, then the transported* $\mathbf{U}$ *is the equivalence class of the transported* $\mathbf{u}$*.* (46)

This assumption is quite natural considering that $\mathbf{u} \in \mathbf{U}$ means that $\mathbf{u}$ and $\mathbf{U}$ correspond to the same

direction in the manifold. It has two consequences, which can be conveniently expressed in terms of connection coefficients (the latter are defined in section 4 of part I).

Let us consider the parallel transport of vectors from an arbitrary point Q to a nearby point Q'. If two five-vectors at Q belong to the same equivalence class, then according to our assumption, the transported five-vectors should also be equivalent. Since parallel transport is a linear operation, this means that vectors from $\mathcal{E}_{\mathrm{at}\ Q}$ are transported into vectors from $\mathcal{E}_{\mathrm{at}\ Q'}$. Consequently, in *any* standard five-vector basis,

$$G^{\alpha}_{5\mu} = 0. \tag{47}$$

Let $\mathbf{e}_A$ be an arbitrary standard five-vector basis and let $\mathbf{E}_\alpha$ be the associated basis of four-vectors. If $\mathbf{E}_\alpha(Q)$ are transported into vectors $\mathbf{E}_\beta(Q')C^\beta_\alpha$, then according to our assumption, $\mathbf{e}_\alpha(Q)$ should be transported into vectors $\mathbf{e}_\beta(Q')C^\beta_\alpha + \mathbf{e}_5(Q')C^5_\alpha$, where the coefficients C^β_α are the same in both cases. This means that in the selected bases,

$$G^{\alpha}_{\beta\mu} = \Gamma^{\alpha}_{\beta\mu}. \tag{48}$$

It is evident that assumption (46) tells one nothing about $G^5_{\alpha\mu}$ and $G^5_{5\mu}$. To get an idea of what these coefficients can be like, let us now consider a particular case where the connection for five-vector fields is such that there exists a certain local symmetry which can be formulated as the following principle:

> For any set of scalar, five-vector and five-tensor fields defined in the vicinity of any point Q in space-time, by means of a certain procedure one can construct a set of fields in the vicinity of any other point Q', such that at Q' these new fields (which will be called *equivalent*) satisfy the same algebraic and first-order differential relations that the original fields satisfy at Q. (49)

The procedure by means of which the equivalent fields are constructed can be formulated as follows:

1. Introduce at Q a system of local Lorentz coordinates x^α.
   Introduce the corresponding *regular coordinate* five-vector basis $\mathbf{e}_A$.
   Introduce the corresponding bases for all other five-tensors.
2. Each scalar field f in the vicinity of Q will then determine and be determined by one real coordinate function $f(x)$.
   Each five-vector field $\mathbf{u}$ in the vicinity of Q will determine and be determined by five real coordinate functions $u^A(x)$ (= components of $\mathbf{u}$ in the basis $\mathbf{e}_A$).
   Each five-tensor field $\mathbf{T}$ in the vicinity of Q will determine and be determined by an appropriate number of real coordinate functions $T^{AB\dots C}_{DE\dots F}(x)$ (= components of $\mathbf{T}$ in the relevant tensor basis corresponding to $\mathbf{e}_A$).
3. Introduce at Q' a system of local Lorentz coordinates x'^α such that $x'^\alpha(Q') = x^\alpha(Q)$.
   Introduce the corresponding regular coordinate five-vector basis $\mathbf{e}'_A$.
   Introduce the corresponding bases for all other five-tensors.
4. Then the equivalent scalar, five-vector and five-tensor fields in the vicinity of Q' will be determined in coordinates x'^α and in the corresponding bases by the *same functions* $f(\cdot)$, $u^A(\cdot)$, …, $T^{AB\dots C}_{DE\dots F}(\cdot)$ that determine the original fields in the vicinity of Q in coordinates x^α and in the corresponding bases.

At $Q' = Q$ the two mentioned systems of local Lorentz coordinates, x^α and x'^α, are related as follows:

$$x'^\alpha(P) = x^\alpha(Q) + \Lambda^\alpha_\beta[x^\beta(P) - x^\beta(Q)] + \text{ terms of order } [x^\alpha(P) - x^\alpha(Q)]^3,$$

where P is an arbitrary point in the vicinity of Q and Λ^α_β is a matrix from O(3,1). Reasoning as in section 4 of part I, one can show that in the regular basis associated with either of these coordinate systems one should have $G^\alpha_{\beta\mu}(Q) = G^5_{5\mu}(Q) = 0$ and $G^5_{\alpha\mu}(Q) \propto \eta_{\alpha\mu}$. Since in these coordinates $g_{\alpha\beta}(Q) \propto \eta_{\alpha\beta}$, too, this means that the connection coefficients $G^5_{\alpha\mu}(Q)$ are proportional to the components of the metric tensor. Denoting the proportionality factor as $-\varsigma$ and using the obvious transformation formulae for five-vector connection coefficients, one can show that in *any* regular five-vector basis

$$G^5_{5\mu} = 0 \tag{50}$$

and

$$G^5_{\alpha\mu} = -\varsigma g_{\alpha\mu}. \tag{51}$$

From requirement (49) it also follows that five-vector connection coefficients should have the same form at any two points in space-time in similar five-vector bases. In the case of four-vector connection coefficients a similar condition is satisfied automatically, and therefore is not necessary. For five-vectors this is a nontrivial requirement, which means that ς in equation (51) should be a constant.

It is evident that the value of ς is not fixed by the symmetry principle. Since for dimensionless coordinates and curve parameters the connection coefficients are dimensionless and $g_{\alpha\beta}$ are measured in the units of interval squared, ς should have the dimension $(interval)^{-2}$. There is no sense in talking about five-vectors if $\varsigma = 0$, for it is impossible to distinguish a five-vector with such rules of parallel transport from a pair consisting of a four-vector and a scalar. Indeed, V_5 is isomorphic to the direct sum of V_4 and the space of scalars (regarded as one-dimensional vectors), and it is apparent that at $\varsigma = 0$ this isomorphism is preserved by parallel transport. Considering this, I will always assume that $\varsigma \neq 0$.

B. *Five-vectors associated with dimensional curve parameters*

So far we have been dealing with dimensionless curve parameters and coordinates. In practice, the latter are usually selected in such a way so that their values would be associated in some particular way with certain lengths, time intervals or angles determined by the space-time metric. For example, any system of dimensionless Lorentz coordinates in flat space-time is such that the square of the interval between any two events A and B, measured in certain units ℓ, equals

$$[x^0(A) - x^0(B)]^2 - [x^1(A) - x^1(B)]^2 - [x^2(A) - x^2(B)]^2 - [x^3(A) - x^3(B)]^2.$$

It is evident that if one changes the unit for measuring the interval as

$$\ell \to k\ell \ \ (k > 0), \tag{52}$$

the dimensionless Lorentz coordinates will change in the inverse proportion. This enables one to consider the latter as numerical values of certain dimensional quantities, $\bar{x}^\alpha$, measured in the units of interval, and it is these latter quantities one usually has in mind when using the term "Lorentz coordinates".

The situation is similar in all other cases and as in the above example, enables one to introduce the corresponding dimensional coordinates. For simplicity, in the following I will suppose that all four coordinates are measured in the units of interval. A convenient property of such dimensional coordinates is that the corresponding metric coefficients, defined by the equation

$$ds^2 = \bar{g}_{\alpha\beta}\, d\bar{x}^\alpha d\bar{x}^\beta,$$

are all dimensionless quantities. It is easy to see that $\bar{g}_{\alpha\beta}$ are the values of the dimensional metric coefficients $g_{\alpha\beta}$ that correspond to the dimensionless coordinates x^α which are the values of $\bar{x}^\alpha$ at the given ℓ.

The same idea can be used to define dimensional curve parameters (for simplicity, let us consider only those of them which are measured in the units of interval). One can then introduce the notion of a tangent four-vector corresponding to a curve parameterized by a given dimensional parameter $\bar{\lambda}$. Such four-vectors behave as dimensional quantities in the sense that at each ℓ they have a certain "value", which, by definition, is the four-vector that corresponds to the dimensionless parameter λ which is the value of $\bar{\lambda}$ for the given ℓ. The algebraic operations and parallel transport for such dimensional four-vectors are defined on the basis of the corresponding operations for four-vectors associated with dimensionless parameters. For example, a sum of two dimensional four-vectors $\mathbf{U}$ and $\mathbf{V}$ is a dimensional four-vector whose value at any ℓ equals the sum of the corresponding values of $\mathbf{U}$ and $\mathbf{V}$. It is evident that when one changes ℓ according to formula (52), the value of each dimensional four-vector changes in the same proportion, owing to which the inner product of any two such four-vectors is a dimensionless quantity. This and other properties of four-vectors associated with dimensional curve parameters are well known, and I will not discuss them any further.

Let us now see how one can define a tangent five-vector corresponding to a curve parametrized by some dimensional parameter $\bar{\lambda}$. Following the same idea that has been used for tangent four-vectors, one should consider such a five-vector as a quantity that has a certain "value" at every choice of ℓ. This "value" is the tangent five-vector that corresponds to the dimensionless parameter λ which is the value of $\bar{\lambda}$ for the given ℓ. Let us now find the operator that corresponds to this latter five-vector.

According to section 2, the general form of the operator representing the five-vector tangent to a curve parametrized by a given dimensionless parameter λ is

$$a \cdot d/d\lambda + b \cdot \lambda \cdot \mathbf{1}, \tag{53}$$

where a and b are some arbitrary nonzero constants. As it has been said above, the overall normalization of the operators representing five-vectors can always be chosen in such a way that a be unity. When dimensionless curve parameters are considered by themselves—not as values of some dimensional parameters, one can take $b = 1$, too, as it has been done in formula (12). However, if operator (53) represents the value of a five-vector associated with a dimensional parameter $\bar{\lambda}$, the value of b has to depend on the choice of ℓ. Indeed, let us suppose that one has a dimensional five-vector, $\mathbf{u}$, represented by a purely differential operator and one parallel transports it from a given space-time point Q to some other

point Q'. By definition, $\mathbf{u}^{\text{transported}}$ is the five-vector at Q' whose value at any ℓ equals the value of $\mathbf{u}$ at Q transported from Q to Q' along the selected path. It is evident that if one changes ℓ according to formula (52), the value of $\mathbf{u}$ will change in the same proportion, and since parallel transport is a linear operation, so will the value of $\mathbf{u}^{\text{transported}}$. Consequently, the algebraic part of the operator representing $\mathbf{u}^{\text{transported}}$, which in the general case will not be zero, should change in the same proportion as ℓ, which is only possible if b changes as $b \to k^2 b$.

We thus see that in the case of five-vectors associated with dimensional curve parameters, the coefficient b in formula (53) has to be the value of some nonzero constant with dimension $(interval)^{-2}$. Apart from being nonzero, this constant is absolutely arbitrary, and it is convenient to choose it equal to the constant ς introduced in the previous subsection. The operator representing a five-vector associated with a dimensional parameter $\bar{\lambda}$ can then be presented in the following form:

$$d/d\bar{\lambda} + \bar{\lambda} \cdot \varsigma \cdot \mathbf{1}. \tag{54}$$

In a similar manner one can introduce five-vectors corresponding to parameters with dimension other than that of the interval. The algebraic and differential properties of all such five-vectors will be practically the same as those of the five-vectors associated with dimensionless parameters, and only the dimension of certain relevant quantities will be different. For example, in the particular case considered above, both the inner product h' induced by the metric and the nondegenerate inner product h are dimensionless. The relation between the two is still given by formula (6), only now ξ has the dimension $(interval)^{-2}$.

In the case of dimensional five-vectors, there exist *three* convenient ways to normalize the fifth basis vector in a standard five-vector basis and, accordingly, there are three ways to define a regular basis.

In those cases where the emphasis is made on parallel transport of five-vectors, it is convenient to choose $\mathbf{e}_5 = \varsigma \cdot \mathbf{1}$. Then, in the corresponding regular basis (in the one where the other four basis five-vectors belong to $\mathcal{Z}$) one will have $G^5{}_{\alpha\mu} = -\,g_{\alpha\mu}$, and the fifth component of any five-vector $\mathbf{u}$ will equal $\lambda_{\mathbf{u}}$. In the following, such a basis will be referred to as an *active* regular basis.

In those cases where the emphasis is made on the action of five-vectors on scalar functions, it is convenient to take $\mathbf{e}_5 = \mathbf{1}$. In the corresponding regular basis one will then have $G^5{}_{\alpha\mu} = -\,\varsigma g_{\alpha\mu}$ and the fifth component of any five-vector $\mathbf{u}$ will equal $\varsigma\lambda_{\mathbf{u}}$. In the following, such a basis will be referred to as a *passive* regular basis.

Finally, in those cases where the emphasis is made on the inner product of five-vectors (at some particular choice of ξ), it is convenient to normalize $\mathbf{e}_5$ by the requirement $h(\mathbf{e}_5, \mathbf{e}_5) = \operatorname{sign}\xi$. It is evident that this equation has two solutions: $\mathbf{e}_5 = +\,|\xi|^{-1/2}\,\varsigma \cdot \mathbf{1}$ and $\mathbf{e}_5 = -\,|\xi|^{-1/2}\,\varsigma \cdot \mathbf{1}$, and to be definite, I will choose the first one. In the corresponding regular basis one will then have $G^5{}_{\alpha\mu} = -\,|\xi|^{1/2} g_{\alpha\mu}$, and the fifth component of any five-vector $\mathbf{u}$ will equal $|\xi|^{1/2}\lambda_{\mathbf{u}}$. In the following, such a basis will be referred to as a *normalized* regular basis and the operator $|\xi|^{-1/2}\,\varsigma \cdot \mathbf{1}$ will be denoted as $\mathbf{n}$.

From now on, unless it is stated otherwise, I will talk only about five-vectors associated with dimensional curve parameters and coordinates, and will omit the bar over the dimensional x^α and λ. It is evident that any result obtained for such five-vectors can readily be reformulated for five-vectors corresponding to dimensionless parameters.

C. *Four-vectors as simple bivectors over V_5*

We are now ready to demonstrate that the five-vectors introduced formally in part I can be identified with the five-dimensional tangent vectors introduced in this paper. More precisely, it will be shown that there can be established a natural isomorphism between the space of four-vectors and one of the maximal vector spaces of simple bivectors over V_5 and that in those cases where the connection for five-vectors possesses the local symmetry considered in subsection A, this isomorphism is preserved by parallel transport. This will mean that the five-vectors considered in this paper have all the formal properties postulated for five-vectors in part I.

Let us fix a nonzero five-vector $\mathbf{e} \in \mathcal{E}$ and consider all simple bivectors of the form $\mathbf{u} \wedge \mathbf{e}$, where $\mathbf{u} \in V_5$. It is evident that $\mathbf{u} \wedge \mathbf{e} = \mathbf{v} \wedge \mathbf{e}$ if and only if $\mathbf{u} - \mathbf{v} \in \mathcal{E}$, which is exactly the equivalence relation R of subsection 1.C. Thus, one is able to establish a one-to-one correspondence between four-vectors and elements of the maximal vector space of simple bivectors over V_5 with the directional vector belonging to $\mathcal{E}$. It is evident that this correspondence is a homomorphism and that it depends on the choice of the arbitrary nonzero vector $\mathbf{e}$. Let us fix the latter by requiring that the considered correspondence be an isomorphism.

Let us consider some particular nondegenerate inner product on V_5, where the constant ξ has been chosen positive, so that h would have the signature $(+---+)$. It is not difficult to check that if $\mathbf{u} \in \mathbf{U}$

and $\mathbf{v} \in \mathbf{V}$, then

$$g(\mathbf{U}, \mathbf{V}) = h(\mathbf{u}, \mathbf{v}) - \frac{h(\mathbf{e}, \mathbf{u}) h(\mathbf{e}, \mathbf{v})}{h(\mathbf{e}, \mathbf{e})}. \qquad (55)$$

On the other hand, the inner product of $\mathbf{u} \wedge \mathbf{e}$ and $\mathbf{v} \wedge \mathbf{e}$ induced by h is

$$h(\mathbf{u} \wedge \mathbf{e}, \mathbf{v} \wedge \mathbf{e}) = h(\mathbf{u}, \mathbf{v}) h(\mathbf{e}, \mathbf{e}) - h(\mathbf{u}, \mathbf{e}) h(\mathbf{v}, \mathbf{e}).$$

For the correspondence $\mathbf{U} \mapsto \mathbf{u} \wedge \mathbf{e}$ to be an isomorphism $g(\mathbf{U}, \mathbf{V})$ should equal $h(\mathbf{u} \wedge \mathbf{e}, \mathbf{v} \wedge \mathbf{e})$ for all $\mathbf{u}$ and $\mathbf{v}$, which is only possible if $h(\mathbf{e}, \mathbf{e}) = 1$. This means that $\mathbf{e}$ is either $+\mathbf{n}$ or $-\mathbf{n}$. We thus see that (for the given $\xi > 0$) there exist *two* isomorphisms of V_4 onto the considered maximal vectors space of simple bivectors, and unless additional requirements are imposed, the choice between the two is a matter of convention. To be definite, I will take $\mathbf{e} = \mathbf{n}$.

The fact that the above isomorphism (actually, both of them) is preserved by parallel transport becomes evident if one considers that the relation $\mathbf{u} \in \mathbf{U}$ is invariant under parallel transport and that $\mathbf{n}$ is transported into $\mathbf{n}$.

One can now use all the results obtained within the formal theory of five-vectors. Most of the definitions made in the present paper correspond to those made in part I. The only essential difference concerns the associated four-vector basis.

When introducing five-vectors formally, one has no means of associating them with four-dimensional tangent vectors other than saying that a five-vector $\mathbf{u}$ corresponds to the four-vector identified with the bivector $\mathbf{u} \wedge \mathbf{e}$, where $\mathbf{e}$ is some directional vector. The only way one can fix $\mathbf{e}$ within the formal theory is to require that it be of certain length. However, since the inner product of five-vectors is an object of study itself, one prefers to have a purely "kinematic" relation between the four- and five-vector bases, and the only sensible choice is to take $\mathbf{E}_\alpha = \mathbf{e}_\alpha \wedge \mathbf{e}_5$. This means that

$$\mathbf{E}_\alpha = \xi^{1/2} \lambda_{\mathbf{e}_5} \times (\text{the equivalence class of } \mathbf{e}_\alpha). \qquad (56)$$

Considering that $\xi\,(\lambda_{\mathbf{e}_5})^2 = h(\mathbf{e}_5, \mathbf{e}_5)$, from formula (55) one obtains the relation between the components of g and h derived in part I:

$$g_{\alpha\beta} = h_{55} h_{\alpha\beta} - h_{\alpha 5} h_{\beta 5}.$$

Furthermore, if $\nabla_\mu \mathbf{n} = 0$, then $G^5_{\ 5\mu} = (\lambda_{\mathbf{e}_5})^{-1} \partial_\mu \lambda_{\mathbf{e}_5}$, and for the four-vector connection coefficients corresponding to basis (56) one has

$$\Gamma^\alpha_{\ \beta\mu} = G^\alpha_{\ \beta\mu} + \delta^\alpha_\beta G^5_{\ 5\mu},$$

which is exactly the relation obtained in part I. Finally, if one assumes that flat space-time possesses the symmetry considered in subsection A, then in any orthonormal standard five-vector basis one will have

$$G^5_{\ 5\mu} = 0 \ \text{ and } \ G^5_{\ \alpha\mu} = -\kappa \eta_{\alpha\mu},$$

where $\kappa = \xi^{1/2}$ (if we had taken $\mathbf{e} = -\mathbf{n}$, we would have had $\kappa = -\xi^{1/2}$).

Let me also say a few words about the equation for the first covariant derivative of h. Straightforward calculations similar to those made in part I give the following result:

$$\begin{aligned} &\{\nabla_{\mathbf{U}} h\}(\mathbf{v}, \mathbf{w}) \\ &\quad = \kappa g(\mathbf{U}, \mathbf{V}) h(\mathbf{w}, \mathbf{n}) \\ &\qquad + \kappa g(\mathbf{U}, \mathbf{W}) h(\mathbf{v}, \mathbf{n}), \end{aligned} \qquad (57)$$

where it is assumed that $\mathbf{v} \in \mathbf{V}$ and $\mathbf{w} \in \mathbf{W}$. Since for any $\mathbf{v}$ one has $\kappa h(\mathbf{v}, \mathbf{n}) = \xi \cdot \lambda_{\mathbf{v}}$, this equation can also be presented as

$$\{\nabla_{\mathbf{U}} h\}(\mathbf{v}, \mathbf{w}) = \xi\, g(\mathbf{U}, \mathbf{V}) \lambda_{\mathbf{w}} + \xi\, g(\mathbf{U}, \mathbf{W}) \lambda_{\mathbf{v}}. \qquad (58)$$

It is easy to see that for an arbitrary nonzero $\mathbf{e} \in \mathcal{E}$ the bivectors $\mathbf{v} \wedge \mathbf{e}$ and $\mathbf{w} \wedge \mathbf{e}$ correspond to the four-vectors $\xi^{1/2} \lambda_{\mathbf{e}} \mathbf{V}$ and $\xi^{1/2} \lambda_{\mathbf{e}} \mathbf{W}$, respectively. Thus, by multiplying both sides of equation (57) by $\xi\,(\lambda_{\mathbf{e}})^2 = h(\mathbf{e}, \mathbf{e})$ one obtains

$$\begin{aligned} &h(\mathbf{e}, \mathbf{e}) \{\nabla_{\mathbf{U}} h\}(\mathbf{v}, \mathbf{w}) \\ &\quad = \kappa g(\mathbf{U}, \mathbf{v} \wedge \mathbf{e}) h(\mathbf{w}, \mathbf{e}) \\ &\qquad + \kappa g(\mathbf{U}, \mathbf{w} \wedge \mathbf{e}) h(\mathbf{v}, \mathbf{e}), \end{aligned}$$

which is exactly the equation for ∇h obtained within the formal theory of five-vectors.

D. *Operator ∇ and matrix g with five-vector indices*

Above I have introduced the covariant derivative operator, $\nabla_{\mathbf{U}}$, which differentiates five-vector fields in the direction specified by its argument—by the four-vector $\mathbf{U}$. As a consequence, the corresponding connection coefficients, $G^A_{B\mu}$, have indices of two kinds: two five-vector indices A and B and one four-vector index μ. This is not very convenient in those cases where indices of different kinds have to be permuted, for any relation with such permutations is valid only if the four- and five-vector bases have been chosen accordingly.

This inconvenience can be easily eliminated if instead of $\nabla_{\mathbf{U}}$ one considers the operator $\nabla_{\mathbf{u}}$, defined by the relation

$$\nabla_{\mathbf{u}} = \nabla_{\mathbf{U}} \ \text{ for } \ \mathbf{u} \in \mathbf{U}. \qquad (59)$$

It is obvious that $\nabla_{\mathbf{u}}$ is absolutely equivalent to $\nabla_{\mathbf{U}}$. However, unlike the latter, it formally depends on a

five-vector. It is evident that $\nabla_{\mathbf{u}} = \nabla_{(\mathbf{u}^{\mathcal{Z}})}$ for any five-vector $\mathbf{u}$, so for any $\mathbf{e} \in \mathcal{E}$ one has $\nabla_{\mathbf{e}} = 0$. Operator $\nabla_{\mathbf{u}}$ is the analog of the operator $\partial_{\mathbf{u}}$ that acts upon scalar functions, and relation (59) is the analog of the relation

$$\partial_{\mathbf{u}} = \partial_{\mathbf{U}} \quad \text{for} \quad \mathbf{u} \in \mathbf{U}.$$

It is natural to introduce the notation $\nabla_A \equiv \nabla_{\mathbf{e}_A}$. Then, in any standard five-vector basis one has $\nabla_5 = 0$ and $\nabla_\mu^{(\text{with a five-vector index})} = \nabla_\mu^{(\text{with a four-vector index})}$. In view of this, I will use the same carrier letter 'G' to denote the connection coefficients corresponding to ∇_A:

$$\nabla_A \mathbf{e}_B = \mathbf{e}_C \, G^C_{\ BA}.$$

Then $G^A_{\ B\mu}$ with a five-vector μ will equal $G^A_{\ B\mu}$ with a four-vector μ in any standard basis, and rules (47), (48), (50), and (51) will apply to $G^A_{\ BC}$ without any changes. In addition, one will have a fifth rule: that in any standard five-vector basis,

$$G^A_{\ B5} = 0.$$

In the usual manner one can derive the transformation formula for $G^A_{\ BC}$, corresponding to the basis transformation $\mathbf{e}'_A = \mathbf{e}_B L^B_{\ A}$:

$$G'^A_{\ BC} = (L^{-1})^A_{\ D} G^D_{\ EF} L^E_{\ B} L^F_{\ C} + (L^{-1})^A_{\ D} (\partial_F L^D_{\ B}) L^F_{\ C}.$$

If both bases are standard, one will have $G'^A_{\ B5} = G^A_{\ B5} = 0$ and

$$G'^A_{\ B\mu} = (L^{-1})^A_{\ D} G^D_{\ E\nu} L^E_{\ B} L^\nu_{\ \mu} + (L^{-1})^A_{\ D} (\partial_\nu L^D_{\ B}) L^\nu_{\ \mu},$$

which is the usual formula for transformation of connection coefficients.

In a similar manner one can deal with four-vector indices in $g_{\mu\nu}$. Actually, I have already defined the corresponding five-vector quantity in subsection 1.F, where it has been denoted as h'. From now on, instead of $h'(\mathbf{u}, \mathbf{v})$ I will use the notation $g(\mathbf{u}, \mathbf{v})$, so formulae (5) and (6) will acquire the form:

$$g(\mathbf{u}, \mathbf{v}) = g(\mathbf{U}, \mathbf{V})$$

for $\mathbf{u} \in \mathbf{U}$ and $\mathbf{v} \in \mathbf{V}$, and

$$h(\mathbf{u}, \mathbf{v}) = g(\mathbf{u}, \mathbf{v}) + \xi \cdot \lambda_{\mathbf{u}} \lambda_{\mathbf{v}}. \tag{60}$$

It is evident that $g(\mathbf{u}, \mathbf{v}) = g(\mathbf{u}^{\mathcal{Z}}, \mathbf{v}^{\mathcal{Z}})$ for any five-vectors $\mathbf{u}$ and $\mathbf{v}$, so for any $\mathbf{e} \in \mathcal{E}$ one has $g(\mathbf{u}, \mathbf{e}) = 0$. If one now introduces the notation $g_{AB} \equiv g(\mathbf{e}_A, \mathbf{e}_B)$, then in any standard five-vector basis one will have

$$g_{55} = g_{\alpha 5} = g_{5\alpha} = 0$$

and

$$g_{\alpha\beta}^{(\text{with five-vector indices})} = g_{\alpha\beta}^{(\text{with four-vector indices})}.$$

From these formulae and equations (47) and (48) of subsection A it follows that in any standard five-vector basis

$$\partial_\mu g_{AB} - g_{CB} G^C_{\ A\mu} - g_{AC} G^C_{\ B\mu} = 0,$$

which means that g *regarded as a five-tensor* satisfies the equation $\nabla g = 0$.

The latter equation and formula (60) enable one to obtain the following expression for the first covariant derivative of the inner product h regarded as a five-tensor:

$$\{\nabla_{\mathbf{u}} h\}(\mathbf{v}, \mathbf{w}) = \xi \, \{\nabla_{\mathbf{u}} \lambda\}_{\mathbf{v}} \lambda_{\mathbf{w}} + \xi \, \lambda_{\mathbf{w}} \{\nabla_{\mathbf{u}} \lambda\}_{\mathbf{v}},$$

where $\{\nabla_{\mathbf{u}} \lambda\}_{\mathbf{v}} \equiv \partial_{\mathbf{u}} \lambda_{\mathbf{v}} - \lambda_{(\nabla_{\mathbf{u}} \mathbf{v})}$. Comparing this expression with equation (58), one can see that the latter is equivalent to the following simpler equation:

$$\{\nabla_{\mathbf{u}} \lambda\}_{\mathbf{v}} = g(\mathbf{u}, \mathbf{v}). \tag{61}$$

E. *Forms associated with five-vectors*

As in the case of any other type of vectors, one can consider linear forms corresponding to five-vectors. Such forms will be denoted with lower-case boldface Roman letters with a tilde: $\widetilde{\mathbf{a}}$, $\widetilde{\mathbf{b}}$, $\widetilde{\mathbf{c}}$, etc., and their space will be denoted as $\widetilde{V}_5$. To distinguish a p-form associated with five-vectors from a p-form associated with four-vectors I will call the former a *five-vector* p-form and the latter a *four-vector* p-form.

Five-vector 1-forms have all the properties common to linear forms in general. In addition, they have several specific features, which are due to their association with five-vectors, and it is these latter properties I will now consider.

The existence of two distinguished subspaces in V_5 results in the existence of two distinguished subspaces in $\widetilde{V}_5$. The first of these subspaces is made up by all those 1-forms from $\widetilde{V}_5$ whose contraction with any five-vector from $\mathcal{E}$ is zero. It is evident that this subspace is four-dimensional, and I will denote it as $\widetilde{\mathcal{Z}}$. The other distinguished subspace is made up by all those 1-forms that have a zero contraction with any five-vector from $\mathcal{Z}$. This subspace is one-dimensional, and I will denote it as $\widetilde{\mathcal{E}}$. It is easy to see that $\widetilde{\mathcal{Z}}$ and $\widetilde{\mathcal{E}}$ have only one common element—the zero 1-form, and that $\widetilde{V}_5$ is the direct sum of $\widetilde{\mathcal{Z}}$ and $\widetilde{\mathcal{E}}$. The components of an arbitrary five-vector 1-form $\widetilde{\mathbf{w}}$ in these two subspaces will be denoted as $\widetilde{\mathbf{w}}^{\widetilde{\mathcal{Z}}}$ and $\widetilde{\mathbf{w}}^{\widetilde{\mathcal{E}}}$, respectively.

If $\mathbf{e}_A$ is a standard five-vector basis and $\widetilde{\mathbf{o}}^A$ is the corresponding dual basis of five-vector 1-forms, then $\widetilde{\mathbf{o}}^\alpha \in \widetilde{\mathcal{Z}}$ for all α. The fifth basis 1-form will not necessarily be an element of $\widetilde{\mathcal{E}}$: this will be the case only if all $\mathbf{e}_\alpha \in \mathcal{Z}$. The same conclusions follow from the transformation formulae for the dual basis of 1-forms, corresponding to the transformation $\mathbf{e}'_A = \mathbf{e}_B L^B_A$ from one standard five-vector basis to another. Since in this case $(L^{-1})^\alpha_5 = 0$, one has

$$\widetilde{\mathbf{o}}'^{\,\alpha} = (L^{-1})^\alpha_{\ B}\, \widetilde{\mathbf{o}}^B = (L^{-1})^\alpha_{\ \beta}\, \widetilde{\mathbf{o}}^\beta,$$

but

$$\widetilde{\mathbf{o}}'^{\,5} = (L^{-1})^5_{\ 5}\, \widetilde{\mathbf{o}}^5 + (L^{-1})^5_{\ \beta}\, \widetilde{\mathbf{o}}^\beta.$$

If $\mathbf{e}_A$ is a passive regular basis, then $\widetilde{\mathbf{o}}^5 \in \widetilde{\mathcal{E}}$ and $< \widetilde{\mathbf{o}}^5, \mathbf{1} > = 1$. This particular five-vector 1-form will be denoted as $\widetilde{\mathbf{J}}$.

The fact that $\mathcal{Z}$ is isomorphic to V_4 enables one to establish a natural isomorphism between $\widetilde{\mathcal{Z}}$ and the space of four-vector 1-forms, which will be denoted as $\widetilde{V}_4$. Namely, to each five-vector 1-form $\widetilde{\mathbf{w}}$ from $\widetilde{\mathcal{Z}}$ one can put into correspondence such a four-vector 1-form $\widetilde{\mathbf{W}}$ that for any five-vector $\mathbf{u} \in \mathcal{Z}$ one will have $< \widetilde{\mathbf{w}}, \mathbf{u} > = < \widetilde{\mathbf{W}}, \mathbf{U} >$, where $\mathbf{u} \in \mathbf{U}$. It is evident that this isomorphism can be extended to a map of $\widetilde{V}_5$ onto $\widetilde{V}_4$, which will be a homomorphism but will not be a one-to-one correspondence. In the standard way, this latter map defines an equivalence relation on $\widetilde{V}_5$:

$$\widetilde{\mathbf{u}} \equiv \widetilde{\mathbf{v}} \ \text{ iff their images in } \widetilde{V}_4 \text{ are equal.} \qquad (62)$$

This enables one to regard $\widetilde{V}_4$ as a quotient set and four-vector 1-forms as equivalence classes. It is not difficult to see that the equality of the images of $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{v}}$ in $\widetilde{V}_4$ is equivalent to $\widetilde{\mathbf{u}} - \widetilde{\mathbf{v}} \in \widetilde{\mathcal{E}}$. The relation between $\widetilde{V}_4$ and $\widetilde{V}_5$ is thus similar to the relation between V_4 and V_5, however, unlike the latter, it is not preserved by parallel transport, as it will be shown below.

The parallel transport of five-vector 1-forms is defined in the standard way: by requiring that it conserve the contraction. Consequently, if $G^A_{B\mu}$ are connection coefficients for a standard five-vector basis, then for the corresponding dual basis of 1-forms one has

$$\nabla_\mu \widetilde{\mathbf{o}}^A = -\, G^A_{B\mu} \widetilde{\mathbf{o}}^B, \qquad (63)$$

and from formulae (47) and (48) one obtains that

$$\nabla_\mu \widetilde{\mathbf{o}}^\alpha = -\, G^\alpha_{B\mu} \widetilde{\mathbf{o}}^B = -\, G^\alpha_{\beta\mu} \widetilde{\mathbf{o}}^\beta = -\, \Gamma^\alpha_{\beta\mu} \widetilde{\mathbf{o}}^\beta.$$

This means that 1-forms from $\widetilde{\mathcal{Z}}$ are transported into 1-forms from $\widetilde{\mathcal{Z}}$ and that the isomorphism between $\widetilde{\mathcal{Z}}$ and $\widetilde{V}_4$ is preserved by parallel transport. From formula (63) it also follows that

$$\nabla_\mu \widetilde{\mathbf{o}}^5 = -\, G^5_{5\mu} \widetilde{\mathbf{o}}^5 - G^5_{\beta\mu} \widetilde{\mathbf{o}}^\beta,$$

which shows that in the general case, 1-forms from $\widetilde{\mathcal{E}}$ are not transported into 1-forms from $\widetilde{\mathcal{E}}$, so equivalence relation (62) is not invariant under parallel transport.

As in the case of any other vector space, each inner product on V_5 defines a certain correspondence between five-vectors and five-vector 1-forms. Since one has two inner products on V_5 — g and h, there are two such correspondences, which I will denote as ϑ_g and ϑ_h, respectively. By definition, $\vartheta_g(\mathbf{u})$ is such a five-vector 1-form that

$$< \vartheta_g(\mathbf{u}), \mathbf{v} > = g\,(\mathbf{u}, \mathbf{v}) \quad \text{for any} \quad \mathbf{v} \in V_5. \qquad (64)$$

The definition of the 1-form $\vartheta_h(\mathbf{u})$ is similar. It is evident that both ϑ_g and ϑ_h are linear maps of V_5 into $\widetilde{V}_5$. If u^A are components of some five-vector $\mathbf{u}$ in a certain five-vector basis, then the components of $\vartheta_g(\mathbf{u})$ and $\vartheta_h(\mathbf{u})$ in the corresponding dual basis of 1-forms are $g_{AB}u^B$ and $h_{AB}u^B$, respectively. Since the matrix h_{AB} is nondegenerate, this means that ϑ_h is a one-to-one correspondence and is a map of V_5 *onto* $\widetilde{V}_5$. It is also easy to see that $\vartheta_h(\mathcal{Z}) = \widetilde{\mathcal{Z}}$ and $\vartheta_h(\mathcal{E}) = \widetilde{\mathcal{E}}$. By contrast, ϑ_g is neither a one-to-one correspondence nor a surjection. It is evident that $\vartheta_g(\mathbf{u}) = \vartheta_g(\mathbf{u}^{\mathcal{Z}}) = \vartheta_h(\mathbf{u}^{\mathcal{Z}})$, so $\vartheta_g(\mathcal{Z}) = \widetilde{\mathcal{Z}}$, but $\vartheta_g(\mathcal{E}) = \{\widetilde{\mathbf{0}}\}$. Consequently, one can use g_{AB} only to lower five-vector indices. Raising indices with g_{AB} is possible only if one confines oneself to five-vectors from $\mathcal{Z}$ and to 1-forms from $\widetilde{\mathcal{Z}}$.

All this is in agreement with the general theorem that asserts that the following three statements are equivalent: (*i*) the correspondence between vectors and linear forms induced by a given inner product is injective; (*ii*) this correspondence is surjective; (*iii*) the inner product is nondegenerate.

Another general theorem states that the correspondence between vectors and linear forms is invariant under parallel transport if and only if the corresponding inner product is covariantly constant. Since g, as a five-tensor, satisfies the equation $\nabla g = 0$, one has

$$[\vartheta_g(\mathbf{u})]^{\text{transported}} = \vartheta_g(\mathbf{u}^{\text{transported}})$$

for any $\mathbf{u}$. Alternatively, this can be expressed as

$$\nabla_{\mathbf{v}}[\vartheta_g(\mathbf{u})] = \vartheta_g(\nabla_{\mathbf{v}}\mathbf{u})$$

for all $\mathbf{u}$ and $\mathbf{v}$, which means that the lowering of five-vector indices with g_{AB} commutes with covariant differentiation.

As it has been discussed earlier, the nondegenerate inner product h is *not* covariantly constant, and so in the general case, $[\vartheta_h(\mathbf{u})]^{\text{transported}}$ does not coincide with $\vartheta_h(\mathbf{u}^{\text{transported}})$. Consequently, the lowering and raising of five-vector indices with h_{AB} does not commute with covariant differentiation, and one should take special care whenever these two operations are performed on the same five-tensor.

In section 5 of part I I have introduced the five-vector 1-form $\widetilde{\mathbf{x}}$, which by definition coincides with the fifth element of the 1-form basis dual to an active regular five-vector basis. Comparing this with the definition of the 1-form $\widetilde{\mathbf{j}}$, one finds that $\widetilde{\mathbf{x}} = \varsigma^{-1} \cdot \widetilde{\mathbf{j}}$. Furthermore, it is easy to see that for any five-vector $\mathbf{v}$,

$$\lambda_{\mathbf{v}} = < \widetilde{\mathbf{x}}, \mathbf{v} > .$$

Substituting this expression for $\lambda_{\mathbf{v}}$ into the definition of $\nabla\lambda$, one finds that

$$\begin{aligned} \{\nabla_{\mathbf{u}}\lambda\}_{\mathbf{v}} &= \partial_{\mathbf{u}} < \widetilde{\mathbf{x}}, \mathbf{v} > - < \widetilde{\mathbf{x}}, \nabla_{\mathbf{u}}\mathbf{v} > \\ &= < \nabla_{\mathbf{u}}\widetilde{\mathbf{x}}, \mathbf{v} > . \end{aligned}$$

Substituting this latter expression and definition (64) into equation (61), one obtains that

$$< \nabla_{\mathbf{u}}\widetilde{\mathbf{x}} - \vartheta_g(\mathbf{u}) , \mathbf{v} > = 0$$

for any five-vector $\mathbf{v}$, which means that

$$\nabla_{\mathbf{u}}\widetilde{\mathbf{x}} = \vartheta_g(\mathbf{u}) \tag{65}$$

for any $\mathbf{u}$, which is nothing but equation (38) of part I. In equation (65) the 1-form $\vartheta_g(\mathbf{u})$ can be presented as a contraction of g regarded as a five-tensor of rank $(0, 2)$, with the five-vector $\mathbf{u}$. Considering also that $\nabla_{\mathbf{u}}\widetilde{\mathbf{x}} = < \nabla\widetilde{\mathbf{x}}, \mathbf{u} >$, one can present equation (65) as

$$\nabla\widetilde{\mathbf{x}} = g.$$

Let me finally say a few words about five-vector p-forms with p other than 1. It is a simple matter to see that any five-vector p-form $\widetilde{\mathbf{s}}$ with $p > 1$ can be uniquely presented as a sum of two terms: (i) a p-form made only of 1-forms from $\widetilde{\mathcal{Z}}$ and (ii) a wedge product of the type $\widetilde{\mathbf{t}} \wedge \widetilde{\mathbf{j}}$, where $\widetilde{\mathbf{t}}$ is a $(p-1)$-form. In the following, these two terms will be referred to as the $\widetilde{\mathcal{Z}}$- and $\widetilde{\mathcal{E}}$-components of $\widetilde{\mathbf{s}}$, respectively, and will be denoted as $\widetilde{\mathbf{s}}^{\widetilde{\mathcal{Z}}}$ and $\widetilde{\mathbf{s}}^{\widetilde{\mathcal{E}}}$. It is easy to see that at $p = 1$ this definition agrees with the definition of the $\widetilde{\mathcal{Z}}$- and $\widetilde{\mathcal{E}}$-components of a 1-form given above. It is obvious that a five-vector 5-form has only the $\widetilde{\mathcal{E}}$-component, and it is convenient to take that for any 0-form f,

$$f^{\widetilde{\mathcal{Z}}} = f \text{ and } f^{\widetilde{\mathcal{E}}} = 0.$$

The application of five-vector forms in exterior differential calculus will be discussed in detail in part IV.

One-sided Lévy stable distributions

Jung Hun Han [1]

Abstract

In this paper, we show new representations of one-sided Lévy stable distributions for irrational Lévy indices of the type $\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}$ which are not covered in [8] : for rational Lévy indices. Furthermore, other equivalent representations for a distribution of a rational Lévy index is described. We also give a simplest proof for the formulae which cover the cases for rational Lévy indices. Finally we introduce the concepts of Lévy smashing and Lévy-smashed gamma stochastic processes.

1 Introduction and preliminary results

In [8], Penson and Górska obtained a general form of one-sided Lévy stable distributions expressed as a Meijer G-function for rational Lévy indices by putting $v = 0$ and $a = 1$ in the formula 2.2.1.19 in vol. 5 of [10]. In [2], the role of Mathai transformation in the theory of fractional calculus, which connects ordinary integral to fractional integral through their kernels, is described and α-level space is defined. Furthermore it is insisted that fractional integral and derivative preserve the Lévy structure defined in [2]. The Lévy structure is nothing but the integrand $\frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}$ of the H-function representation of the Lévy distribution with α known as the Lévy index which lies between 0 and 1. Hence the case of simple irrational Lévy indices of the type $\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}$ can be handled.

In this paper, we provide formulae for one-sided Lévy distribution of irrational Lévy index for $0 < \alpha < 1$ and some techniques to generate infinitely many new representations of one-sided Lévy distribution. Furthermore, we introduce the concept of *Lévy smashing* as a consequence of Lévy effect on the family of gamma density functions and Lévy-smashed stochastic processes.

We will use the following integral representation of the gamma function:

$$\begin{aligned}\Gamma(z) &= p^z \int_0^\infty t^{z-1} e^{-pt} dt, \quad \Re(p) > 0, \Re(z) > 0 \\ &= \lim_{n\to\infty} \frac{n!\, n^z}{z(z+1)\cdots(z+n)}, \quad z \neq 0, 1, 2, 3, \cdots .\end{aligned}$$

Pochammer symbol is defined as

$$\begin{aligned}(b)_k &= b(b+1)\cdots(b+k-1),\ (b)_0 = 1,\ b \neq 0 \\ &= (a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} \text{ whenever the gammas exist.}\end{aligned}$$

[1]Corresponding Author Address: Centre for Mathematical Sciences, Pala Campus, Arunapuram P.O., Pala, Kerala-686 574, India, Email : jhan176@yahoo.com, jhan176@gmail.com

For H-function representations and their convergence conditions, consult with [6], [7]:

$$H_{p,q}^{m,n}\left[z\Big|_{(b_1,B_1),((b_2,B_2),\cdots,((b_q,B_q)}^{(a_1,A_1),(a_2,A_2),\cdots,(a_p,A_p)}\right]=\frac{1}{2\pi i}\oint_L\frac{\{\prod_{j=1}^{m}\Gamma(b_j+B_js)\}\{\prod_{j=1}^{n}\Gamma\left(1-a_j-A_js\right)\}}{\{\prod_{j=m+1}^{q}\Gamma\left(1-b_j-B_js\right)\}\{\prod_{j=n+1}^{p}\Gamma\left(a_j+A_js\right)\}}z^{-s}ds.$$

Generalized Wright function, which is well explained in [7] and which is a particular case of the H-function, is

$${}_p\Psi_q\left[z\Big|_{(b_1,B_1),(b_2,B_2),\cdots,(b_q,B_q)}^{(a_1,A_1),(a_2,A_2),\cdots,(a_p,A_p)}\right]=\sum_{n=0}^{\infty}\frac{\prod_{i=1}^{p}\Gamma\left(a_i+A_in\right)}{\prod_{j=1}^{q}\Gamma(b_j+B_jn)\ n!}z^n$$

where $a_i,b_j\in\mathbb{C}$ and $A_i,B_j\in\mathbb{R}$: $A_i\neq 0,B_j\neq 0,i=1,\cdots,p,j=1,\cdots,q$; $\sum_{j=1}^{q}B_j-\sum_{i=1}^{p}A_i>-1$ for absolute convergence.

Hypergeometric series, which is well explained in [6] and which is a particular case of the H-function, is

$${}_pF_q\left(a_1,\cdots,a_p;b_1,\cdots,b_q;z\right)=\sum_{n=0}^{\infty}\frac{\prod_{i=1}^{p}(a_i)}{\prod_{j=1}^{q}(b_j)\ n!}z^n$$

which is absolutely convergent for all z in $\mathbb{C}$ if $p\leq q$.

Definition 1.1 *Lévy jump function is defined as follows*

$$\left[\frac{j}{q}\right]_n=\begin{cases}\frac{j}{q}, & j<n,\ j,n\in\mathbb{Z}^+\\ \frac{j+1}{q}, & j\geq n,\ j,n\in\mathbb{Z}^+.\end{cases}$$

Lemma 1.1 *[Stirling asymptotic formula] [6]*
For $|z|\to\infty$ and α a bounded quantity,

$$\Gamma(z+\alpha)\approx(2\pi)^{1/2}z^{z+\alpha-1/2}e^{-z}. \tag{1.1}$$

Lemma 1.2

$$\begin{aligned}\Gamma\left(-v+\frac{i}{q}-\left[\frac{j}{q}\right]_i\right) &= \frac{(-1)^v\Gamma\left(\frac{i}{q}-\left[\frac{j}{q}\right]_i\right)}{\left(1-\frac{i}{q}+\left[\frac{j}{q}\right]_i\right)_v}\\ \Gamma\left(-v+\frac{iq-jp}{pq}\right) &= \frac{(-1)^v\Gamma\left(\frac{iq-jp}{pq}\right)}{\left(1-\frac{iq-jp}{pq}\right)_v}\end{aligned}$$

and

$$H_{p,q}^{m,n}\left[z\Big|_{(b_q,B_q)}^{(a_p,A_p)}\right]=k\ H_{p,q}^{m,n}\left[z^k\Big|_{(b_q,kB_q)}^{(a_p,kA_p)}\right]$$

where $v,n,p,q,i,j\in\mathbb{Z}^+$ and $k>0$.

Proof. Use the formula in [6], p44, $\Gamma(\beta+1-v)=\frac{(-1)^v\Gamma(\beta+1)}{(-\beta)_v}$ and the properties in [7], which complete the proof.

2 Lévy stable distribution of Lévy index $\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}$

Let

$$f_\alpha(x)=\frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s_1}{\alpha})}{\alpha\Gamma(1-s_1)}x^{-s_1}ds_1,\quad 1>\alpha>Re(s_1)>0. \tag{2.1}$$

$\frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}$ appears firstly in [5] in the literature. Then $f_\alpha(x)$ is the well known Levy density function having the Laplace transform e^{-t^α}.

Theorem 2.1 *Let $0<\alpha<1$ and p,q,l_1,l_2 be arbitrary positive integers such that $0<\alpha=\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}<1,\quad p<q$ and $l=\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}-1}$.*

If l is not a positive integer and $l_1\neq l_2$, then

$$\begin{aligned}
f_{(\alpha,l)}(x)&=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ H^{q,0}_{p,q}\left[\frac{p^{pl}}{x^{pl}q^q}\Big|^{(1,l),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}_{(1,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]\\
&=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\left(\frac{p^{lp}}{x^{lp}q^q}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(-\frac{q-1}{q},-1)\cdots(-\frac{2}{q},-1)(-\frac{1}{q},-1)}_{(1-l,-l)(\frac{q-plq}{pq},-l)\cdots(\frac{(p-1)q-plq}{pq},-l)}\right]+\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{jl\frac{p}{q}}}{x^{jl\frac{p}{q}}q^j}\right)\\
&\times\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{\left(1-\left[\frac{j}{q}\right]_q,-1\right)\left(\frac{2}{q}-\left[\frac{j}{q}\right]_2,-1\right)\left(\frac{3}{q}-\left[\frac{j}{q}\right]_3,-1\right)\cdots\left(\frac{q-1}{q}-\left[\frac{j}{q}\right]_{q-1},-1\right)}_{\left(\frac{pq-jpl}{pq},-l\right)\left(\frac{q-jpl}{pq},-l\right)\cdots\left(\frac{(p-1)q-jpl}{pq},-l\right)}\right].
\end{aligned}\tag{2.2}$$

If $l=1$ and $l_1=l_2$, then

$$\begin{aligned}
f_{(\alpha,1)}(x)&=\frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ H^{q-1,0}_{p-1,q-1}\left[\frac{p^p}{x^pq^q}\Big|^{\left(\frac{1}{p},1\right),\cdots,\left(\frac{p-1}{p},1\right)}_{\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]\\
&=\frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{j\frac{p}{q}}}{x^{j\frac{p}{q}}q^j}\right)\frac{\prod_{i_1=2}^{q-1}\Gamma\left(\frac{i_1}{q}-\left[\frac{j}{q}\right]_{i_1}\right)}{\prod_{i_2=1}^{p-1}\Gamma\left(\frac{i_2q-jp}{pq}\right)}\\
&\times\ {}_{p-1}F_{q-2}\left[(-1)^{q-p}\frac{p^p}{x^pq^q}\Big|^{\left(1-\frac{q-jp}{pq}\right)\cdots\left(1-\frac{(p-1)q-jp}{pq}\right)}_{\left(1-\frac{2}{q}+\left[\frac{j}{q}\right]_2\right)\left(1-\frac{3}{q}+\left[\frac{j}{q}\right]_3\right)\cdots\left(1-\frac{q-1}{q}+\left[\frac{j}{q}\right]_{q-1}\right)}\right].
\end{aligned}\tag{2.3}$$

If l belongs to the set $\{2,3,4,\cdots\}$ and $l_1\neq l_2$, then

$$\begin{aligned}
f_{(\alpha,l)}(x)&=\frac{\sqrt{pql}}{x(2\pi)^{\frac{q+1-p-l}{2}}}\ H^{q-1,0}_{p+l-2,q-1}\left[\frac{p^{pl}l^l}{x^{pl}q^q}\Big|^{\left(\frac{1}{l},1\right),\cdots,\left(\frac{l-1}{l},1\right),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}_{\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]\\
&=\frac{\sqrt{pql}}{x(2\pi)^{\frac{q+1-l-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{jl\frac{p}{q}}l^{j\frac{l}{q}}}{x^{jl\frac{p}{q}}q^j}\right)\\
&\times\ {}_{q-2}\Psi_{p-1+l-1}\left[-\frac{p^{pl}l^l}{x^{pl}q^q}\Big|^{\left(\frac{2}{q}-\left[\frac{j}{q}\right]_2,-1\right)\left(\frac{3}{q}-\left[\frac{j}{q}\right]_3,-1\right)\cdots\left(\frac{q-1}{q}-\left[\frac{j}{q}\right]_{q-1},-1\right)}_{\left(\frac{q-jl}{lq},-1\right)\cdots\left(\frac{(l-1)q-jl}{lq},-1\right)\left(\frac{q-jpl}{pq},-l\right)\cdots\left(\frac{(p-1)q-jpl}{pq},-l\right)}\right].
\end{aligned}\tag{2.4}$$

Proof. Here we use a transformation $1-s_1=\alpha s$, then

$$f_\alpha(x)=\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(\alpha s)}x^{\alpha s-1}ds,\quad \frac{1}{\alpha}>Re(s)>\frac{1-\alpha}{\alpha}>0. \tag{2.5}$$

Now, by using the Gauss-Legendre formula (Multiplicative formula) for gamma function, namely

$$\Gamma(mz) \;=\; (2\pi)^{\frac{1-m}{2}}m^{mz-\frac{1}{2}}\Gamma(z)\Gamma(z+\frac{1}{m})\cdots\Gamma(z+\frac{m-1}{m}),\ m=1,2,\cdots,$$

we have

$$\Gamma(\alpha s) \;=\; \Gamma\left(\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}s\right)=\Gamma\left(p\frac{ls}{q}\right)=(2\pi)^{\frac{1-p}{2}}p^{p\frac{ls}{q}-\frac{1}{2}}\Gamma\left(\frac{ls}{q}\right)\Gamma\left(\frac{ls}{q}+\frac{1}{p}\right)\cdots\Gamma\left(\frac{ls}{q}+\frac{p-1}{p}\right).$$

Apply to the integrand in (2.5), then

$$f_{(\alpha,l)}(x)=\frac{1}{x}\frac{1}{2\pi i}\oint_L\frac{\Gamma(s)\ x^{\frac{p}{q}ls}}{(2\pi)^{\frac{1-p}{2}}p^{p\frac{ls}{q}-\frac{1}{2}}\Gamma(\frac{ls}{q})\Gamma(\frac{ls}{q}+\frac{1}{p})\cdots\Gamma(\frac{ls}{q}+\frac{p-1}{p})}ds$$

where $\alpha=\frac{p}{q}l=\frac{p}{q}\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}-1}=\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}$.
We use

$$H_{p,q}^{m,n}\left[z\Big|_{(b_q,B_q)}^{(a_p,A_p)}\right]=k\ H_{p,q}^{m,n}\left[z^k\Big|_{(b_q,kB_q)}^{(a_p,kA_p)}\right]$$

in Lemma 1.2. Then

$$\begin{aligned}
f_{(\alpha,l)}(x)&=\frac{q}{2\pi ix}\oint_L\frac{\Gamma(qs)\ x^{pls}}{(2\pi)^{\frac{1-p}{2}}p^{pls-\frac{1}{2}}\Gamma(ls)\Gamma(ls+\frac{1}{p})\cdots\Gamma(ls+\frac{p-1}{p})}ds\\
&=\frac{q}{2\pi ix}\oint_L\frac{(2\pi)^{\frac{1-q}{2}}q^{qs-\frac{1}{2}}\Gamma(s)\Gamma(s+\frac{1}{q})\cdots\Gamma(s+\frac{q-1}{q})\ x^{pls}}{(2\pi)^{\frac{1-p}{2}}p^{pls-\frac{1}{2}}\Gamma(ls)\Gamma(ls+\frac{1}{p})\cdots\Gamma(ls+\frac{p-1}{p})}ds\\
&=\frac{1}{2\pi i}\oint_L\frac{\frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ ls\Gamma(s)\Gamma\left(s+\frac{1}{q}\right)\cdots\Gamma\left(s+\frac{q-1}{q}\right)}{ls\Gamma\left(ls\right)\Gamma\left(ls+\frac{1}{p}\right)\cdots\Gamma\left(ls+\frac{p-1}{p}\right)}\left[\frac{p^{pl}}{x^{pl}q^q}\right]^{-s}ds\\
&=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ \frac{1}{2\pi i}\oint_L\frac{\Gamma(s+1)\Gamma\left(s+\frac{1}{q}\right)\cdots\Gamma\left(s+\frac{q-1}{q}\right)}{\Gamma\left(ls+1\right)\Gamma\left(ls+\frac{1}{p}\right)\cdots\Gamma\left(ls+\frac{p-1}{p}\right)}\left[\frac{p^{pl}}{x^{pl}q^q}\right]^{-s}ds\\
&=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}H_{p,q}^{q,0}\left[\frac{p^{pl}}{x^{pl}q^q}\Big|_{(1,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}^{(1,l),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}\right]
\end{aligned}$$

Since we have the following

$$\begin{aligned}
&H_{p,q}^{q,0}\left[\frac{p^{pl}}{x^{pl}q^q}\Big|_{(1,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}^{(1,l),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}\right]\\
&=\sum_{k_0=1}^{\infty}\frac{(-1)^{k_0}\Gamma(-k_0-\frac{q-1}{q})\cdots\Gamma(-k_0-\frac{2}{q})\Gamma(-k_0-\frac{1}{q})}{k_0!\ \Gamma(-lk_0+\frac{pq-plq}{pq})\Gamma(-lk_0+\frac{q-plq}{pq})\cdots\Gamma(-lk_0+\frac{(p-1)q-plq}{pq})}\left(\frac{p^{pl}}{x^{pl}q^q}\right)^{k_0+1}
\end{aligned}$$

$$+\sum_{k_1=0}^{\infty}\frac{(-1)^{k_1}\Gamma(-k_1+\frac{q-1}{q})\Gamma(-k_1+\frac{1}{q})\cdots\Gamma(-k_1+\frac{q-2}{q})}{k_1!\,\Gamma(-lk_1+\frac{pq-pl}{pq})\Gamma(-lk_1+\frac{q-pl}{pq})\cdots\Gamma(-lk_1+\frac{(p-1)q-pl}{pq})}\left(\frac{p^{pl}}{x^{pl}q^q}\right)^{k_1+\frac{1}{q}}$$
$$+\sum_{k_2=0}^{\infty}\frac{(-1)^{k_2}\Gamma(-k_2+\frac{q-2}{q})\Gamma(-k_2-\frac{1}{q})\Gamma(-k_2+\frac{1}{q})\cdots\Gamma(-k_2+\frac{q-3}{q})}{k_2!\,\Gamma(-lk_2+\frac{pq-2pl}{pq})\Gamma(-lk_2+\frac{q-2pl}{pq})\cdots\Gamma(-lk_2+\frac{(p-1)q-2pl}{pq})}\left(\frac{p^{pl}}{x^{pl}q^q}\right)^{k_2+\frac{2}{q}}$$
$$+\cdots$$
$$+\sum_{k_{q-1}=0}^{\infty}\frac{(-1)^{k_{q-1}}\Gamma(-k_{q-1}+\frac{1}{q})\Gamma(-k_{q-1}-\frac{q-2}{q})\cdots\Gamma(-k_{q-1}-\frac{1}{q})}{k_{q-1}!\,\Gamma(-lk_{q-1}+\frac{pq-(q-1)pl}{pq})\Gamma(-lk_{q-1}+\frac{q-(q-1)pl}{pq})\cdots\Gamma(-lk_{q-1}+\frac{(p-1)q-(q-1)pl}{pq})}$$
$$\times\left(\frac{p^{pl}}{x^{pl}q^q}\right)^{k_{q-1}+\frac{q-1}{q}}=\left(\frac{p^{lp}}{x^{lp}q^q}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(-\frac{q-1}{q},-1)\cdots(-\frac{2}{q},-1)(-\frac{1}{q},-1)}_{(1-l,-l)(\frac{q-plq}{pq},-l)\cdots(\frac{(p-1)q-plq}{pq},-l)}\right]$$
$$+\left(\frac{p^{l\frac{p}{q}}}{x^{l\frac{p}{q}}q}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(\frac{q-1}{q},-1)(\frac{1}{q},-1)\cdots(\frac{q-2}{q},-1)}_{(\frac{pq-pl}{p},-l)(\frac{q-pl}{pq},-l)\cdots(\frac{(p-1)q-pl}{p},-l)}\right]$$
$$+\left(\frac{p^{2l\frac{p}{q}}}{x^{2l\frac{p}{q}}q^2}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(-\frac{1}{q},-1)(\frac{1}{q},-1)\cdots(\frac{q-3}{q},-1)(\frac{q-2}{q},-1)}_{(\frac{q-2pl}{pq},-l)\cdots(\frac{(p-1)q-2pl}{p},-l)(\frac{pq-2pl}{p},-l)}\right]$$
$$+\cdots$$
$$+\left(\frac{p^{(q-1)l\frac{p}{q}}}{x^{(q-1)l\frac{p}{q}}q^{(q-1)}}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(\frac{1}{q},-1)(-\frac{(q-2)}{q},-1)(\frac{(q-3)}{q},-1)\cdots(-\frac{1}{q},-1)}_{(\frac{pq-(q-1)pl}{p},-l)(\frac{q-(q-1)pl}{pq},-l)\cdots(\frac{(p-1)q-(q-1)pl}{p},-l)}\right],$$

we get

$$f_{(\alpha,l)}(x)=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ H_{p,q}^{q,0}\left[\frac{p^{pl}}{x^{pl}q^q}\Big|^{(1,l),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}_{(1,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]$$
$$=\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\left(\frac{p^{lp}}{x^{lp}q^q}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{(-\frac{q-1}{q},-1)\cdots(-\frac{2}{q},-1)(-\frac{1}{q},-1)}_{(1-l,-l)(\frac{q-plq}{pq},-l)\cdots(\frac{(p-1)q-plq}{pq},-l)}\right]$$
$$+\frac{l\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{jl\frac{p}{q}}}{x^{jl\frac{p}{q}}q^j}\right)\ {}_{q-1}\Psi_p\left[-\frac{p^{pl}}{x^{pl}q^q}\Big|^{\left(1-\left[\frac{j}{q}\right]_q,-1\right)\left(\frac{2}{q}-\left[\frac{j}{q}\right]_2,-1\right)\left(\frac{3}{q}-\left[\frac{j}{q}\right]_3,-1\right)\cdots\left(\frac{q-1}{q}-\left[\frac{j}{q}\right]_{q-1},-1\right)}_{\left(\frac{pq-jpl}{pq},-l\right)\left(\frac{q-jpl}{pq},-l\right)\cdots\left(\frac{(p-1)q-jpl}{pq},-l\right)}\right]$$

where $\left[\frac{j}{q}\right]_n$ is the Lévy jump function.
Note that the series are absolutely convergent satisfying the condition $\sum_{j=1}^{q}B_j-\sum_{i=1}^{p}A_i>-1\Rightarrow$ $-lp+q-1>-1\Rightarrow q>lp\Rightarrow\frac{q}{p}>l\Rightarrow\frac{q}{p}>\frac{p}{q}\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}-1}\Rightarrow\frac{q}{p}>1>\left(\frac{p}{q}\right)^{\frac{l_2}{l_1}}=\alpha$ since $\alpha<1$.

If l belongs to the set $\{2,3,4,\cdots\}$, then we have

$$f_{(\alpha,l)}(x)=\frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}H_{p,q}^{q,0}\left[\frac{p^{pl}}{x^{pl}q^q}\Big|^{(0,l),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}_{(0,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]\tag{2.6}$$
$$=\frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\frac{1}{2\pi i}\oint_L\frac{\Gamma(s)\Gamma\left(s+\frac{1}{q}\right)\cdots\Gamma\left(s+\frac{q-1}{q}\right)}{\Gamma(ls)\,\Gamma\left(ls+\frac{1}{p}\right)\cdots\Gamma\left(ls+\frac{p-1}{p}\right)}\left[\frac{p^{pl}}{x^{pl}q^q}\right]^{-s}ds$$

We use

$$\Gamma(ls)=(2\pi)^{\frac{1-l}{2}}l^{ls-\frac{1}{2}}\Gamma(s)\Gamma\left(s+\frac{1}{l}\right)\cdots\Gamma\left(s+\frac{l-1}{l}\right).$$

Hence we have

$$= \frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)\Gamma\left(s+\frac{1}{q}\right)\cdots\Gamma\left(s+\frac{q-1}{q}\right)}{(2\pi)^{\frac{1-l}{2}}l^{ls-\frac{1}{2}}\Gamma(s)\Gamma\left(s+\frac{1}{l}\right)\cdots\Gamma\left(s+\frac{l-1}{l}\right)\Gamma\left(ls+\frac{1}{p}\right)\cdots\Gamma\left(ls+\frac{p-1}{p}\right)}\left[\frac{p^{pl}}{x^{pl}q^q}\right]^{-s}ds$$

$$= \frac{\sqrt{pql}}{x(2\pi)^{\frac{q+1-p-l}{2}}}\frac{1}{2\pi i}\oint_L \frac{\Gamma\left(s+\frac{1}{q}\right)\cdots\Gamma\left(s+\frac{q-1}{q}\right)}{\Gamma\left(s+\frac{1}{l}\right)\cdots\Gamma\left(s+\frac{l-1}{l}\right)\Gamma\left(ls+\frac{1}{p}\right)\cdots\Gamma\left(ls+\frac{p-1}{p}\right)}\left[\frac{p^{pl}l^l}{x^{pl}q^q}\right]^{-s}ds$$

$$= \frac{\sqrt{pql}}{x(2\pi)^{\frac{q+1-p-l}{2}}}H^{q-1,0}_{p+l-2,q-1}\left[\frac{p^{pl}l^l}{x^{pl}q^q}\Big|^{\left(\frac{1}{l},1\right),\cdots,\left(\frac{l-1}{l},1\right),\left(\frac{1}{p},l\right),\cdots,\left(\frac{p-1}{p},l\right)}_{\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]$$

Note that for $l \in \{2,3,4,\cdots\}$, q should be of the form kp, $k = 4,5,6,\cdots$ and $l_2 < l_1$. $\sum_{j=1}^{q} B_j - \sum_{i=1}^{p} A_i > 0$ means $q-1-(l-1)-l(p-1) > 0 \Rightarrow q > lp \Rightarrow kp > lp \Rightarrow k > l$. But this condition is always satisfied since $l = (k)^{1-\frac{l_2}{l_1}}$.

If we put $l = 1$ in (2.6), then

$$f_{(\alpha,1)}(x) = \frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ H^{q,0}_{p,q}\left[\frac{p^p}{x^pq^q}\Big|^{(0,1),\left(\frac{1}{p},1\right),\cdots,\left(\frac{p-1}{p},1\right)}_{(0,1),\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right] = \frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\ H^{q-1,0}_{p-1,q-1}\left[\frac{p^p}{x^pq^q}\Big|^{\left(\frac{1}{p},1\right),\cdots,\left(\frac{p-1}{p},1\right)}_{\left(\frac{1}{q},1\right),\cdots,\left(\frac{q-1}{q},1\right)}\right]$$

$$= \frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{j\frac{p}{q}}}{x^{j\frac{p}{q}}q^j}\right)\ {}_{q-2}\Psi_{p-1}\left[-\frac{p^p}{x^pq^q}\Big|^{\left(\frac{2}{q}-\left[\frac{j}{q}\right]_2,-1\right)\left(\frac{3}{q}-\left[\frac{j}{q}\right]_3,-1\right)\cdots\left(\frac{q-1}{q}-\left[\frac{j}{q}\right]_{q-1},-1\right)}_{\left(\frac{q-jp}{pq},-1\right)\cdots\left(\frac{(p-1)q-jp}{pq},-1\right)}\right]$$

by applying the formula $\Gamma(\beta+1-v) = \frac{(-1)^v\Gamma(\beta+1)}{(1-(\beta+1))_v}$ in Lemma 1.1

$$= \frac{\sqrt{pq}}{x(2\pi)^{\frac{q-p}{2}}}\sum_{j=1}^{q-1}\left(\frac{p^{j\frac{p}{q}}}{x^{j\frac{p}{q}}q^j}\right)\frac{\prod_{i_1=2}^{q-1}\Gamma\left(\frac{i_1}{q}-\left[\frac{j}{q}\right]_{i_1}\right)}{\prod_{i_2=1}^{p-1}\Gamma\left(\frac{i_2q-jp}{pq}\right)}$$

$$\times\ {}_{p-1}F_{q-2}\left[(-1)^{q-p}\frac{p^p}{x^pq^q}\Big|^{\left(1-\frac{q-jp}{pq}\right)\cdots\left(1-\frac{(p-1)q-jp}{pq}\right)}_{\left(1-\frac{2}{q}+\left[\frac{j}{q}\right]_2\right)\left(1-\frac{3}{q}+\left[\frac{j}{q}\right]_3\right)\cdots\left(1-\frac{q-1}{q}+\left[\frac{j}{q}\right]_{q-1}\right)}\right]$$

For the case of $l = 1$, the condition for their convergence is trivial since $p < q$.

Some special cases will be given. We will use (2.3) to show some known results.

Example 2.1 *When $p = 1,\ q = 4,\ \alpha = \frac{1}{4}$, we have*

$$f_{\left(\frac{1}{4},1\right)}(x) = \frac{2}{x(2\pi)^{\frac{3}{2}}}\ H^{3,0}_{0,3}\left[\frac{1}{4^4x}\Big|^{-}_{\left(\frac{1}{4},1\right),\left(\frac{2}{4},1\right),\left(\frac{3}{4},1\right)}\right]$$

$$= \frac{2}{x(2\pi)^{\frac{3}{2}}}\sum_{j=1}^{3}\left(\frac{1}{x^{j\frac{1}{4}}4^j}\right)\frac{\prod_{i_1=2}^{3}\Gamma\left(\frac{i_1}{4}-\left[\frac{j}{4}\right]_{i_1}\right)}{\prod_{i_2=1}^{0}\Gamma\left(\frac{i_24-j}{4}\right)}\ {}_0F_2\left[(-1)^3\frac{1}{4^4x}\Big|^{-}_{\left(1-\frac{2}{4}+\left[\frac{j}{4}\right]_2\right)\left(1-\frac{3}{4}+\left[\frac{j}{4}\right]_3\right)}\right]$$

$$= \frac{1}{2x^{\frac{5}{4}}(2\pi)^{\frac{3}{2}}}\Gamma\left(\frac{1}{4}\right)\Gamma\left(\frac{1}{2}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)}\right] + \frac{1}{8x^{\frac{3}{2}}(2\pi)^{\frac{3}{2}}}\Gamma\left(-\frac{1}{4}\right)\Gamma\left(\frac{1}{4}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{\left(\frac{3}{4}\right)\left(\frac{5}{4}\right)}\right]$$

$$+\frac{1}{32x^{\frac{7}{4}}(2\pi)^{\frac{3}{2}}}\Gamma\left(-\frac{1}{4}\right)\Gamma\left(-\frac{1}{2}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{(\frac{5}{4})(\frac{3}{2})}\right]$$
$$=\frac{2}{(2\pi)^{\frac{3}{2}}x}\{\frac{2}{x^{\frac{1}{4}}}\Gamma\left(1+\frac{1}{4}\right)\Gamma\left(1+\frac{2}{4}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{(\frac{3}{4})(\frac{2}{4})}\right]$$
$$-\frac{1}{x^{\frac{1}{2}}}\Gamma\left(\frac{3}{4}\right)\Gamma\left(1+\frac{1}{4}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{(\frac{3}{4})(1+\frac{1}{4})}\right]+\frac{1}{8x^{\frac{3}{4}}}\Gamma\left(\frac{2}{4}\right)\Gamma\left(\frac{3}{4}\right)\ {}_0F_2\left[\frac{-1}{256x}\Big|^{-}_{(1+\frac{1}{4})(1+\frac{2}{4})}\right]\}.$$

We will use (2.2) to show some new results.

Example 2.2 *Let* $p=1,\ q=2,\ l_1=2,\ l_2=1,\ \alpha=\frac{1}{\sqrt{2}},\ l=\sqrt{2}$, *then*

$$f_{(\frac{1}{\sqrt{2}},\sqrt{2})}(x)=\frac{\sqrt{2}}{x\sqrt{\pi}}\ H^{2,0}_{1,2}\left[\frac{1}{4x^{\sqrt{2}}}\Big|^{(1,\sqrt{2})}_{(1,1),(\frac{1}{2},1)}\right]$$
$$=\frac{\sqrt{2}}{x\sqrt{\pi}}\left(\frac{1}{4x^{\sqrt{2}}}\ {}_1\Psi_1\left[-\frac{1}{4x^{\sqrt{2}}}\Big|^{(-\frac{1}{2},-1)}_{(1-\sqrt{2},-\sqrt{2})}\right]+\frac{1}{2x^{\frac{1}{\sqrt{2}}}}\ {}_1\Psi_1\left[-\frac{1}{4x^{\sqrt{2}}}\Big|^{(\frac{1}{2},-1)}_{(1-\frac{1}{\sqrt{2}},-\sqrt{2})}\right]\right).$$

3 A process to generate more representations

For the same rational α, more representations of one distribution can be generated by using (2.3) and (2.4).

Example 3.1 *When* $p=1,\ q=2,\ \alpha=\frac{1}{2}$ *in* (2.3), *we have*

$$f_{(\frac{1}{2},1)}(x)=\frac{1}{x\sqrt{\pi}}\ H^{1,0}_{0,1}\left[\frac{1}{4x}\Big|^{-}_{(\frac{1}{2},1)}\right]=\frac{1}{2x^{\frac{3}{2}}\sqrt{\pi}}\ {}_0F_0\left[\frac{-1}{4x}\Big|^{-}_{-}\right]=\frac{1}{2\sqrt{\pi}x^{\frac{3}{2}}}exp\left(\frac{-1}{4x}\right)$$

Example 3.2 *Let* $p=1,\ q=4,\ l_1=2,\ l_2=1,\ \alpha=\frac{1}{4}^{\frac{1}{2}}=\frac{1}{2},\ l=2$ *in* (2.4), *then we have*

$$f_{(\frac{1}{2},2)}(x)=\frac{\sqrt{8}}{2\pi x}H^{3,0}_{1,3}\left[\frac{4}{x^2 4^4}\Big|^{(\frac{1}{2},1)}_{(\frac{1}{4},1),(\frac{2}{4},1),(\frac{3}{4},1)}\right]=\frac{1}{2\pi x^{\frac{3}{2}}}\sum_{v=0}^{\infty}\frac{\Gamma(-v+\frac{1}{2})}{v!}\left(\frac{-1}{4^3x^2}\right)^v$$
$$+\frac{1}{2^4\pi x^{\frac{5}{2}}}\sum_{v=0}^{\infty}\frac{\Gamma(-v-\frac{1}{2})}{v!}\left(\frac{-1}{4^3x^2}\right)^v=\frac{1}{2\sqrt{\pi}x^{\frac{3}{2}}}\left(\ {}_0F_1\left[\frac{1}{4^3x^2}\Big|^{-}_{(\frac{1}{2})}\right]-\frac{1}{4x}\ {}_0F_1\left[\frac{1}{4^3x^2}\Big|^{-}_{(\frac{3}{2})}\right]\right)$$

Note that $exp\left(\frac{-1}{4x}\right)=\ {}_0F_1\left[\frac{1}{4^3x^2}\Big|^{-}_{(\frac{1}{2})}\right]-\frac{1}{4x}\ {}_0F_1\left[\frac{1}{4^3x^2}\Big|^{-}_{(\frac{3}{2})}\right]$ and by setting $\alpha=\left(\frac{2}{8}\right)^{\frac{1}{2}}=\frac{1}{2}$ in (2.4), another representation can be born in the sum of faster convergent series.

4 Lévy smashing on the family of gamma density functions and Lévy-smashed gamma stochastic process

Mellin transform of a density function in statistics shows its statistical structure and this technique can be used as a tool to blend two independently distributed random variables. In this section,

we show what Lévy effect could be and how we should understand it. To start with, consider the Lévy density function $\frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}x^{-s}ds$ and the gamma density functions $\frac{x^{\gamma-1}}{\Gamma(\gamma)}e^{-x}$ where $0<\alpha<1,\ 0<\gamma$ and $0<x<\infty$. $\frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}x^{-s}ds$ is the one-sided Lévy density function found in [5] constructed in a different way in [2]. We will use the Mellin transformation of the form $\int_t h_1(\frac{x}{t})h_2(t)\frac{dt}{t}$, where $h_1(x)$ and $h_2(x)$ are certain density functions. Then we have

$$
\begin{aligned}
f_{(\alpha)}(x) &= \int_0^\infty \frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}x^{-s}t^s ds\ \frac{t^{\gamma-1}}{\Gamma(\gamma)}e^{-t}\frac{dt}{t} = \frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}x^{-s}\int_0^\infty \frac{t^{s+\gamma-1-1}}{\Gamma(\gamma)}e^{-t}dtds \\
&= \frac{1}{2\pi i}\oint_L \frac{\Gamma(\frac{1}{\alpha}-\frac{s}{\alpha})}{\alpha\Gamma(1-s)}\frac{\Gamma(s+\gamma-1)}{\Gamma(\gamma)}x^{-s}ds \ \text{ by transformation } s=1-\alpha s_1, \\
&= \frac{1}{2\pi i}\oint_L \frac{\Gamma(s_1)}{\Gamma(\alpha s_1)}\frac{\Gamma(\gamma-\alpha s_1)}{\Gamma(\gamma)}x^{\alpha s_1-1}ds_1 \qquad (4.1)
\end{aligned}
$$

and its Laplace transform

$$
\begin{aligned}
L_f(y) &= \int_0^\infty e^{-yx}\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(\alpha s)}\frac{\Gamma(\gamma-\alpha s)}{\Gamma(\gamma)}x^{\alpha s-1}dsdx = \frac{1}{2\pi i}\oint_L \frac{\Gamma(s)\Gamma(\gamma-\alpha s)}{\Gamma(\gamma)}y^{-\alpha s}ds \\
&= \sum_{k=0}^\infty \frac{(-1)^k\Gamma(\gamma+\alpha k)}{k!\,\Gamma(\gamma)}(y)^{\alpha k}. \qquad (4.2)
\end{aligned}
$$

When $\alpha=1$ in (4.1), (4.1) becomes gamma density $\frac{x^{\gamma-1}}{\Gamma(\gamma)}e^{-x}$. So $f_{(1)}(x)$ is a one-sided function concentrated at $x=1$ for $\mathbb{R}^+$. To understand its effect, put $\alpha=\frac{1}{2}$ in (4.1), then we have

$$
\begin{aligned}
&\frac{1}{2\pi i}\oint_L \frac{\Gamma(s_1)}{\Gamma(\frac{1}{2}s_1)}\frac{\Gamma(\gamma-\frac{1}{2}s_1)}{\Gamma(\gamma)}x^{\frac{1}{2}s_1-1}ds_1, s_1=2s \qquad (4.3)\\
&= \frac{1}{2\pi i}\oint_L \frac{\Gamma(2s)}{\Gamma(s)}\frac{\Gamma(\gamma-s)}{\Gamma(\gamma)}x^{s-1}2ds = \frac{1}{2\pi i}\oint_L \frac{2^{2s-1}\Gamma(s)\Gamma(s+\frac{1}{2})}{\sqrt{\pi}\Gamma(s)}\frac{\Gamma(\gamma-s)}{\Gamma(\gamma)}x^{s-1}2ds \qquad (4.4)\\
&= \sum_{k=0}^\infty \frac{(-1)^k\Gamma(\gamma+\frac{1}{2}+k)}{k!\sqrt{\pi}\Gamma(\gamma)}x^{-k-\frac{3}{2}}4^{-k-\frac{1}{2}} = \frac{\Gamma(\gamma+\frac{1}{2})}{2\sqrt{\pi}\Gamma(\gamma)x^{\frac{3}{2}}}\left(1+\frac{1}{4x}\right)^{-\gamma-\frac{1}{2}} \qquad (4.5)\\
&= \frac{4\Gamma(\gamma+\frac{1}{2})(4x)^{\gamma-1}}{\sqrt{\pi}\Gamma(\gamma)}(1+4x)^{-\gamma-\frac{1}{2}} \qquad (4.6)
\end{aligned}
$$

Fig. 1 shows the impact on the family of gamma functions.

	gamma family		Lévy smashed gamma family
(a) $\gamma=1$	e^{-x}	$\leftrightarrow$	$2(1+4x)^{-3/2}$
(b) $\gamma=2$	xe^{-x}	$\leftrightarrow$	$12x(1+4x)^{-5/2}$
(c) $\gamma=3$	$\frac{1}{2}x^2e^{-x}$	$\leftrightarrow$	$60x^2(1+4x)^{-7/2}$
(d) $\gamma=4$	$\frac{1}{3!}x^3e^{-x}$	$\leftrightarrow$	$280x^3(1+4x)^{-9/2}$

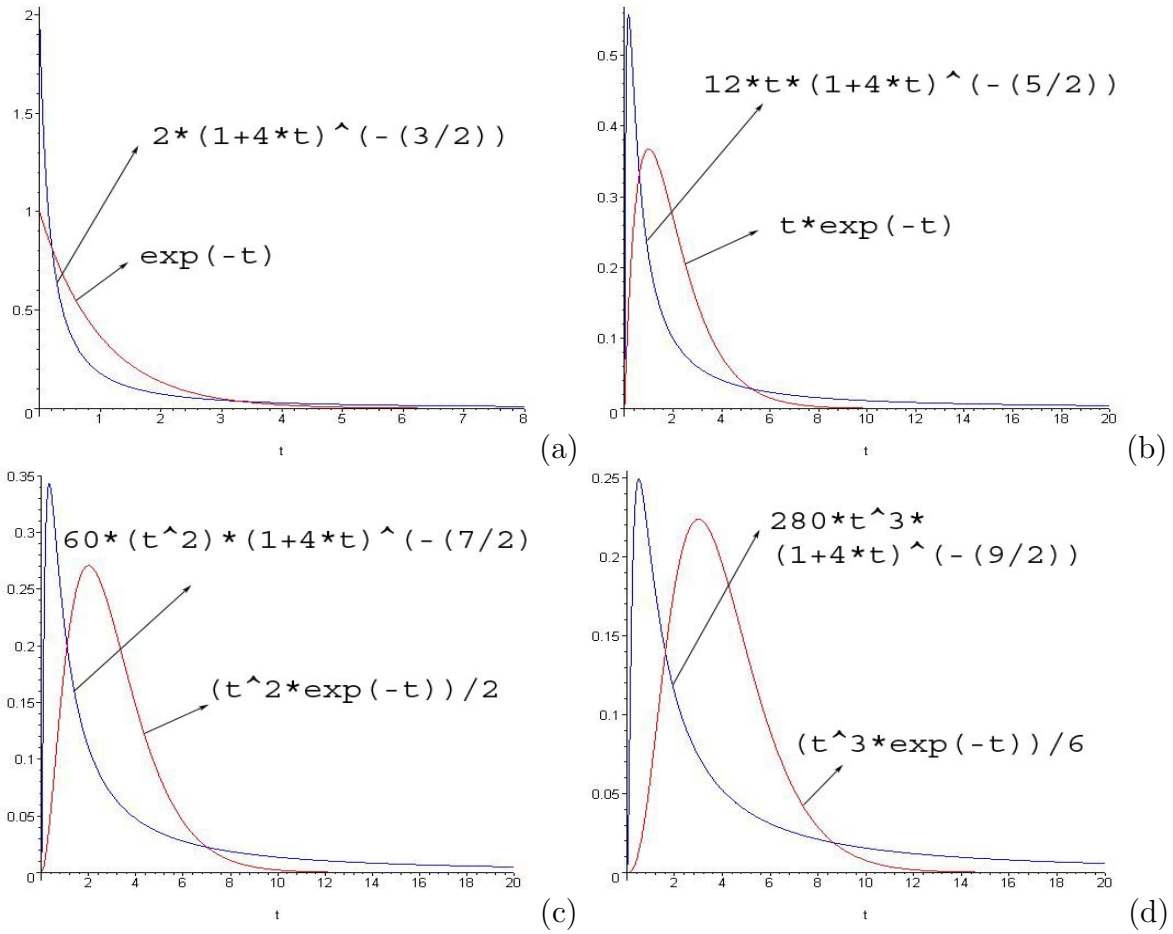


Figure 1.

$f_{(\alpha)}(x) = \frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(\alpha s)}\frac{\Gamma(\gamma-\alpha s)}{\Gamma(\gamma)}x^{\alpha s-1}ds$ is absolutely convergent series for all x since $\mu = \alpha-\alpha+1 = 1 > 0$, see [7]. From the fig. 1, $f_{(\alpha)}(x)$ will be called as Lévy-smashed gamma density functions especially when the parameter $0 < \alpha < 1$. Note that α can be any positive real number.

The stochastic process $X(t), t > 0$ with $X(0) = 0$ and having stationary and independent increments, where $X(t)$ has the density function $\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(\alpha s)}\frac{\Gamma(t-\alpha s)}{\Gamma(t)}x^{\alpha s-1}ds,\ 0 < \alpha \leq 1$, will be called Lévy-smashed gamma stochastic process. The Lévy-smashed gamma stochastic process $X(t)$ has the distribution function,for $t > 0,\ \ 0 < \alpha < 1$, $F_{(\alpha,t)}(x) = \frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(1+\alpha s)}\frac{\Gamma(t-\alpha s)}{\Gamma(t)}x^{\alpha s}ds$. From [1] and [9], we can prove that the Lévy-smashed gamma distribution with parameter α is attracted to the stable distribution with exponent $\alpha,\ \ 0 < \alpha < 1$. Namely,

$$\lim_{n\to\infty} L_f(\frac{y}{n}) = \lim_{n\to\infty}\int_0^\infty e^{-yx}\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)}{\Gamma(\alpha s)}\frac{\Gamma(n-\alpha s)}{\Gamma(n)}n^{\alpha s}x^{\alpha s-1}dsdx$$
$$= \lim_{n\to\infty}\frac{1}{2\pi i}\oint_L \frac{\Gamma(s)\Gamma(n-\alpha s)}{\Gamma(n)}n^{\alpha s}y^{-\alpha s}ds = \lim_{n\to\infty}\sum_{k=0}^{\infty}\frac{(-1)^k\Gamma(n+\alpha k)}{k!\,\Gamma(n)}\left(\frac{y}{n}\right)^{\alpha k} = e^{-y^\alpha}$$

5 Remarks

In [3], they consider the signalling problem for the standard diffusion equation $\frac{\partial}{\partial t}u(x,t) = D\frac{\partial^2}{\partial x^2}u(x,t)$ with the conditions $u(x,0^+) = 0\ x > 0; u(0^+,t) = h(t), u(+\infty,t) = 0,\ t > 0$. And they say "$\cdots$ Then the solution is as follows $u(x,t) = \int_0^t \mathcal{G}_s^d(x,\tau)h(t-\tau)d\tau$, where $\mathcal{G}_s^d(x,t) = \frac{x}{2\sqrt{\pi D}}t^{-\frac{3}{2}}exp{-\frac{x^2}{4Dt}}$. Here $\mathcal{G}_s^d(x,t)$ represents the fundamental solution (or Green function) of the signalling problem, since it corresponds to $h(t) = \delta(t)$. We note that

$$\mathcal{G}_s^d(x,t) = p_{LS}(t;\mu) := \frac{\sqrt{\mu}}{\sqrt{2\pi}t^{\frac{3}{2}}}exp(-\frac{\mu}{2t}),\ t \geq 0,\ \mu = \frac{x^2}{2D} \tag{5.1}$$

where $p_{LS}(t;\mu)$ denotes the one-sided Lévy-Smirnov pdf spread out over all non-negative t (the time variable). The Lévy-Smirnov pdf has all moments of integer order infinite, since it decays at infinity as $t^{-\frac{3}{2}}$. However, we note that the absolute moments of real order ν are finite only if $0 \leq \nu < \frac{1}{2}$. In particular, for this pdf the mean is infinite, for which we can take the median as expected value. From $\mathcal{P}_{LS}(t_{med};\mu) = \frac{1}{2}$, it turns out that $t_{med} \approx 2\mu$, since the complementary error function gets the value $\frac{1}{2}$ as its argument is approximatively $\frac{1}{2}$. $\cdots$".

With the inspiration from the above paragraph, take the Lévy density function with the parameter $\alpha = \frac{1}{2}$, then the density function is well known to be $\frac{1}{2\sqrt{\pi}t^{\frac{3}{2}}}exp(-\frac{1}{4t})$. Put this in the mellin convolution formula $\int_t h_1(\frac{x}{t})h_2(t)\frac{dt}{t}$, then it becomes $\int_0^\infty \frac{\sqrt{t}}{2\sqrt{\pi}x^{\frac{3}{2}}}exp\left(-\frac{t}{4x}\right)\ h_2(t)dt$. $\frac{\sqrt{t}}{2\sqrt{\pi}x^{\frac{3}{2}}}exp\left(-\frac{t}{4x}\right)$ has the same form with (5.1) when $t = \mu$. Therefore the cases of Lévy smashing treated in section 4 can be thought of as superstatistics in Physics and Bayesian statistical analysis, subordination in statistics, namely,

$$f_{(\frac{1}{2})}(x) = \int_0^\infty \frac{1}{2\pi i}\oint_L \frac{\Gamma(2-2s)}{\frac{1}{2}\Gamma(1-s)}x^{-s}y^s ds\ \frac{y^{t-1}}{\Gamma(t)}e^{-y}\frac{dy}{y} = \int_0^\infty \frac{\sqrt{y}}{2\sqrt{\pi}x^{\frac{3}{2}}}exp\left(-\frac{y}{4x}\right)\ \frac{y^{t-1}}{\Gamma(t)}e^{-y}dy$$
$$= \int_0^\infty \frac{\sqrt{y}}{2\sqrt{\pi}x^{\frac{3}{2}}}\sum_{k=0}^\infty \frac{(-1)^k y^k}{k!\ (4x)^k}\ \frac{y^{t-1}}{\Gamma(t)}e^{-y}dy = \frac{1}{2\sqrt{\pi}\Gamma(t)x^{\frac{3}{2}}}\sum_{k=0}^\infty \frac{(-1)^k}{k!\ (4x)^k}\ \int_0^\infty y^{t-1+k+\frac{1}{2}}e^{-y}dy$$
$$\frac{1}{2\sqrt{\pi}\Gamma(t)x^{\frac{3}{2}}}\sum_{k=0}^\infty \frac{(-1)^k\Gamma(t+k+\frac{1}{2})}{k!\ (4x)^k} = \frac{4\Gamma(t+\frac{1}{2})(4x)^{t-1}}{\sqrt{\pi}\Gamma(t)}\left(1+4x\right)^{-t-\frac{1}{2}}$$

But in this paper, the concept of Lévy smashing is totally different from the point of view of superstatistics in Physics and Bayesian statistical analysis, subordination in statistics.

Acknowledgement

The author would like to thank to the Department of Science and Technology, Government of India, New Delhi, for the financial assistance under Project No. SR/S4/MS:287/05 and the Centre for Mathematical Sciences for providing all facilities.

References

[1] Feller, W. An Introduction to Probability Theory and Its Applications, Vol. II, Wiley, New York (1966).

[2] Jung Hun Han, On the Lévy density function, arXiv:1102.2709v1 [math.ST] 14 Feb 2011.

[3] Mainardi F., Paradisi P and Gorenflo R., Probability distributions generated by fractional diffusion equations, FRACALMO preprint, www.fracalmo.org, 1997.

[4] Mainardi F., Pagnini G. and Gorenflo R., Mellin convolution for subordinated stable processes, J. Math. Sci., Vol. 132, No. 5, 2006.

[5] Mathai A.M., Some properties of Mittag-Leffler functions and matrix-variate analogues: a statistical perspective, Preprint.

[6] Mathai A.M. and Haubold Hans J., Special Functions for Applied Scientists, Springer, New York, 2008.

[7] Mathai A.M., Saxena R.K. and Haubold Hans J., The H-function: Theory and Applications, Springer, New York, 2010.

[8] Penson K. A. and Górska K., Exact and explicit probability densities for one-sided Lévy stable distributions, arXiv:1007.0193v1 [cond-mat.stat-mech] 1 Jul 2010.

[9] Pillai R. N., On mittag-leffler functions and related distributions, Ann. Inst. Statist. Math. Vol. 42, No. I, 157-161 (1990).

[10] Prudnikov A. P., Brychkov Yu. A. and Marichev O. I., Integrals and Series, Vols. 1-5 (Gordon and Breach, Amsterdam, 1992-1998).

Approximation by Semigroups of Spherical Operators*

Yu Guang Wang Feilong Cao†

Department of Mathematics, China Jiliang University,
Hangzhou 310018, Zhejiang Province, P R China

Abstract

This paper discusses the approximation by a class of so called exponential-type multiplier operators. It is proved that such operators form a strongly continuous semigroup of contraction operators of class $(\mathscr{C}_0)$, from which the equivalence between approximation for these operators and K-functionals introduced by the operators is given. As examples, the constructed r-th Boolean of generalized spherical Abel-Poisson operator and r-th Boolean of generalized spherical Weierstrass operator denoted by $\oplus^r V_t^\gamma$ and $\oplus^r W_t^\kappa$ separately (r is any positive integer, $0 < \gamma, \kappa \leq 1$ and $t > 0$) satisfy that $\| \oplus^r V_t^\gamma f - f\|_{\mathcal{X}} \approx \omega^{r\gamma}(f, t^{1/\gamma})_{\mathcal{X}}$ and $\| \oplus^r W_t^\kappa f - f\|_{\mathcal{X}} \approx \omega^{2r\gamma}(f, t^{1/(2\kappa)})_{\mathcal{X}}$, for all $f \in \mathcal{X}$, where $\mathcal{X}$ is a Banach space of continuous functions or $\mathcal{L}^p$-integrable functions $(1 \leq p < \infty)$ and $\|\cdot\|_{\mathcal{X}}$ is the norm on $\mathcal{X}$ and $\omega^s(f,t)_{\mathcal{X}}$ is the moduli of smoothness of degree $s > 0$ for $f \in \mathcal{X}$. The saturation order and saturation class of the regular exponential-type multiplier operators with positive kernels are also obtained. Moreover, it is proved that $\oplus^r V_t^\gamma$ and $\oplus^r W_t^\kappa$ have the same saturation class if $\gamma = 2\kappa$.

MSC(2000): 42C10; 41A25

Keywords: sphere; semigroup; approximation; moduli of smoothness; multiplier

1 Introduction

The study on spherical approximation started early in 1960's when Butzer, Berens and Pawelke (see [2]) studied the saturation properties of singular integrals on the sphere. Since 1980's, the approximation theory on the sphere has been developed by Nikol'skiĭ, Lizorkin et al. and since 1990's, Wang, Li and Dai et al. pursued the research of

*The research was supported by the National Natural Science Foundation of China (No. 60873206), the Natural Science Foundation of Zhejiang Province of China (No. Y7080235) and the Innovation Foundation of Post-Graduates of Zhejiang Province of China (No. YK2008066).

†Corresponding author: Feilong Cao, E-mail: feilongcao@gmail.com

approximation theory on the sphere (see for example [25]). Some classical theorems in the case of one dimension, for example, Jackson-type theorem, were generalized to the sphere (see [23]). The approximation tools then were mainly polynomial operators, say, spherical Jackson operators (see [22]) and de la Vallée Poussin means (see [3]). In recent years, the study on spherical approximation has attracted more and more attention of researchers. There have been many interesting works in this field, such as [8]-[13], [17]. In particular, we notice that as a non-polynomial operator on the sphere, the classical Abel-Poisson operator was studied by Dai and Ditzian [8], where the equivalence between approximation by Abel-Poisson operators on the sphere and the K-functional introduced by their infinitesimal generator was given.

This paper is mainly about the approximation by non-polynomial operators on the sphere, namely, a class of exponential-type operators denoted by $\{T_p^\gamma(t)|0 \le t < \infty\}$ with polynomial $p(x)$ and $0 < \gamma \le 1$, in the form of $T_p^\gamma(t)f = \sum\limits_{k=0}^{\infty} e^{-(p(x))^\gamma t} Y_k f$ $(f \in \mathcal{X})$, where $Y_k f$ is the k-th term of the Laplace expansion of f on the sphere. $T_p^\gamma(t)$ is called regular if the coefficient of the first item of $p(x)$ is positive, $p(0) = 0$ and the degree of $p(x)$ is larger than 0. These regular operators with positive kernels are proved to form a semigroup of class $(\mathscr{C}_0)$.

The semigroups of operators were early studied by Hille and Phillips (see [19]). Later, in 1960's, Butzer and Berens studied the approximation properties of semigroups of operators (see [2]). By Ditzian and Ivanov's method (see [14, Sec.5]), we will prove that for a class of exponential-type multiplier operators that form a semigroup of class $(\mathscr{C}_0)$ and for $r \in \mathbb{Z}_+$ there holds

$$\|(T(t) - I)^r f\|_{\mathcal{X}} \approx \inf_{g \in \mathcal{D}_1(\mathcal{A}^r)} (\|f - g\|_{\mathcal{X}} + \|\mathcal{A}^r g\|_{\mathcal{X}}) \tag{1}$$

if $\{T(s)|s > 0\}$ possesses additive properties that $T(t)f \in \mathcal{D}_1(\mathcal{A})$ for all $t > 0$ and $f \in \mathcal{X}$ and $\mathcal{A}$ satisfies the further following Bernstein-type inequality

$$t\|\mathcal{A}T(t)f\|_{\mathcal{X}} \le N\|f\|_{\mathcal{X}} \quad (t > 0,\ f \in \mathcal{X}), \tag{2}$$

where $\mathcal{D}_1(\mathcal{A}^r) = \{g \in \mathcal{X} |\ \mathcal{A}^r g \in \mathcal{X}\}$ and r-th power of multiplier operator $\mathcal{A}$.

It will be proved that the regular $T_p^\gamma(t)$ with positive kernel also satisfies (2). Hence (1) holds for $T_p^\gamma(t)$. Noticing $(I - T(t))^r f = f - \oplus^r T(t)f$, here $\oplus^r T(t)$ is defined as the r-th Boolean of $T(t)$, we shall obtain the equivalence between the approximation of $\oplus^r T_p^\gamma(t)$ and K-functional introduced by the multiplier operator, actually infinitesimal generator of $\left\{T_p^\gamma(t)|0 \le t < \infty\right\}$. In terms of convolution, we shall also obtain the saturation order and saturation class of Booleans of the regular $\oplus^r T_p^\gamma(t)$. The generalized spherical Abel-Poisson operators V_t^γ and generalized spherical Weierstrass operators

W_t^κ that were introduced by Bochner in [5, P. 43-47] and [4, P. 84] respectively, are actually the examples of regular exponential-type operators with positive kernels. Thus follows the equivalence between approximation of $\oplus^r V_t^\gamma$ or $\oplus^r W_t^\kappa$ and the moduli of smoothness on the sphere. It is also discussed that the saturation properties of $\oplus^r V_t^\gamma$ and $\oplus^r W_t^\kappa$.

The paper is structured as follows. In Section 2, some basic concepts will be introduced. Section 3 discusses the properties of some spherical function classes as well as K-functionals that are introduced by multiplier sequences. Section 4 consists of our main results and their proofs. In Section 5, the results of previous sections are applied to $\oplus^r V_t^\gamma$ and $\oplus^r W_t^\kappa$. Our main results are Theorem 4.2.3, Theorem 4.2.4, Theorem 5.3, Theorem 5.4 and Theorem 5.5.

2 Preliminaries

Denote by letters C, C_i or $C(i)$ positive constants, where i is either a positive integer, variable, function or space on which C depends only. Their values may be different at different occurrences, even within one formula. The notation $a \approx b$ means that there exists a positive constant C such that $C^{-1}b \leq a \leq Cb$. Denote by $f(t) = \mathcal{O}(t)$ which means there exists some constant C independent of t such that $|f(t)| \leq C|t|$, here $f(t)$ is a function with respect to t and we write $f(t) = o(t)$ if $f(t)/t$ tends to zero as $t \to \infty$ or as $t \to t_0$ where t_0 is a real number. The collection of all positive integers are denoted by $\mathbb{Z}_+$. Let $\mathbb{S}^n$ be the unit sphere in $(n+1)$-dimensional Euclidean space $\mathbb{R}^{n+1}$. x and y are denoted as the points on $\mathbb{S}^n$ and $x \cdot y$ denotes the inner product in $\mathbb{R}^{n+1}$. Denote by $d\omega_n(x)$ the elementary surface piece on $\mathbb{S}^n$ and by $d\omega(x)$ for convenience if there's no confusion. The volume of $\mathbb{S}^n$ is

$$|\mathbb{S}^n| := \int_{\mathbb{S}^n} d\omega(x) = \frac{2\pi^{n/2}}{\Gamma(n/2)}.$$

Denote by $\mathcal{L}^p(\mathbb{S}^n)$ the Banach space of p-th integrable functions $f : \mathbb{S}^n \to \mathbb{C}$ ($\mathbb{C}$ is the collection of all complexes) with norm $\|f\|_\infty := \|f\|_{\mathcal{L}^\infty(\mathbb{S}^n)} := \operatorname{ess\,sup}_{x\in\mathbb{S}^n} |f(x)|$ and

$$\|f\|_{\mathcal{L}^p} := \|f\|_{\mathcal{L}^p(\mathbb{S}^n)} := \left\{ \int_{\mathbb{S}^n} |f(x)|^p d\omega(x) \right\}^{1/p} < \infty \quad (1 \leq p < \infty).$$

Denote by $\mathcal{C}(\mathbb{S}^n)$ the Banach space consisting of all continuous functions $f : \mathbb{S}^n \to \mathbb{C}$ with norm $\|f\|_{\mathcal{C}} := \max_{x\in\mathbb{S}^n} |f(x)|$. Denote by $\mathcal{M}(\mathbb{S}^n)$ the collection of all finite regular Borel measures on $\mathbb{S}^n$ (the range is also in $\mathbb{C}$) and it is a Banach space with norm $\|\mu\|_{\mathcal{M}} := \int_{\mathbb{S}^n} |d\mu(x)|$. $\mathcal{L}^p(\mathbb{S}^n)$ $(1 \leq p \leq \infty)$, $\mathcal{C}(\mathbb{S}^n)$ and $\mathcal{M}(\mathbb{S}^n)$ may be replaced by $\mathcal{L}^p$,

$\mathcal{C}$ and $\mathcal{M}$ for convenience. Denote by $\mathcal{X}$ either $\mathcal{L}^p(\mathbb{S}^n)$ $(1 \le p < \infty)$ or $\mathcal{C}(\mathbb{S}^n)$. The dual space of $\mathcal{X}$, the collection of all bounded linear functionals on $\mathcal{X}$, is denoted by $\mathcal{X}^*$.

A function f on $\mathbb{R}^{n+1}$ is called harmonic if $\Delta f = 0$ where $\Delta = \frac{\partial}{\partial x_1^2} + \frac{\partial}{\partial x_2^2} + \cdots + \frac{\partial}{\partial x_{n+1}^2}$ is the classical Laplace operator on $\mathbb{R}^{n+1}$. The collection of all homogeneous and harmonic polynomials on $\mathbb{R}^{n+1}$ is denoted by $\mathcal{A}_k^n$. And the collection of restrictions on $\mathbb{S}^n$ of all functions in $\mathcal{A}_k^n$ is denoted by $\mathcal{H}_k^n$. Denote by Π_k^n the collection of restrictions on $\mathbb{S}^n$ of all polynomials on $\mathbb{R}^{n+1}$ which is dense in $\mathcal{X}$ and any polynomial restricted on $\mathbb{S}^n$ with degree $k \in \mathbb{Z}_+$ is in $\text{span}\{\mathcal{H}_j^n | 0 \le j \le k\}$, the linear combination of $\mathcal{H}_1^n$, $\mathcal{H}_2^n$, $\ldots$, $\mathcal{H}_k^n$.

Definition 2.1 *The r-th $(r \in \mathbb{Z}_+)$ Boolean of an operator T on $\mathcal{X}$ (an operator from $\mathcal{X}$ to $\mathcal{X}$) is defined as*

$$\oplus^r T := I - (I - T)^r = -\sum_{i=1}^{r} (-1)^i \binom{r}{i} T^i,$$

where $\binom{r}{k} := \dfrac{r!}{k!(r-k)!}$ *and* $T^0 := I$.

The projection on $\mathcal{H}_k^n$ of $f \in \mathcal{L}^1(\mathbb{S}^n)$ is defined by (see [2, Chap.1] and [25, Chap.1])

$$Y_k(f)(x) := \frac{\Gamma(\lambda)(k+\lambda)}{2\pi^{\lambda+1}} \int_{\mathbb{S}^n} P_k^\lambda(x \cdot y) f(y)\, d\omega_n(y)$$

and for $\mu \in \mathcal{M}(\mathbb{S}^n)$, $Y_k(d\mu)(x) := \frac{\Gamma(\lambda)(k+\lambda)}{2\pi^{\lambda+1}} \int_{\mathbb{S}^n} P_k^\lambda\,(x \cdot y)\, d\mu(y)$, where $2\lambda = n-2$, and $P_k^\nu(t)$, $|t| \le 1$, $k = 0, 1, 2, \ldots$, $\nu > -1/2$ is the ultraspherical polynomial (Gegenbauer polynomial) of degree k with ν and is generated by $(1 - 2tr + r^2)^{-\nu} = \sum\limits_{k=0}^{\infty} P_k^\nu(t) r^k$ $(0 \le r < 1)$. $\{P_k^\nu(t)\}_{k=0}^\infty$ forms an orthogonal system with the weight $(1-t^2)^{\nu-1/2}$, that is, for $\nu > -1/2$, $\nu \ne 0$ (see [24, P. 81]),

$$\begin{aligned} &\int_{-1}^{1} P_k^\nu(t) P_j^\nu(t) (1-t^2)^{\nu-1/2}\, dt \\ = &\int_0^\pi P_k^\nu(\cos\theta) P_j^\nu(\cos\theta) (\sin\theta)^{2\nu}\, d\theta = \begin{cases} (c(k,\nu))^{-1} & (k=j), \\ 0 & (k \ne j). \end{cases} \end{aligned} \tag{3}$$

where $c(k,\nu) := (2^{2\nu-1}(\Gamma(\nu))^2(k+\nu)\Gamma(k+1))/(\pi\Gamma(k+2\nu))$. Now let $\nu = \lambda := (n-2)/2 > 0$, then (see [24, P. 171])

$$\left| P_k^\lambda(t) \right| = \mathcal{O}(k^{2\lambda-1}). \tag{4}$$

A function $f \in \mathcal{X}$ is called a zonal function with x_0 on $\mathbb{S}^n$ if for some fixed $x_0 \in \mathbb{S}^n$, $f(x_0 \cdot y)$ is a constant when $x_0 \cdot y$ is unchanged. The collection of all zonal functions with

x_0 in $\mathcal{L}^p$ or $\mathcal{C}$ is denoted by $\mathcal{L}^p_\lambda(\mathbb{S}^n, x_0)$ $(1 \le p < \infty)$, $\mathcal{C}_\lambda(\mathbb{S}^n, x_0)$ ($\mathcal{L}^p_\lambda$, $\mathcal{C}_\lambda$ for convenience if there's no confusion). $\mathcal{L}^p_\lambda(\mathbb{S}^n, x_0)$ $(1 \le p < \infty)$ with norm

$$\|\varphi\|_{\mathcal{L}^p_\lambda} := \left\{ \int_{\mathbb{S}^n} |\varphi(x \cdot y)|^p d\omega(y) \right\}^{1/p} = \left\{ \left|\mathbb{S}^{n-1}\right| \int_0^\pi |\varphi(\cos\theta)|^p (\sin\theta)^{2\lambda} d\theta \right\}^{1/p}, \quad (5)$$

$\mathcal{C}_\lambda(\mathbb{S}^n, x_0)$ with norm $\|\varphi\|_{\mathcal{C}_\lambda} := \sup_{0 \le \theta \le \pi} |\varphi(\cos\theta)|$, and $\mathcal{M}_\lambda(\mathbb{S}^n, x_0)$ with norm $\|\mu\|_{\mathcal{M}_\lambda} := \left|\mathbb{S}^{n-1}\right| \int_0^\pi |d\mu^*(\theta)|$, where μ^* is the corresponding function in $\mathcal{M}[0,\pi]$ of the measure $\mu \in \mathbb{S}^n$ (actually there is a bijection between $\mathcal{M}_\lambda(\mathbb{S}^n, x_0)$ and some subset of $\mathcal{M}[0,\pi]$, see [15, P. 250-252] and [2, P. 204 and P. 209]), are all Banach spaces.

For $f \in \mathcal{L}^1(\mathbb{S}^n)$ and $\varphi \in \mathcal{L}^1_\lambda(\mathbb{S}^n)$, the convolution of f and the zonal function φ is defined by

$$(f * \varphi)(x) := \int_{\mathbb{S}^n} f(y)\varphi(x \cdot y)\, d\omega(y) \quad (x \in \mathbb{S}^n). \quad (6)$$

The convolution of $\psi \in \mathcal{L}^1_\lambda(\mathbb{S}^n)$ and $\mu \in \mathcal{M}(\mathbb{S}^n)$ is defined by $(\psi * d\mu)(x) := \int_{\mathbb{S}^n} \psi(x \cdot y) d\mu(y)$. The convolution of $f \in \mathcal{L}^1(\mathbb{S}^n)$ and the zonal measure $\mu \in \mathcal{M}_\lambda(\mathbb{S}^n)$ with x_0 is defined by

$$(f * d\mu)(x) := \int_{\mathbb{S}^n} f(y)\, d\varphi_x\mu(y) \quad (x \in \mathbb{S}^n), \quad (7)$$

where $\varphi_x\mu(E) := \mu(\rho E)$, $\rho x = x_0$, for all measurable subsets $E \subset \mathbb{S}^n$ (Please refer to [2, Chapter 1], [25, Chapter 1] and [15] for further details on convolution).

Remark 2.2 *In this paper, we follow the definition of convolution in [25] and Young's inequality still holds on the sphere.*

Definition 2.3 (see [2, P. 254]) *Let two function spaces $\mathcal{Y}$ and $\mathcal{Z}$ either be $\mathcal{C}(\mathbb{S}^n)$, $\mathcal{L}^p(\mathbb{S}^n)$ $(1 \le p < \infty)$ or $\mathcal{M}(\mathbb{S}^n)$. A sequence $\{a_k \in \mathbb{C} |\ k = 0, 1, 2, \dots\}$ is called a multiplier sequence from $\mathcal{Y}$ to $\mathcal{Z}$ if for each $f \in \mathcal{Y}$, there exists $g \in \mathcal{Z}$ whose Laplace expansion is as follows $Y_k g = \lambda/(k+\lambda)\, a_k Y_k f$ $(k = 0, 1, 2, \dots)$. The collection of all multiplier sequences from $\mathcal{Y}$ to $\mathcal{Z}$ is denoted by $(\mathcal{Y}, \mathcal{Z})$. For $\mathcal{Y} = \mathcal{Z}$, $\{a_k\}_{k=0}^\infty$ is called a multiplier sequence on $\mathcal{Y}$.*

Remark 2.4 *By* [20, P. 222-P. 231]

$$(\mathcal{M}, \mathcal{M}) = (\mathcal{C}, \mathcal{C}) = (\mathcal{L}^1, \mathcal{L}^1) \subset (\mathcal{L}^p, \mathcal{L}^p) \subset (\mathcal{L}^2, \mathcal{L}^2) \quad (1 < p < \infty).$$

Definition 2.5 *An operator T on $\mathcal{X}$ is called a multiplier operator with a sequence $\{a_k\}_{k=0}^\infty$ on $\mathcal{X}$ if for each $f \in \mathcal{X}$, $Tf \in \mathcal{X}$ and $Y_k(Tf) = a_k Y_k(f)$ $(k = 0, 1, 2, \dots)$. The operator T^α $(\alpha > 0)$ defined by $T^\alpha f \sim \sum_{k=0}^\infty -(a(k))^\alpha Y_k(f)$, where "$\sim$" is in the sense of distribution (see [12, P. 323-325] and [18, Section.1]), is called the fractional differential operator if $T^\alpha f \in \mathcal{X}$. Denote the domain of T^α by $\mathcal{D}_1(T^\alpha) = \left\{ f \in \mathcal{X} \middle|\ T^\alpha f \in \mathcal{X} \right\}$.*

Denote by $\|T\|_{\mathcal{X}}$ the norm of an operator T on $\mathcal{X}$. Then the collection of endomorphisms of $\mathcal{X}$ denoted by $\mathscr{E}(\mathcal{X})$ is a Banach algebra with norm $\|T\|_{\mathcal{X}}$ (see [6, P. 7]).

Definition 2.6 (see [6, P. 7-8]) *If $T(t)$ is an operator function on the non-negative real axis $0 \le t < \infty$ to the Banach algebra $\mathscr{E}(\mathcal{X})$, in the following conditions, if* (8) *is satisfied, $\{T(t)|\, 0 \le t < \infty\}$ is called one-parameter semigroup of operators in $\mathscr{E}(\mathcal{X})$ and it is said to be of class $(\mathscr{C}_0)$ if it satisfies the further property* (9),

$$T(t_1 + t_2) = T(t_1) \circ T(t_2)\ (t_1, t_2 \ge 0), \quad T(0) = I, \tag{8}$$

$$s\text{-}\lim_{t\to 0+} T(t) = I, \tag{9}$$

where $A \circ B$ means the composition of operators A and B, I is the identity operator on $\mathcal{X}$ and $s\text{-}\lim\limits_{t\to 0+} f_t = f$ denotes the strongly convergence which means $\|f_t - f\|_{\mathcal{X}}$ tends to zero as $t \to 0+$. $T(t)$ is called to have contraction if it satisfies

$$\|T(t)f\|_{\mathcal{X}} \le \|f\|_{\mathcal{X}} \quad (f \in \mathcal{X}). \tag{10}$$

Definition 2.7 (see [6, P. 11]) *The infinitesimal generator $\mathcal{A}$ of the semigroup $\{T(t)|\, 0 \le t < \infty\}$ is defined by $\mathcal{A} := s\text{-}\lim\limits_{t\to 0+} \dfrac{T(t)f - f}{t}$, whenever the limit exists; the domain of $\mathcal{A}$ is, in symbols $\mathcal{D}(\mathcal{A})$, being the set of elements $f \in \mathcal{X}$ for which the limit exists; for $r = 0, 1, 2, \dots$, the r-th power of $\mathcal{A}$ denoted by $\mathcal{A}^r$ is defined inductively by the relations $\mathcal{A}^0 = I$, $\mathcal{A}^1 = \mathcal{A}$, and*

$$\begin{aligned} &\mathcal{D}(\mathcal{A}^r) := \Big\{ f|\ f \in \mathcal{D}(\mathcal{A}^{r-1}) \text{ and } \mathcal{A}^{r-1} f \in \mathcal{D}(\mathcal{A}) \Big\}, \\ &\mathcal{A}^r f := \mathcal{A}(\mathcal{A}^{r-1} f) = s\text{-}\lim_{t\to 0+} \frac{T(t) - I}{t} \mathcal{A}^{r-1} f \quad (f \in \mathcal{D}(\mathcal{A}^r)). \end{aligned} \tag{11}$$

For $r \in \mathbb{Z}_+$, $\mathcal{D}(\mathcal{A}^r)$ is a linear subspace and $\mathcal{A}^r$ is a linear operator.

If an operator T in $\mathscr{E}(\mathcal{X})$ can be expressed in the form of convolution (6) or (7), then $\varphi, \psi \in \mathcal{L}^1_\lambda$ or $\mu \in \mathcal{M}_\lambda$ there is called the kernel of T. The Cesàro mean of $f \in \mathcal{X}$ denoted by $\sigma^\alpha_k(f)$ is defined by (see for instance [25, P. 49]) $\sigma^\alpha_k(f) = (1/A^\alpha_k) \sum_{j=0}^k A^\alpha_{k-j} Y_j f$, where α is a complex whose real part is not less than -1, $k \in \mathbb{Z}_+$ and $A^\alpha_k = \binom{k+\alpha}{\alpha} = \Gamma(k+\alpha+1)/(\Gamma(\alpha+1)\Gamma(k+1))$ $(k \in \mathbb{Z}_+)$, is the generalized combination number. For $\alpha > \lambda = (n-2)/2$, there holds (see for instance Theorem 2.3.10 in [25, P. 54-55])

$$\|\sigma^\alpha_k(f)\|_{\mathcal{X}} \le C(n, \alpha, \mathcal{X}) \|f\|_{\mathcal{X}} \quad (k \in \mathbb{Z}_+,\ f \in \mathcal{X}). \tag{12}$$

For any sequence $\{\mu_k\}_{k=0}^\infty$, denote $\delta\mu_k = \mu_k - \mu_{k+1}$ $(k = 0, 1, \dots)$ and $\delta^{i+1}\mu_k = \delta(\delta^i \mu_k)$ $(i = 1, 2, \dots)$.

The definitions of moduli of smoothness on the sphere are given as follows (see for instance [25, P. 56- P.57, P. 183-184]). The translation operator on $\mathcal{L}^1(\mathbb{S}^n)$ is defined by $S_\theta(f)(x) := \big(|\mathbb{S}^{n-2}| \sin^{n-1}\theta\big)^{-1} \displaystyle\int_{x\cdot y = \cos\theta} f(y)\, d\omega_{n-1}(y)$ $(0 < \theta \le \pi)$. Let

$\alpha > 0$, $\theta > 0$. The multiplier operator on $\mathcal{X}$ is called the finite difference of degree α with step θ, defined by $\Delta_\theta^\alpha := (I - S_\theta)^{\frac{\alpha}{2}} = \sum_{k=0}^{\infty}(-1)^k\binom{\frac{\alpha}{2}}{k}(S_\theta)^k$, where $\binom{\frac{\alpha}{2}}{k} := \frac{1}{k!}\frac{\alpha}{2}\left(\frac{\alpha}{2}-1\right)\cdots\left(\frac{\alpha}{2}-k+1\right)$.

Definition 2.8 *Let $f \in \mathcal{X}$, $\alpha > 0$. The moduli of smoothness of degree α of f is defined as* $\omega^\alpha(f,t)_\mathcal{X} := \sup\left\{\|\Delta_\theta^\alpha f\|_\mathcal{X} : 0 < \theta \le t\right\}$ $(0 < t \le \pi)$.

Definition 2.9 (see for instance [12, P. 323-325]) *Let $f \in \mathcal{X}$, $t > 0$. The K-functional introduced by multiplier operator $\mathcal{A}$ with multiplier sequence $\{a_k\}_{k=0}^\infty$ is defined as* $K_\mathcal{A}(f,t)_\mathcal{X} := \inf_{g\in\mathcal{D}_1(\mathcal{A})}\left\{\|f-g\|_\mathcal{X} + t\|\mathcal{A}g\|_\mathcal{X}\right\}$, here

$$\mathcal{D}_1(\mathcal{A}) = \left\{f \in \mathcal{X} | \text{there exists } \ g \in \mathcal{X} \ \text{ such that } \ a(k)Y_k f = Y_k g,\ k = 0,1,2,\dots\right\}. \quad (13)$$

Particularly, for the K-functional introduced by $(\alpha/2)$-th Laplace-Beltrami operator $D^{\alpha/2}$ with the multiplier sequence $\left\{\left(-k(k+2\lambda)\right)^{\alpha/2}\right\}_{k=0}^\infty$ $(\alpha > 0)$, there holds

$$K_{D^{\alpha/2}}(f,t^\alpha)_\mathcal{X} \approx \omega^\alpha(f,t)_\mathcal{X}, \quad (14)$$

which was finally proved by Riemenschneider and Wang (see [23]). One might also define K-functional introduced by infinitesimal generator as follows.

Definition 2.10 (see [6, Section 3.4]) *Suppose $\{T(t)|0 \le t < \infty\}$ a semigroup of operator of class $(\mathscr{C}_0)$ in $\mathscr{E}(\mathcal{X})$ and let $\mathcal{A}$ be its infinitesimal generator. For $f \in \mathcal{X}$, the r-th K-functional introduced by $\mathcal{A}$ is defined by* $K^*_{\mathcal{A}^r}(f,t)_\mathcal{X} := \inf_{g\in\mathcal{D}(\mathcal{A}^r)}\left\{\|f-g\|_\mathcal{X} + t\|\mathcal{A}^r g\|_\mathcal{X}\right\}$, *here $\mathcal{D}(\mathcal{A}^r)$ is defined by Definition 2.7 and $\mathcal{A}^r$ denotes the r-th power of infinitesimal generator $\mathcal{A}$.*

Finally, it is worth mentioning here the concept of saturation for operators on $\mathcal{X}$ (see [2, P. 217]), which was first proposed by Favard in [16].

Definition 2.11 *Let $\varphi(\rho)$ be a positive function with respect to ρ, $0 < \rho < \infty$, tending monotonely to zero as $\rho \to \infty$. For a sequence of operators $\{I_\rho\}_{\rho>0}$ if there exists $\mathcal{K} \subsetneqq \mathcal{X}$ such that*

(i) *If $\|I_\rho(f) - f\|_p = o(\varphi(\rho))$, then $I_\rho f = f$ for all $\rho > 0$;*

(ii) *$\|I_\rho(f) - f\|_p = \mathcal{O}(\varphi(\rho))$ if and only if $f \in \mathcal{K}$;*

then I_ρ is said to be saturated on $\mathcal{X}$ with order $\mathcal{O}(\varphi(\rho))$ and $\mathcal{K}$ is called its saturation class.

3 Classes and K-Functionals Introduced by Multiplier Sequences on the Sphere

Let $\psi(x)$ be a function from $\mathbb{R}$ to $\mathbb{C}$, define $\mathcal{H}\Big(\big\{\psi(k)\big\}_{k=0}^{\infty};\mathcal{X}\Big) := \Big\{f\in\mathcal{X}\big|$ there exists $g\in\mathcal{X}$ such that $\psi(k)Y_kf = Y_kg$ for $k=0,1,\dots\Big\}$, and denoted by $\mathcal{H}\big(\psi(k);\mathcal{X}\big)$ for convenience.

Theorem 3.1 *Suppose that $\psi_0(x)$ and $\varphi_0(x)$ are functions from $[0,+\infty)$ to $\mathbb{C}$ and there exist v_1, $v_2\in\mathbb{R}$ such that $\psi(x)=e^{iv_1\pi}\psi_0(x)$ and $\varphi(x)=e^{iv_2\pi}\varphi_0(x)$ are both real valued functions. And $0<\lim\limits_{x\to+\infty}\big(\psi(x)/\varphi(x)\big)=c_0<+\infty$ and $\psi(0)=\varphi(0)=0$, setting*

$$g(t) := \begin{cases} \dfrac{\psi(t^{-1})}{\varphi(t^{-1})}, & 0<t<+\infty,\\ c_0, & t=0, \end{cases}$$

if $g(t)$, $(g(t))^{-1}\in\mathcal{C}^{2\lambda+2}[0,+\infty)$ ($\mathcal{C}^{2\lambda+2}[0,+\infty)$ is the collection of real functions on $[0,+\infty)$ that are $(2\lambda+2)$ times continuously differentiable), then for $-\infty<s<+\infty$, there holds

$$\mathcal{H}\big((\psi_0(k))^s;\mathcal{X}\big)=\mathcal{H}\big((\varphi_0(k))^s;\mathcal{X}\big).$$

Proof. We first prove that for any $s\in(-\infty,+\infty)$,

$$\left\{C_k^s=\frac{k+\lambda}{\lambda}\left(\frac{\psi(k)}{\varphi(k)}\right)^s, k=1,2,\dots;C_0^s=\varphi(0)\right\}$$

belongs to $(\mathcal{M},\mathcal{M})$. In fact, for any $k\in\mathbb{Z}_+$, $g_1(t)=(g(t))^s\in\mathcal{C}^{2\lambda+2}[0,+\infty)$ allows us to use Taylor's formula for $g_1(t)$ on $[0,1/k]$ at $t=0$, that is, there exists $0<\xi_k<1/k$ such that

$$\begin{aligned}\left(\frac{\psi(k)}{\varphi(k)}\right)^s &= g_1\left(\frac{1}{k}\right)=g_1(0)+g_1^{(1)}(0)\frac{1}{k}+\cdots+\frac{g_1^{(2\lambda+1)}(0)}{(2\lambda+1)!}\left(\frac{1}{k}\right)^{2\lambda+1}\\ &\quad+\frac{g_1^{(2\lambda+2)}(\xi_k)}{(2\lambda+2)!}\left(\frac{1}{k}\right)^{2\lambda+2}. \end{aligned}\tag{15}$$

We deduce from the assumption that $g_1^{(i)}(0)$, $i=0,1,\dots,2\lambda+1$, are constants depending only on ψ, φ, s and n, and

$$\left|g_1^{(2\lambda+2)}(\xi_k)\right|\le C(\varphi,\psi,s,n).\tag{16}$$

Multiply (15) by $(n+\lambda)/\lambda$, then according to Definition 2.3, one can verify that the sequence consisting of the first term $\big(g_1(0)(k+\lambda)\big)/\lambda=\big(c_0(k+\lambda)\big)/\lambda$ $(k=1,2,\dots)$ belongs to $(\mathcal{M},\mathcal{M})$. [1, I, P. 202-203] proved that $(k+\lambda)/k^\alpha$ $(\alpha>0)$ are Gegenbauer-Stieltjes-coefficients of some measure in $\mathcal{M}$. For the last term of (15), we estimate the

following series that

$$\left|\frac{1}{|\mathbb{S}^n|}\sum_{k=1}^{\infty}\frac{g_1^{(2\lambda+2)}(\xi_k)}{(2\lambda+2)!}\left(\frac{1}{k}\right)^{2\lambda+2}\frac{k+\lambda}{\lambda}P_k^{\lambda}(\cos\theta)(\sin\theta)^{2\lambda}\right| \;\leq\; C(\varphi,\psi,s,n)\sum_{k=0}^{\infty}\frac{1}{k^2}<\infty,$$

here the inequality is due to (16) and (4). Thus, there exits $\mu_1\in\mathcal{M}_\lambda(\mathbb{S}^n)$ such that

$$d\mu_1^*(\theta)=\frac{1}{|\mathbb{S}^n|}\sum_{k=1}^{\infty}\frac{g_1^{(2\lambda+2)}(\xi_k)}{(2\lambda+2)!}\left(\frac{1}{k}\right)^{2\lambda+2}\frac{k+\lambda}{\lambda}P_k^{\lambda}(\cos\theta)(\sin\theta)^{2\lambda}d\theta.$$

It follows that the Gegenbauer-Stieltjes-coefficients of μ_1 are

$$\begin{aligned}\check{\mu}_1(j) &= |\mathbb{S}^n|c(j,\lambda)\int_0^{\pi}P_j^{\lambda}(\cos\theta)d\mu_1^*(\theta)\\ &= \frac{j+\lambda}{\lambda}\frac{g_1^{(2\lambda+2)}(\xi_j)}{(2\lambda+2)!}\left(\frac{1}{j}\right)^{2\lambda+2}\qquad(j=1,2,\dots).\end{aligned}$$

Lemma 5.3.1 in [2, P. 255] tells us that a sequence is Gegenbauer-Stieltjes-coefficients of some zonal measure on $\mathbb{S}^n$ if and only if it belongs to $(\mathcal{M},\mathcal{M})$, hence,

$$\left\{\frac{k+\lambda}{\lambda}\frac{g_1^{(2\lambda+2)}(\xi_k)}{(2\lambda+2)!}\left(\frac{1}{k}\right)^{2\lambda+2}\right\}_{k=1}^{\infty}$$

are Gegenbauer-Stieltjes-coefficients of μ_1, which implies that

$$\left\{C_k^s=\frac{k+\lambda}{\lambda}\left(\frac{\psi(k)}{\varphi(k)}\right)^s,\ k=1,2,\dots;C_0^s=0\right\}$$

belongs to $(\mathcal{M},\mathcal{M})$. By Remark 2.4, we obtain

$$\left\{C_k^s\right\}_{k=0}^{\infty}\in(\mathcal{M},\mathcal{M}) \;=\; (\mathcal{C},\mathcal{C})\subset(\mathcal{L}^p,\mathcal{L}^p)\ (1\leq p<\infty). \tag{17}$$

Now we prove $\mathcal{H}\big((\varphi(k))^s;\mathcal{X}\big)=\mathcal{H}\big((\psi(k))^s;\mathcal{X}\big)$. We will just take account of the case of $\mathcal{X}=\mathcal{L}^p(\mathbb{S}^n)$ $(1\leq p<\infty)$ and the proof of the case of $\mathcal{X}=\mathcal{C}(\mathbb{S}^n)$ is analogous. For $f\in\mathcal{H}\big((\psi(k))^s;\mathcal{L}^p(\mathbb{S}^n)\big)$ $(1\leq p<\infty,\ s\in\mathbb{R})$, there exists $g_1\in\mathcal{L}^p(\mathbb{S}^n)$ such that $(\psi(k))^s\,Y_kf=Y_k\,g_1$ $(k=0,1,2,\dots)$. Thus,

$$(\varphi(k))^s\,Y_kf \;=\; \frac{\lambda}{k+\lambda}C_k^{-s}\,Y_k\,g_1\quad(k=1,2,\dots).$$

It follows from (17) that $C_k^{-s}\in(\mathcal{L}^p,\mathcal{L}^p)$. So, there exists $g_2\in\mathcal{L}^p(\mathbb{S}^n)$ such that

$$\frac{\lambda}{k+\lambda}C_k^{-s}\,Y_k\,g_1=Y_k\,g_2\quad(k=0,1,2,\dots),$$

that is, $(\varphi(k))^s\,Y_kf=Y_k\,g_2$ $(k=1,2,\dots)$, in addition, $(\varphi(0))^sY_kf=0=Y_0\,g_2$, so $f\in\mathcal{H}\big((\varphi(k))^s;\mathcal{X}\big)$. Thus, $\mathcal{H}\big((\psi(k))^s;\mathcal{X}\big)\subset\mathcal{H}\big((\varphi(k))^s;\mathcal{X}\big)$. One will similarly obtain that $\mathcal{H}\big((\varphi(k))^s;\mathcal{X}\big)\subset\mathcal{H}\big((\psi(k))^s;\mathcal{X}\big)$. Hence, $\mathcal{H}\big((\varphi_0(k))^s;\mathcal{X}\big)=\mathcal{H}\big((\varphi(k))^s;\mathcal{X}\big)=$

$\mathcal{H}\big((\psi(k))^s;\mathcal{X}\big)=\mathcal{H}\big((\psi_0(k))^s;\mathcal{X}\big)$. This completes the proof of Theorem 3.1. $\square$

Remark 3.2 *Define*

$$\mathcal{H}_1\Big(\{\psi(k)\}_{k=0}^{\infty};\mathcal{X}\Big):=\begin{cases}\Big\{f\in\mathcal{X}\Big|\ \text{there exists } g\in\mathcal{L}^p(\mathbb{S}^n) \text{ such that } \psi(k)Y_kf=Y_kg\\ \qquad\qquad \text{for } k=0,1,\dots\Big\}\ (\mathcal{X}=\mathcal{L}^p(\mathbb{S}^n),\ 1<p<\infty),\\ \Big\{f\in\mathcal{X}\Big|\ \text{there exists } \mu\in\mathcal{M}(\mathbb{S}^n) \text{ such that}\\ \qquad\qquad \psi(k)Y_kf=Y_k(d\mu) \text{ for } k=0,1,\dots\Big\}\ \big(\mathcal{X}=\mathcal{L}^1(\mathbb{S}^n)\big),\\ \Big\{f\in\mathcal{X}\Big|\ \text{there exists } g\in\mathcal{L}^\infty(\mathbb{S}^n) \text{ such that}\\ \qquad\qquad \psi(k)Y_kf=Y_kg \text{ for } k=0,1,\dots\Big\}\ (\mathcal{X}=\mathcal{C}(\mathbb{S}^n)).\end{cases}$$

Suppose $\varphi_0(x)$ and $\psi_0(x)$ satisfy the hypothese of Theorem 3.1, then one can analogously prove that

$$\mathcal{H}_1\big((\varphi_0(k))^s;\mathcal{X}\big)=\mathcal{H}_1\big((\psi_0(k))^s;\mathcal{X}\big)\quad(-\infty<s<\infty),$$

by the fact $(\mathcal{M},\mathcal{M})=(\mathcal{C},\mathcal{C})\subset(\mathcal{L}^p,\mathcal{L}^p)$ $(1\le p\le\infty)$ (see Remark 2.4).

Theorem 3.3 *Let $a(x)$ and $b(x)$ be polynomials with the same degree d. Suppose $\mathcal{A}$ and $\mathcal{B}$ are operators in $\mathcal{X}$ with multiplier sequences $\{a(k)\}_{k=0}^{\infty}$ and $\{b(k)\}_{k=0}^{\infty}$ respectively and both possess α-th power $(\alpha>0)$. If $a(x)$ and $b(x)$ satisfy the hypotheses of Theorem 3.1 and neither of $a(x)$ and $b(x)$ have any zero points on $(0,+\infty)$, then there hold*

$$\mathcal{D}_1(\mathcal{A}^\alpha)=\mathcal{D}_1(\mathcal{B}^\alpha),\quad K_{\mathcal{A}^\alpha}(f,\delta)_{\mathcal{X}}\approx K_{\mathcal{B}^\alpha}(f,\delta)_{\mathcal{X}}$$

for all $\alpha>0$ and $\delta>0$.

Proof. The idea comes from [7]. By Theorem 3.1, there holds

$$\mathcal{D}_1(\mathcal{A}^\alpha)=\mathcal{H}((a(k))^\alpha;\mathcal{X})=\mathcal{H}((b(k))^\alpha;\mathcal{X})=\mathcal{D}_1(\mathcal{B}^\alpha).\tag{18}$$

For $g\in\mathcal{D}_1(\mathcal{A}^\alpha)$, set $h:=\sum\limits_{k=0}^{\infty}\Big(\frac{b(k)}{a(k)}\Big)^\alpha Y_k(\mathcal{A}^rg)=\sum_{k=0}^{\infty}\big(b(k)\big)^\alpha Y_k(g)\sim\mathcal{B}^rg$. We show that $\|h\|_{\mathcal{X}}\le C(a,b,\alpha,n_0)\|\mathcal{A}^rg\|_{\mathcal{X}}$. Setting $\psi(x)=\big(b(x)/a(x)\big)^\alpha$ $(x\in[0,+\infty))$, it can be verified that $\Big|\big(\psi(x)\big)^{(l+1)}\Big|\le C(a,b,\alpha,l)(1+x)^{-(l+2)}$ $(x\ge1)$, from which it follows that

$$\begin{aligned}\Big|\delta^{l+1}\mu_k\Big| &\le \left|\int_0^1\cdots\int_0^1\psi^{(l+1)}(x)\Big|_{x=k+u_1+u_2+\cdots+u_{l+1}}du_1\cdots du_2\right|\\ &\le C(a,b,\alpha,l)\frac{1}{(1+k)^{l+2}}.\end{aligned}\tag{19}$$

Thus, for $l>\lambda$, one has

$$\|h\|_{\mathcal{X}}\le\sum_{k=0}^{\infty}\big|\delta^{l+1}\mu_k\big|\binom{k+l}{l}\Big\|\sigma_k^l(\mathcal{A}^\alpha g)\Big\|_{\mathcal{X}}\le C(a,b,\alpha,l)\,\|\mathcal{A}^\alpha g\|_{\mathcal{X}},$$

here the first inequality uses Abel transformations $(l+1)$ times, and the second one is by (12) and (19). That is, $\|\mathcal{B}^\alpha g\|_{\mathcal{X}} = \left\| \sum_{k=0}^{\infty} (b(k))^\alpha Y_k g \right\|_{\mathcal{X}} \le C(a,b,\alpha,l)\,\|\mathcal{A}^\alpha g\|_{\mathcal{X}}$. In the same way, one has $\|\mathcal{A}^\alpha g\|_{\mathcal{X}} \le C(a,b,\alpha,l)\,\|\mathcal{B}^\alpha g\|_{\mathcal{X}}$. So, $\|\mathcal{A}^\alpha g\|_{\mathcal{X}} \approx \|\mathcal{B}^\alpha g\|_{\mathcal{X}}$ for all $g \in \mathcal{X}$. In addition, taking into account (18), one obtains that for $f \in \mathcal{X}$,

$$\begin{aligned} K_{\mathcal{A}^\alpha}(f,\delta) &= \inf_{g_1 \in \mathcal{D}_1(\mathcal{A}^\alpha)} \left\{ \|f-g_1\|_{\mathcal{X}} + \delta^\alpha \|\mathcal{A}^\alpha g_1\|_{\mathcal{X}} \right\} \\ &\approx \inf_{g_2 \in \mathcal{D}_1(\mathcal{B}^\alpha)} \left\{ \|f-g_2\|_{\mathcal{X}} + \delta^\alpha \|\mathcal{B}^\alpha g_2\|_{\mathcal{X}} \right\} = K_{\mathcal{B}^\alpha}(f,\delta) \quad (\delta > 0). \end{aligned}$$

This proves Theorem 3.3. □

4 Approximation for Semigroups of Contraction Operators of Class $(\mathscr{C}_0)$ on the Sphere

4.1 Equivalence between approximation by semigroups of exponential-type multiplier operators and K-functional

First, we prove the following lemma.

Lemma 4.1.1 *Let $\{T(t)|0 \le t < \infty\}$ be a semigroup of class $(\mathscr{C}_0)$ in $\mathscr{E}(\mathcal{X})$ and also be multiplier operators with an exponential-type sequence $\{a_t(k)\}_{k=0}^\infty$ on $\mathcal{X}$, that is, there exits $\{a(k)\}_{k=0}^\infty$ such that $a_t(k) = e^{a(k)t}$ $(k = 0, 1, \dots)$. Then for $r \in \mathbb{Z}_+$,*

$$\mathcal{D}(\mathcal{A}^r) \subset \mathcal{D}_1(\mathcal{A}^r) \tag{20}$$

and $\mathcal{A}^r f \sim \sum_{k=0}^{\infty} (a(k))^r Y_k f$ $(f \in \mathcal{D}(\mathcal{A}^r))$. Particularly, $\mathcal{D}(\mathcal{A}) = \mathcal{D}_1(\mathcal{A})$. Moreover, for $f \in \mathcal{D}(\mathcal{A}^r)$ and $g \in \mathcal{X}$ such that $(a(k))^r Y_k f = Y_k g$ $(k = 0, 1, \dots)$, there holds

$$(T(t) - I)^r f = \int_0^t \cdots \int_0^t T(u_1 + \cdots + u_r) g \, du_1 \cdots du_r \quad a.e., \tag{21}$$

here $\mathcal{D}(\mathcal{A})$ and $\mathcal{D}_1(\mathcal{A})$ is defined by (11) *and* (13) *respectively.*

Proof. First we prove $\mathcal{D}_1(\mathcal{A}) \subset \mathcal{D}(\mathcal{A})$. Set $f \in \mathcal{D}_1(\mathcal{A})$ and $Tf \in \mathcal{X}$ such that $Tf = \sum_{k=0}^{\infty} a(k) Y_k f$ $(f \in \mathcal{D}_1(\mathcal{A}))$. For each fixed $x \in \mathbb{S}^n$, $Y_k(f)(x)$ $(k = 0, 1, \dots,)$ is a bounded linear functional on $\mathcal{X}$, which can commute with Bochner integral Then, for $k = 0, 1, 2, \dots,$

$$\begin{aligned} & Y_k \left(\int_0^t T(\tau)(Tf)\, d\tau \right)(x) = \int_0^t Y_k \left(T(\tau)(Tf) \right)(x)\, d\tau = \int_0^t e^{a(k)\tau} Y_k \big(Tf\big)(x)\, d\tau \\ = & \int_0^t e^{a(k)\tau} a(k) \left(Y_k f \right)(x)\, d\tau = \left(e^{a(k)t} - 1 \right) \left(Y_k f \right)(x) = Y_k \left(T(t) f - f \right)(x), \end{aligned}$$

hence by uniqueness theorem, we have

$$\frac{T(t)-I}{t}f=\frac{1}{t}\int_0^t T(\tau)\big(Tf\big)d\tau \quad a.e.. \tag{22}$$

$\{T(t)|0\le t<\infty\}$ is of class $(\mathscr{C}_0)$ then by (9),

$$\begin{aligned}\left\|\frac{T(t)-I}{t}f-Tf\right\|_{\mathcal{X}} &= \left\|\frac{1}{t}\int_0^t\Big(T(\tau)\big(Tf\big)-Tf\Big)d\tau\right\|_{\mathcal{X}}\\ &\le \sup_{0\le\tau<t}\big\|T(\tau)\big(Tf\big)-Tf\big\|_{\mathcal{X}}\to 0 \quad (t\to 0+).\end{aligned}$$

Therefore, $f\in\mathcal{D}(\mathcal{A})$ and $\mathcal{A}f=s-\lim\limits_{t\to 0+}\dfrac{T(t)-I}{t}f=Tf\sim\sum\limits_{k=0}^{\infty}a(k)Y_kf$ $(f\in\mathcal{D}(\mathcal{A}))$, thus $\mathcal{D}_1(\mathcal{A})\subset\mathcal{D}(\mathcal{A})$.

Conversely, for $f\in\mathcal{D}(\mathcal{A}^r)$ $(r\in\mathbb{Z}_+)$,

$$\begin{aligned}&\left(\frac{e^{a(k)t}-1}{t}\right)^r\left(Y_kf\right)(x)=Y_k\left(\left(\frac{T(t)-I}{t}\right)^r f\right)(x)\\ =\;& Y_k\left(\int_0^t\cdots\int_0^t T(u_1+\cdots+u_r)\mathcal{A}^rf du_1\cdots du_r\right)(x)\\ =\;&\int_0^t\cdots\int_0^t e^{a(k)(u_1+\cdots+u_r)}du_1\cdots du_rY_k\left(\mathcal{A}^rf\right)(x)\\ =\;&\left(\frac{e^{a(k)t}-1}{t}\right)^r\left(a(k)\right)^{-r}Y_k\left(\mathcal{A}^rf\right)(x)\quad(k=0,1,2,\dots).\end{aligned}$$

where the second equality is by Proposition 1.1.6 in [6, P. 11-12]. Hence, $Y_k\left(\mathcal{A}^rf\right)(x)=(a(k))^rY_kf$ $(k=0,1,2,\dots)$. Thus $f\in\mathcal{D}_1(\mathcal{A}^r)$. So $\mathcal{D}(\mathcal{A}^r)\subset\mathcal{D}_1(\mathcal{A}^r)$.

To prove (21), we notice that

$$\begin{aligned}&Y_k\left(\int_0^t\cdots\int_0^t T(u_1+\cdots+u_r)gdu_1\cdots du_r\right)(x)\\ =\;&\int_0^t\cdots\int_0^t e^{a(k)(u_1+\cdots+u_r)}\left(a(k)\right)^r\left(Y_kf\right)(x)du_1\cdots du_r\\ =\;&\Big(e^{a(k)t}-1\Big)^r\left(Y_kf\right)(x)=Y_k\big((T(t)-I)^rf\big)(x)\end{aligned}\tag{23}$$

with which uniqueness theorem for Laplace series yields the result. This completes the proof of Lemma 4.1.1. □

Remark 4.1.2 *Suppose the hypotheses of Lemma 4.1.1 are satisfied, by* (20)*,*

$$K_{\mathcal{A}^r}(f,t)_{\mathcal{X}}\le K^*_{\mathcal{A}^r}(f,t)_{\mathcal{X}}\quad(f\in\mathcal{X},\,t>0).$$

By Ditzian and Ivanov's method (see [14, P. 73-76]), we obtain the following theorem.

Theorem 4.1.3 *Let $\mathcal{X}$ and $\mathscr{E}(\mathcal{X})$ be defined in Section 2. Suppose that $\{T(t)|0 \le t < \infty\}$ is a strongly continuous semigroup of contraction operators of class $(\mathscr{C}_0)$ in $\mathscr{E}(\mathcal{X})$ and $\mathcal{A}$ is its infinitesimal generator and also $T(t)$ is an exponential-type multiplier operator for each $t > 0$ defined in Lemma 4.1.1. For $f \in \mathcal{X}$ and $t > 0$, $T(t)f \in \mathcal{D}(\mathcal{A}) = \mathcal{D}_1(\mathcal{A})$ and there exists some constant N independent of t and f such that*

$$t\|\mathcal{A}T(t)f\|_{\mathcal{X}} \le N\|f\|_{\mathcal{X}} \quad (\textit{for all } t > 0). \tag{24}$$

Then, for any $r \in \mathbb{Z}_+$, there holds

$$\| \oplus^r T(t)f - f\|_{\mathcal{X}} \approx K_{\mathcal{A}^r}(f, t^r)_{\mathcal{X}}.$$

To prove Theorem 4.1.3, we need the following remark that are not difficult to verify.

Remark 4.1.4 *Let $\{T(t)|\, 0 \le t < \infty\}$ be a semigroup of operators of class $(\mathscr{C}_0)$. Then for any $f \in \mathcal{D}(\mathcal{A}^r)$ $(r \in \mathbb{Z}_+)$ and $t > 0$, there holds $T(t)f \in \mathcal{D}(\mathcal{A}^r)$ and $\mathcal{A}^r T(t)f = T(t)\mathcal{A}^r f$. If $T(t)f \in \mathcal{D}(\mathcal{A})$ for all $f \in \mathcal{X}$ and all $t > 0$, then for $r \in \mathbb{Z}_+$, $T(t)f \in \mathcal{D}(\mathcal{A}^r)$ for all $f \in \mathcal{X}$.*

Proof of Theorem 4.1.3. The inequality $\| \oplus^r T(t)f - f\|_{\mathcal{X}} \le CK_{\mathcal{A}^r}(f, t^r)_{\mathcal{X}}$ is not hard to obtain. We just give the proof of the converse inequality $\| \oplus^r T(t)f - f\|_{\mathcal{X}} \ge C(r)K_{\mathcal{A}^r}(f, t^r)$. For any $g \in \mathcal{D}_1(\mathcal{A}^r)$, there exists $h \in \mathcal{X}$ such that $Y_k h = a(k)Y_k g$ $(k = 0, 1, 2, \dots)$, then by Lemma 4.1.1,

$$(T(t) - I)^r g = \int_0^t \cdots \int_0^t \int_0^t T(u_1 + u_2 + \cdots + u_r)h \, du_1 du_2 \cdots du_r, \tag{25}$$

which is a Bochner integral on $[0, t]^r$. In the rest part of proof of Theorem 4.1.3, we also view $\mathcal{A}^r$ as r-th power of the infinitesimal generator of $\{T(t)|\, 0 \le t < \infty\}$. By the corollary of Hahn-Banach theorem, for $f \in \mathcal{D}(\mathcal{A}^{r+1})$, there exists $\varphi \in \mathcal{X}^*$ such that

$$\varphi\left(f + \sum_{k=1}^{r}(-1)^r T(kt)f - (-1)^r t^r \mathcal{A}^r f\right) = \left\| f + \sum_{k=1}^{r}(-1)^r T(kt)f - (-1)^r t^r \mathcal{A}^r f \right\|_{\mathcal{X}} \tag{26}$$

and $\|\varphi\|_{\mathcal{X}^*} = 1$. Define $F : [0, \infty) \to \mathbb{C}$ by $F(x) = \varphi(T(x)f)$ $(0 \le x < \infty)$, then $F \in \mathcal{C}^{r+1}[0, \infty)$ ($\mathcal{C}^{r+1}[0, \infty)$ is the space consisting of all r times continuously-differentiable functions on $[0, \infty)$). By induction, we obtain that

$$F^{(i)}(x) = \varphi\left(T(x)\mathcal{A}^i f\right) \quad (i = 1, 2, \dots, r + 1). \tag{27}$$

Thus, $\left\|F^{(r+1)}(x)\right\|_{\mathcal{C}[0,\infty)} \le \left\|\mathcal{A}^{r+1} f\right\|_{\mathcal{X}}$. Therefore,

$$\sup_{x \ge 0}\left|F(x) + \sum_{k=1}^{r}(-1)^k \binom{r}{k} F(x + kt) - (-1)^r t^r F^{(r)}(x)\right| \le \frac{r}{2}\, t^{r+1} \left\|\mathcal{A}^{r+1} f\right\|_{\mathcal{X}}, \tag{28}$$

here we use the relation between finite differences and derivatives and the mean value theorem. Setting $x = 0$, then for $t > 0$ and $r \in \mathbb{Z}_+$, there holds, by (26), (27) and (28) that

$$\left\| f + \sum_{k=1}^{r} (-1)^k T(kt) f - (-1)^r t^r \mathcal{A}^r f \right\|_{\mathcal{X}} \leq \frac{r t^{r+1}}{2} \left\| \mathcal{A}^{r+1} f \right\|_{\mathcal{X}}. \tag{29}$$

On the other hand,

$$t\|\mathcal{A}^2 T((N+2)t) f\|_{\mathcal{X}} \quad \leq \quad \frac{N}{N+1} \|\mathcal{A} T(t) f\|_{\mathcal{X}}, \tag{30}$$

here Remark 4.1.4 and (24) are used. Then, there holds

$$\begin{aligned} t\|\mathcal{A} T((N+2)t) f\|_{\mathcal{X}} &\leq \|T((N+2)t) f - T((N+2)t+t) f + t\mathcal{A} T((N+2)t) f\|_{\mathcal{X}} \\ &\quad + \|T((N+2)t)(T(t) f - f)\|_{\mathcal{X}} \\ &\leq \frac{t^2}{2} \|\mathcal{A}^2 T((N+2)t) f\|_{\mathcal{X}} + \|T(t) f - f\|_{\mathcal{X}} \\ &\leq \frac{N}{2(N+1)} t\|\mathcal{A} T((N+2)t) f\|_{\mathcal{X}} + \left(\frac{N^2}{2} + 1 \right) \|T(t) f - f\|_{\mathcal{X}}, \end{aligned}$$

where the second inequality is because of (29) in the case $r = 1$ and the third is due to (30). Therefore,

$$t\|\mathcal{A} T((N+2)t) f\|_{\mathcal{X}} \leq C(N) \|T(t) f - f\|_{\mathcal{X}}, \tag{31}$$

where $C(N) = \big((N+1)(N^2+2)\big)/(N+2)$. Set $m = r(N+2)$ and $g = -\sum_{k=1}^{r} (-1)^k T(kmt) f$. By $T(t) f \in \mathcal{D}(\mathcal{A})$ and Remark 4.1.4, there holds $g \in \mathcal{D}(\mathcal{A}^k)$ for any $k \in \mathbb{Z}_+$. Then, it is deduced from (31) that

$$\begin{aligned} t^r \|\mathcal{A}^r g\|_{\mathcal{X}} &\leq 2^r t^r \left\| \mathcal{A}^r T(mt) f \right\|_{\mathcal{X}} \\ &= 2^r t^{r-1} \left(t \left\| \mathcal{A} T((N+2)t) \big(\mathcal{A}^{r-1} T((m-N-2)t) f \big) \right\|_{\mathcal{X}} \right) \\ &\leq 2^r C(N) t^{r-1} \left\| \mathcal{A}^{r-1} T((m-N-2)t)(T(t) - I) f \right\|_{\mathcal{X}} \\ &\quad \cdots \\ &\leq 2^r (C(N))^r \left\| (T(t) - I)^r f \right\|_{\mathcal{X}}. \end{aligned}$$

Thus, by Remark 4.1.2, one has $K_{\mathcal{A}^r}(f, t^r)_{\mathcal{X}} \leq K^*_{\mathcal{A}^r}(f, t^r)_{\mathcal{X}} \leq \|f - g\|_{\mathcal{X}} + t^r \|\mathcal{A}^r g\|_{\mathcal{X}} \leq \big(m^r + (2C(N))^r \big) \|(T(t) - I)^r f\|_{\mathcal{X}}$, which completes the proof of Theorem 4.1.3. $\square$

4.2 Approximation by operators with exponential-type multiplier sequences

Definition 4.2.1 *Let $p(x)$ be a polynomial from $\mathbb{R}$ to $\mathbb{R}$ and $0 < \gamma \leq 1$, the exponential-type multiplier operator on $\mathcal{X}$ with $p(x)$ and γ defined by*

$$T_p^{\gamma}(t)f := \sum_{k=0}^{\infty} e^{-(p(k))^{\gamma}t} Y_k f \quad (f \in \mathcal{X}) \tag{32}$$

and

$$T_p^{\gamma}(0)f := f, \tag{33}$$

is called regular if the coefficient of first term is positive, $p(0) = 0$ and the degree of $p(x)$ is larger than 0.

Remark 4.2.2 *For $f \in \mathcal{X}$, $T_p^{\gamma}(t)f = f * \varphi_{p,t}^{\gamma}$, here*

$$\varphi_{p,t}^{\gamma}(\cos\theta) = \frac{1}{|\mathbb{S}^n|} \sum_{k=0}^{\infty} e^{-(p(k))^{\gamma}t} \frac{k+\lambda}{\lambda} P_k^{\lambda}(\cos\theta).$$

Then, $\varphi_{p,t}^{\gamma}(\cos\theta) \in \mathcal{L}_{\lambda}^1$ and $T_p^{\gamma}(t) \in \mathscr{E}(\mathcal{X})$. For $r \in \mathbb{Z}_+$,

$$(\mathcal{A}_p^{\gamma})^r f \sim \sum_{k=0}^{\infty} (-(p(x))^{\gamma})^r Y_k f \quad \left(\left(\mathcal{A}_p^{\gamma}\right)^r f \in \mathcal{X} \right). \tag{34}$$

Theorem 4.2.3 *Let $\{T_p^{\gamma}(t) | 0 \leq t < \infty\}$ defined by (32) and (33) be regular exponential-type multiplier operators on $\mathcal{X}$ with $p(x)$ and γ. For $t \geq 0$, the kernel $\varphi_{p,t}^{\gamma}(\cos\theta)$ of $T_p^{\gamma}(t)$ is positive. Then $\{T_p^{\gamma}(t) | 0 \leq t < \infty\}$ forms a strongly continuous semigroup of contraction operators of class $(\mathscr{C}_0)$ and for $t > 0$ and $f \in \mathcal{X}$, $T_p^{\gamma}(t)f \in \mathcal{D}(\mathcal{A})$. Moreover,*

$$\left\| \mathcal{A}_p^{\gamma} T_p^{\gamma}(t) f \right\|_{\mathcal{X}} \leq \frac{N}{t} \|f\|_{\mathcal{X}}, \tag{35}$$

here N is a constant depending only upon n, γ, $p(x)$ and $\mathcal{X}$.

Proof. For $t_1, t_2 > 0$ and $f \in \mathbb{Z}_+$, one has that

$$T_p^{\gamma}(t_1) \circ T_p^{\gamma}(t_2) f = T_p^{\gamma}(t_1 + t_2) f, \tag{36}$$

and $\left\| \varphi_{p,t}^{\gamma}(\cos(\cdot)) \right\|_{\mathcal{L}_{\lambda}^1} = \left| \mathbb{S}^{n-1} \right| \int_0^{\pi} \varphi_{p,t}^{\gamma}(\cos\theta)(\sin\theta)^{2\lambda} \, d\theta = 1$, which is by (5), the positivity of $\varphi_{p,t}^{\gamma}(\cos\theta)$ and (3). Thus,

$$\left\| T_p^{\gamma}(t) f \right\|_{\mathcal{X}} \leq \left\| \varphi_{p,t}^{\gamma}(\cos(\cdot)) \right\|_{\mathcal{L}_{\lambda}^1} \|f\|_{\mathcal{X}} = \|f\|_{\mathcal{X}}. \tag{37}$$

and also,

$$\lim_{t \to 0+} \|T_p^{\gamma}(t) f - f\|_{\mathcal{X}} = 0 \quad (\text{for all } f \in \mathcal{X}), \tag{38}$$

which is by (37), the contraction of $T_p^\gamma(t)$ and Banach-Steinhaus theorem as well as the fact that the collection of all spherical polynomials is dense in $\mathcal{X}$. By Lemma 4.1.1, there holds for any $f \in \mathcal{X}$ and $t > 0$, $T_p^\gamma(t)f \in \mathcal{D}_1(\mathcal{A}_p^\gamma) = \mathcal{D}(\mathcal{A}_p^\gamma)$, and

$$\mathcal{A}_p^\gamma T_p^\gamma(t)f = \sum_{k=0}^{\infty} -(p(k))^\gamma Y_k \left(T_p^\gamma(t)f\right) = \sum_{k=0}^{\infty} -(p(k))^\gamma e^{-(p(k))^\gamma t} Y_k f \quad a.e.. \tag{39}$$

Then we conclude by (33), (36), (37) and (38) that $T_p^\gamma(t)$ forms a strongly continuous semigroup of contraction operators of class $(\mathscr{C}_0)$.

Now we go to prove (35). There exists constants c and c' such that

$$cx^\beta \le (p(x))^\gamma \le c'x^\beta \quad (0 < x < \infty,\ \beta = d\gamma), \tag{40}$$

where d is the degree of $p(x)$. Then,

$$\begin{aligned} \left\|\mathcal{A}_p^\gamma T_p^\gamma(t)f\right\|_{\mathcal{X}} &= \left\| \sum_{k=1}^{\infty} \delta^{l+1}\big((p(k))^\gamma e^{-(p(k))^\gamma t}\big) A_k^l \sigma_k^l f \right\|_{\mathcal{X}} \\ &\le C \sum_{k=1}^{\infty} \left| \delta^{l+1}\big((p(k))^\gamma e^{-(p(k))^\gamma t}\big) k^l \right| \|f\|_{\mathcal{X}}, \end{aligned} \tag{41}$$

where the first equality uses $p(0) = 0$ and Abel transformations $(l+1)$ times and l is a positive integer larger than $\lambda = (n-2)/2$.

It is necessary to estimate $\left|\sum_{k=1}^{\infty} \delta^{l+1}\big((p(k))^\gamma e^{-(p(k))^\gamma t}\big) k^l\right|$ this moment. One can verify by induction that

$$\begin{aligned} \left(\frac{d}{dx}\right)^l \Big(\big(p(x)\big)^\gamma e^{-(p(x))^\gamma t}\Big) &= \sum_{i=0}^{l} e^{-(p(x))^\gamma t} \sum_{v=1}^{N'_{l-i}} \sum_{j=1}^{N_i} t^{s_{iv}} \big(p(x)\big)^{(s_{iv}+1)\gamma - (m_{iv}+r_{ij})} \\ &\quad \times Q_{ivj}^{d(m_{iv}+r_{ij})-(n_{iv}+i)}(x), \end{aligned}$$

where $0 \le r_{ij} \le i$, $0 \le s_{iv}, m_{iv} \le l-i$, $n_{iv} \ge l-i$ and N_i, N'_i are all positive integers, Q_{ivj}^d $(d = 0, 1, 2, \dots)$ is a polynomial with degree d and $d(m_{iv} + r_{ij}) - (n_{iv} + i) \ge 0$.

Thus, for $x \ge 1$, one has

$$\begin{aligned} &\left|\left(\frac{d}{dx}\right)^l \Big(\big(p(x)\big)^\gamma e^{-(p(x))^\gamma t}\Big)\right| \le \sum_{i=0}^{l} e^{-(p(x))^\gamma t} \sum_{v=1}^{N'_{l-i}} \sum_{j=1}^{N_i} C_{ivj} t^{s_{iv}} x^{(s_{iv}+1)\beta - (n_{iv}+i)} \\ \le\ & \sum_{i=0}^{l} e^{-cx^\beta t} \sum_{v=1}^{N'_{l-i}} C_{iv} t^{s_{iv}} x^{(s_{iv}+1)\beta - l} = \sum_{i=0}^{l} C_i t^i x^{(i+1)\beta - l} e^{-cx^\beta t}. \end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\delta^{l+1}\big((p(k))^{\gamma}e^{-(p(k))^{\gamma}t}\big)\right| \\
\leq\ &\left|\int_0^1\cdots\int_0^1\left(\frac{d}{dx}\right)^{l+1}\Big(\big(p(x)\big)^{\gamma}e^{-(p(x))^{\gamma}t}\Big)\Big|_{x=k+u_1+u_2+\cdots+u_{l+1}}du_1\cdots du_{l+1}\right| \\
\leq\ &\sum_{i=0}^{l+1}C_i't^ik^{(i+1)\beta-(l+1)}e^{-ck^{\beta}t},
\end{aligned}$$

where C_i' $(i=0,1,\cdots,l+1)$ are positive constants depending only upon $p(x)$, γ, i and l. Hence,

$$\left|\sum_{k=1}^{\infty}\delta^{l+1}\big((p(k))^{\gamma}e^{-(p(k))^{\gamma}t}\big)k^l\right|\leq\sum_{i=0}^{l+1}C_i't^i\sum_{k=1}^{\infty}k^{(i+1)\beta-1}e^{-ck^{\beta}t}. \tag{42}$$

Now consider function $a(x)=x^{(i+1)\beta-1}e^{-cx^{\beta}t}$ $(i=0,1,\ldots,m)$, $\dfrac{d}{dx}\big(a(x)\big)=\Big(\big((i+1)\beta-1\big)+(-c\beta t)x^{\beta}\Big)x^{(i+1)\beta-2}e^{-cx^{\beta}t}$, thus there exists integer $k_i\geq 0$ (may depend on t) such that $a(k+1)\geq a(x)\geq a(k)$ $(1\leq k\leq x\leq k+1\leq k_i)$ and $a(k+1)\leq a(x)\leq a(k)$ $(k+1\geq x\geq k>k_i)$, here k is a positive integer. Then from (42),

$$\begin{aligned}
&\left|\sum_{k=1}^{\infty}\delta^{l+1}\big(k^{\beta}e^{-ck^{\beta}t}\big)k^l\right|\leq\sum_{i=0}^{l+1}C_i't^i\sum_{k=1}^{\infty}k^{(i+1)\beta-1}e^{-ck^{\beta}t} \\
\leq\ &\sum_{i=0}^{l+1}C_i't^i\left(\sum_{k=1}^{k_i}\int_k^{k+1}e^{-cx^{\beta}t}x^{(i+1)\beta-1}dx+\sum_{k=k_i+1}^{\infty}\int_{k-1}^{k}e^{-cx^{\beta}t}x^{(i+1)\beta-1}dx\right) \\
\leq\ &\sum_{i=0}^{l+1}\Big(2C_i't^i\Big)\int_0^{\infty}e^{-cx^{\beta}t}x^{(i+1)\beta-1}dx=\frac{N_1}{t},
\end{aligned}$$

where $N_1=\left(\sum\limits_{i=0}^{l+1}2C_i'c^{-(i+1)}i!\right)\beta^{-1}$. Therefore, by (41), we obtain that $\big\|\mathcal{A}_p^{\gamma}T_p^{\gamma}(t)f\big\|_{\mathcal{X}}\leq\dfrac{N}{t}\|f\|_{\mathcal{X}}$, here $N=CN_1=N(p(x),\gamma,n,\mathcal{X})$. The proof of Theorem 4.2.3 is completed. □

Theorem 4.2.4 *Let $T_p^{\gamma}(t)$ $(t\geq 0)$ be regular exponential-type multiplier operators on $\mathcal{X}$ with $p(x)$ and $0<\gamma<\infty$ and their kernels are positive, then there holds for $r\in\mathbb{Z}_+$ that*

$$\|\oplus^r T_p^{\gamma}(t)f-f\|_{\mathcal{X}}\approx K_{(\mathcal{A}_p^{\gamma})^r}(f,t^r)_{\mathcal{X}} \tag{43}$$

for all $f\in\mathcal{X}$, here $\mathcal{A}_p^{\gamma}$ with multiplier operators is the infinitesimal generator of $\{T_p^{\gamma}(t)|0\leq t<\infty\}$ (see (34)). Moreover, $\{\oplus^r T_p^{\gamma}(t)|0\leq t<\infty\}$ is saturated with order $\mathcal{O}(t^r)$ and its saturation class is $\mathcal{H}_1\big(-(p(k))^{r\gamma};\mathcal{X}\big)$.

Proof. (43) follows from Theorem 4.2.3 and Theorem 4.1.3. We now discuss the saturation property of $\{\oplus^r T_p^\gamma(t)|0 \le t < \infty\}$. Let $\mathcal{A}_p^\gamma$ be the infinitesimal generator of $T_p^\gamma(t)$ and $\left\{\widehat{\varphi}_{r,p,t}^\gamma(k)\right\}_{k=0}^\infty$ be the Gegenbauer coefficients of the kernel $\varphi_{r,p,t}^\gamma(\cos\theta)$ of $\oplus^r T_p^\gamma(t)$. Then, $\varphi_{r,p,t}^\gamma(\cos\theta) = \left(1/|\mathbb{S}^n|\right)\sum\limits_{k=0}^\infty \Big(1-\left(1-e^{-(p(k))^\gamma t}\right)^r\Big)\dfrac{k+\lambda}{\lambda}P_k^\lambda(\cos\theta) \in \mathcal{C}_\lambda(\mathbb{S}^n)$, and $\widehat{\varphi}_{r,p,t}^\gamma(k) = ((k+\lambda)/\lambda)\Big(1-\left(1-e^{-(p(k))^\gamma t}\right)^r\Big)$ $(k = 0, 1, 2, \dots)$. Thus,

$$\lim_{t\to 0+}\frac{\dfrac{\lambda}{k+\lambda}\,\widehat{\varphi}_{r,p,t}^\gamma(k)-1}{t^r} = -(p(k))^{r\gamma} \quad (k = 0, 1, 2, \dots). \tag{44}$$

In addition,

$$\left|\mathbb{S}^{n-1}\right|\int_0^\pi \varphi_{r,p,t}^\gamma(\cos\theta)\,(\sin\theta)^{2\lambda}\;d\theta = 1. \tag{45}$$

and

$$\|\oplus^r T_p^\gamma(t)f\|_{\mathcal{X}} \le 2^r\|f\|_{\mathcal{X}}. \tag{46}$$

Using Theorem 3.1 of [2, P. 220] (the $c(0,\lambda)$ there is actually $|\mathbb{S}^{n-1}|/|\mathbb{S}^n|$ in (45)), by (44), (45) and (46), one obtains that $f \in \mathcal{H}_1\big(-(p(k))^{r\gamma};\mathcal{X}\big)$ if $\|\oplus^r T_p^\gamma(t)f - f\|_{\mathcal{X}} = \mathcal{O}(t^r)$ and f is a constant if $\|\oplus^r T_p^\gamma(t)f - f\|_{\mathcal{X}} = o(t^r)$.

Conversely, suppose $f \in \mathcal{H}_1\big(-(p(k))^{r\gamma};\mathcal{X}\big)$. First, for the case of $\mathcal{X} = \mathcal{L}^p(\mathbb{S}^n)$, there exists $g \in \mathcal{L}^p(\mathbb{S}^n)$ such that $-(p(x))^{r\gamma}Y_k f = Y_k g$ $(k = 0, 1, 2, \dots)$. By Lemma 4.1.1, $\|\oplus^r T_p^\gamma(t)f - f\|_p = \left\|\left(T_p^\gamma(t)-I\right)^r f\right\|_p = \left\|\int_0^t\cdots\int_0^t T(u_1+u_2+\cdots+u_r)g\,du_1\cdots du_r\right\|_p$ $\le \|g\|_p\,t^r = \mathcal{O}(t^r)$. The proofs for the cases of $\mathcal{X} = \mathcal{L}^1(\mathbb{S}^n)$ and $\mathcal{C}(\mathbb{S}^n)$ are similar. So we just take the case of $\mathcal{X} = \mathcal{L}^1(\mathbb{S}^n)$ for example (the framework of the proof below is from [2, P. 229-231]). By hypothesis that $f \in \mathcal{H}_1\big(-(p(k))^{r\gamma};\mathcal{L}^1(\mathbb{S}^n)\big)$, there exists $\mu \in \mathcal{M}(\mathbb{S}^n)$ such that

$$-(p(x))^{r\gamma}Y_k f = Y_k(d\mu) \quad (k = 0, 1, \dots). \tag{47}$$

By Remark 2.2, the convolution $(\varphi_{p,t}^\gamma * d\mu)(x) := \int_{\mathbb{S}^n}\varphi_{p,t}^\gamma(x\cdot y)d\mu(y) \in \mathcal{L}^1(\mathbb{S}^n)$ and $\|\varphi_{p,t}^\gamma * d\mu\|_1 \le \|\varphi_{p,t}^\gamma\|_{\mathcal{L}_\lambda^1}\|\mu\|_{\mathcal{M}} = \|\mu\|_{\mathcal{M}}$. For given $d\mu$, $h(t) = \varphi_{p,t}^\gamma * d\mu$ defines a vector valued function from $(0,\infty)$ to $\mathcal{L}^1(\mathbb{S}^n)$ and it can be verified that for any $\varepsilon > 0$, $h(t) = \varphi_{p,t-\varepsilon}^\gamma * (\varphi_{p,\varepsilon}^\gamma * d\mu) = T_p^\gamma(t-\varepsilon)h(\varepsilon)$. Then for $0 < \varepsilon \le t_2 < t_1 < \infty$, one has $\|h(t_1) - h(t_2)\|_1 = \|T_p^\gamma(t_1-\varepsilon)h(\varepsilon) - T_p^\gamma(t_2-\varepsilon)h(\varepsilon)\|_1 \le \|T_p^\gamma(t_1-t_2)f - f\|_1 \to 0$ $(t_1 \to t_2)$, where (38) is used. Therefore, $h(t)$ is strongly continuous in $[\varepsilon,\infty)$ $(\varepsilon > 0)$.

Now, by $\int_\varepsilon^t \|h(\tau)\|_1 d\tau \le \int_\varepsilon^t \|\mu\|_{\mathcal{M}} d\tau < \|\mu\|_{\mathcal{M}}\,t$, it follows that $h(t)$ is Bochner integrable on $(0,t]$ for any $t > 0$. Similar with the proof of (23), for $k = 0, 1, 2, \dots$, $Y_k\left(\int_0^t\cdots\int_0^t\left(\varphi_{p,\,(u_1+u_2+\cdots+u_r)}^\gamma * d\mu\right)\,du_1\cdots du_r\right) = Y_k\left(\oplus^r T_p^\gamma(t)f - f\right)$, hence,

$\oplus^r T_p^\gamma(t)f - f = \int_0^t \cdots \int_0^t \left(\varphi^\gamma_{p,\,(u_1+u_2+\cdots+u_r)} * d\mu\right) du_1 \cdots du_r$, from which it follows that $\| \oplus^r T_p^\gamma(t)f - f\|_1 \le \|\mu\|_{\mathcal{M}}\, t^r = \mathcal{O}(t^r)$. This completes the proof of Theorem 4.2.4. □

5 Approximation for Generalized Spherical Abel-Poisson and Weierstrass Operators and Their Booleans

We now apply the results of Section 4 to two special operators, the generalized spherical Abel-Poisson operators and the generalized spherical Weierstrass operators.

The generalized spherical Abel-Poisson operators (also called generalized Abel-Poisson singular integrals) in $\mathscr{E}(\mathcal{X})$ are defined as (see [5, P. 43-47])

$$V_t^\gamma f := \sum_{k=0}^{\infty} \exp(-k^\gamma t) Y_k f = f * v_t^\gamma \quad (0 < \gamma \le 1,\ f \in \mathcal{X}),$$

where $\exp(\cdot)$ is the exponential function and $v_t^\gamma(\cos\theta)$ is the kernel that $v_t^\gamma(\cos\theta) = \big(1/|\mathbb{S}^n|\big) \sum\limits_{k=0}^{\infty} \exp(-k^\gamma t) \dfrac{k+\lambda}{\lambda} P_k^\lambda(\cos\theta)$ $(0 \le \theta \le \pi)$. For $\gamma = 1$, set $u = e^{-t}$, one has (see [2, P. 212-213]) $V_t^1 f = \sum\limits_{k=0}^{\infty} u^k Y_k f = \int_{\mathbb{S}^n} \dfrac{1}{\left|\mathbb{S}^n\right|} \dfrac{1-u^2}{(u^2 - 2u(x\cdot y)+1)^{\lambda+1}} f(y)\, d\omega_n(y)$ $(0 \le u < 1)$, which is the classical Abel-Poisson summation on the sphere. For $r \in \mathbb{Z}_+$, the r-th Boolean of V_t^γ is

$$\oplus^r V_t^\gamma f = f - (I - V_t^\gamma)^r f = \sum_{k=0}^{\infty} \left(1 - \left(1 - e^{-k^\gamma t}\right)^r\right) Y_k f \quad (f \in \mathcal{X}).$$

The generalized spherical Weierstrass operators (also called generalized spherical Weierstrass singular integrals) are given by (see [4, P. 84])

$$W_t^\kappa f = \sum_{k=0}^{\infty} \exp\big(-(k(k+2\lambda))^\kappa t\big) Y_k f = f * w_k^\gamma \quad (0 < \kappa \le 1,\ f \in \mathcal{X}),$$

here w_k^γ is the kernel that $w_k^\gamma(\cos\theta) = \big(1/|\mathbb{S}^n|\big) \sum\limits_{k=0}^{\infty} \exp\big(-(k(k+2\lambda))^\kappa t\big) \dfrac{k+\lambda}{\lambda} P_k^\lambda(\cos\theta)$ $(0 \le \theta \le \pi)$. For $r \in \mathbb{Z}_+$, the r-th Boolean of W_t^κ is

$$\oplus^r W_t^\kappa f = f - (I - W_t^\kappa)^r f = \sum_{k=0}^{\infty} \left(1 - \left(1 - e^{-(k(k+2\lambda))^\kappa t}\right)^r\right) Y_k f \quad (f \in \mathcal{X}).$$

The kernel of V_t^γ and W_t^κ are both positive, that is, for $0 \le \theta \le \pi,\ t > 0,\ 0 < \gamma \le 1\ 0 < \kappa \le 1$, $v_t^\gamma(\cos\theta) \ge 0$ and $w_t^\kappa(\cos\theta) \ge 0$ which were proved in [4], [5, P. 43-47] and [21]. Therefore, by Theorem 4.2.3, one has the follow lemma.

Lemma 5.1 *V_t^γ and W_t^κ are both regular exponential-type multiplier operators with positive kernels and with $p(x) = x$ and $p(x) = x(x+2\lambda)$ respectively and are both strongly continuous semigroups of contraction operators of class $(\mathscr{C}_0)$ and satisfy the Bernstein-type inequality* (35).

Now we prove the equivalence between the approximation of the two operators and moduli of smoothness on the sphere.

Theorem 5.3 *Let $\{V_t^\gamma | 0 \le t < \infty\}$ $(0 < \gamma \le 1)$ be generalized spherical Abel-Poisson operators in $\mathscr{E}(\mathcal{X})$. Then for any $0 < \gamma \le 1$ and $r \in \mathbb{Z}_+$, there holds for all $f \in \mathcal{X}$ that*

$$\|\oplus^r V_t^\gamma f - f\|_{\mathcal{X}} \approx \omega^{r\gamma}(f, t^{1/\gamma})_{\mathcal{X}}.$$

Proof. Set $\mathcal{V}^\gamma$ the infinitesimal generator of $\{V_t^\gamma | 0 \le t < \infty\}$. By (34), for $(\mathcal{V}^\gamma)^r f \in \mathcal{X}$ and $r \in \mathbb{Z}_+$, there holds $(\mathcal{V}^\gamma)^r f \sim \sum\limits_{k=0}^{\infty} (-k^\gamma)^r Y_k f$. By Lemma 5.1 and Theorem 4.2.4, $\|\oplus^r V_t^\gamma f - f\|_{\mathcal{X}} \approx K_{(\mathcal{V}^\gamma)^r}(f, t^r)_{\mathcal{X}}$.

In Theorem 3.3, set $a(x) = (-1)^{2/\gamma} x^2$, $b(x) = -x(x+2\lambda)$ and $\alpha = (r\gamma)/2$, as $\lim\limits_{x\to\infty} \big((-1)^{2/\gamma} a(x)\big)/\big((-1)b(x)\big) = 1$, $a(0) = b(0) = 0$ and $g(t) = (1+2\lambda t)$, $(g(t))^{-1} = (1+2\lambda t)^{-1} \in C^{(2\lambda+2)}[0, +\infty)$, we obtain that $K_{(\mathcal{V}^\gamma)^r}(f,t)_{\mathcal{X}} \approx K_{D^{\frac{r\gamma}{2}}}(f,t)_{\mathcal{X}}$ $(t > 0)$, then with (14) one obtains $\|\oplus^r V_t^\gamma f - f\|_{\mathcal{X}} \approx K_{(\mathcal{V}^\gamma)^r}(f,t^r)_{\mathcal{X}} \approx K_{D^{\frac{r\gamma}{2}}}(f,t^r)_{\mathcal{X}} \approx \omega^{r\gamma}(f, t^{1/\gamma})_{\mathcal{X}}$. The proof of Theorem 5.3 is completed. □

We have the following similar result for generalized spherical Weierstrass operators.

Theorem 5.4 *Let $\{W_t^\kappa | 0 \le t < \infty\}$ $(0 < \kappa \le 1)$ be generalized spherical Weierstrass operators in $\mathscr{E}(\mathcal{X})$. Then for any $0 < \kappa \le 1$ and $r \in \mathbb{Z}_+$, there holds for all $f \in \mathcal{X}$ that*

$$\|\oplus^r W_t^\kappa f - f\|_{\mathcal{X}} \approx \omega^{2r\kappa}(f, t^{1/(2\kappa)})_{\mathcal{X}}.$$

Finally, we discuss the saturation properties for the Booleans of V_t^γ and W_t^κ.

Theorem 5.5 *For $r \in \mathbb{Z}_+$, $t > 0$, $\oplus^r V_t^\gamma$ $(0 < \gamma \le 1)$ and $\oplus^r W_t^\kappa$ $(0 < \kappa \le 1)$, the following statements are true.*

(i) $\oplus^r V_t^\gamma$ is saturated with $\mathcal{O}(t^r)$ and its saturation class is $\mathcal{H}_1(-k^{r\gamma}; \mathcal{X})$;

(ii) $\oplus^r W_t^\kappa$ is saturated with $\mathcal{O}(t^r)$ and its saturation class is $\mathcal{H}_1(-(k(k+2\lambda))^{r\kappa}; \mathcal{X})$;

(iii) $\mathcal{H}_1(-k^{r\gamma}; \mathcal{X}) = \mathcal{H}_1(-(k(k+2\lambda))^{r\kappa}; \mathcal{X})$ if $0 < \gamma = 2\kappa \le 1$.

Proof. (i) and (ii) are deduced from Remark 5.1 and Theorem 4.2.4. (iii) follows from Remark 3.2 by setting $\psi_0(x) = (-1)^{1/(r\kappa)} x^2$, $\varphi_0(x) = (-1)^{1/(r\kappa)} x(x+2\lambda)$ and $s = r\kappa$. □

References

[1] R. Askey, S. Wainger, On the behavior of special classes of ultraspherical expansions, I, II, J. d'Analyse Math., 15 (1965), I, 193-220; II, 221-244.

[2] H. Berens, P. L. Butzer, S. Pawelke, Limitierungsverfahren von reihen mehrdimensionaler kugelfunktionen und deren saturationsverhalten, Publ. Res. Inst. Math. Sci. Ser. A, 4(2) (1968), 201-268.

[3] H. Berens, L. Q. Li, On the de la Vallée-Poussin means on the sphere, Results in Math., 24(1-2) (1993), 12-26.

[4] S. Bochner, Quasi analytic functions, Laplace operator, positive kernels, Ann. Math., 51(1) (1950), 68-91.

[5] S. Bochner, Sturm-Liouville and heat equations whose eigenfunctions are ultraspherical polynomials or associated Bessel functions, Proceedings of the Conference on Differential Equations, 23-48. University of Maryland, 1955.

[6] P. L. Butzer, H. Berens, Semi-groups of operators and approximation, Springer, Berlin, 1967.

[7] F. Dai, Some equivalence theorems with K-functionals, J. Approx. Theory, 121 (2003), 143-157.

[8] F. Dai, Z. Ditzian, Strong converse inequality for Poisson sums, Proc. Amer. Math. Soc., 133(9) (2005), 2609-2611

[9] F. Dai, K. Y. Wang, Convergence rate of spherical harmonic expansions of smooth funcion, J. Math. Anal. Appl., 348 (2008), 28-33.

[10] F. Dai, H. P. Wang, Positive cubature formulas and Marcinkiewicz-Zygmund inequality on spherical caps, Constr. Approx., 31 (2010), 1-36.

[11] F. Dai, Y. Xu, Moduli of smoothness and approximation on the unit sphere and the unit ball, Adv. Math., (2010), doi:10.1016/j.aim.2010.01.001.

[12] Z. Ditzian, Fractional derivatives and best approximation, Acta Math. Hungar., 81(4) (1998), 323-348.

[13] Z. Ditzian, Approximation on Banach spaces of functions on the sphere, J. Approx. Theory, 140 (2006), 31-45.

[14] Z. Ditzian, K. Ivanov, Strong converse inequalities, J. d'Analyse Math., 61 (1993), 61-111.

[15] C. F. Dunkl, Operators and harmonic analysis on the sphere, Trans. Amer. Math. Soc., 125(2) (1966), 250-263.

[16] J. Favard, Sur l'approximation des fonctions d'une variable reelle, Colloque d'Anal. Harmon. Publ. C. N. R. S., Paris, 15 (1949), 97-110.

[17] Q. T. Le Gia, F. J. Narcowich, J. D. Ward, H. Wendland, Continuous and discrete least-squares approximation by radial basis functions on spheres, J. Approx. Theory, 143 (2006), 124-133.

[18] L. Grafakos, Classical and Modern Fourier Analysis, China Machine Press, Beijing, 2005.

[19] E. Hille, R. S. Phillips, Functional Analysis and Semi-groups, Amer. Math. Soc. Coll. Publ., New York, 31, 1957.

[20] S. Kaczmarz, H. Steinhaus, Theorie der Orthogonalreihen, Warsaw, 1935.

[21] B. Kuttner, On positive Riesz and Abel typical means, Proc. London Math. Soc. Ser., 2(49) (1947), 328-352.

[22] L. Q. Li, R. Y. Yang, Approximation by spherical Jackson polynomials, J. Beijing Normal Univ. Nat. Sci., 27(1) (1991), 1-12 (in Chinese).

[23] S. Riemenschneider, K. Y. Wang, Approximation theorems of Jackson type on the sphere, Adv. Math. (China), 24(2) (1995), 184-186.

[24] G. Szegö, Orthogonal Polynomials, Colloquium Publications, Amer. Math. Soc., 2003.

[25] K. Y. Wang, L. Q. Li, Harmonic analysis and approximation on the unit sphere, Science Press, Beijing, 2006.