# MODERN MATHEMATICAL METHODS

VOLUME III

ISSUE II

MMM

MODERN MATHEMATICAL METHODS

ISSN: 3023-5294

modernmathmeth.com

VOLUME III ISSUE II

August 2025
https://modernmathmeth.com

# MODERN MATHEMATICAL METHODS

## Layout & Language Editor



## Editorial Staff

# Contents

MODERN MATHEMATICAL METHODS
**3** (2025), No. 2, pp. 57-71
https://modernmathmeth.com/
ISSN 3023 - 5294

*Research Article*

# Microwave remote sensing of soil moisture, above ground biomass and freeze-thaw dynamic: Modeling and empirical approaches

LAURA ANGELONI, DOMENICO DANIELE BLOISI, PAOLO BURGHIGNOLI, DAVIDE COMITE, DANILO COSTARELLI, MICHELE PICONI*, ANNA RITA SAMBUCINI, ALESSIO TROIANI, AND ALESSANDRO VENERI

ABSTRACT. Human actions have accelerated changes in global temperature, precipitation patterns, and other critical Earth systems. Key markers of these changes can be linked to the dynamic of Essential Climate Variables (ECVs) and related measures, such as Soil Moisture (SM), Above Ground Biomass (AGB), and Freeze-Thaw (FT) Dynamics. ECVs are crucial for understanding global climate changes, including hydrological and carbon cycles. Moreover, monitoring ECVs helps to validate climate models and inform policy decisions. Monitoring activities can be carried out at a global scale by using technologies like microwave remote sensing. However, other than proper technological developments, the study of ECVs requires suitable theoretical retrieval tools, which leads to the solutions of inverse problems. In this survey, we analyze and summarize the main retrieval techniques available in the literature for SM, AGB, and FT, performed on data collected with microwave remote sensing sensors. Furthermore, we present the project *RETINA* (*REmote sensing daTa INversion with multivariate functional modeling for essential climAte variables characterization*), recently funded by the European Union under the Italian National Recovery and Resilience Plan of NextGenerationEU, under the Italian Ministry of University and Research. The main goal of *RETINA*, is to create innovative techniques for analyzing data generated by the interaction of electromagnetic waves with the Earth's surface, applying theoretical insights to address real-world challenges.

**Keywords:** Microwave remote sensing, essential climate variables, probabilistic cellular automata, neural network operators, Bayesian inversion.

**2020 Mathematics Subject Classification:** 47A58, 47A63, 47A57, 41A25, 41A05.

## 1. INTRODUCTION

Human activities such as burning fossil fuels, deforestation, agriculture, and industrial processes are responsible for releasing significant amounts of carbon dioxide ($CO_2$), methane ($CH_4$) and other greenhouse gases, driving the rapid modifications that the Earth's climate is experiencing. These actions have caused long-term changes in temperature, precipitation patterns, and other Earth system dynamics.

Key markers of these changes include variations in Essential Climate Variables (ECVs) such as Soil Moisture (SM), Above Ground Biomass (AGB), and Freeze-Thaw (FT) Dynamics. SM dynamics, a crucial part of the global hydrological cycle, are impacted by human-induced climate changes. These dynamics affect water availability, agricultural productivity, and even natural disaster patterns (e.g., droughts and floods). Large-scale deforestation reduces the

Received: 19.03.2025; Accepted: 26.06.2025; Published Online: 29.07.2025
*Corresponding author: Michele Piconi; michele.piconi@unipg.it
DOI: 10.64700/mmm.54
https://gcos.wmo.int/en/essential-climate-variables/about


FIGURE 1. The official logo of the *RETINA* project

Earth's capacity to sequester carbon, directly impacting carbon cycles and exacerbating climate change. AGB is a critical measure to monitor carbon stocks and understand the effects of deforestation. Finally, the FT dynamics in polar and boreal regions are influenced by human activity emissions, accelerating permafrost melting and methane release, both of which have significant implications for climate change.

A central role to track these changes is played by Earth observation technologies such as Microwave Remote Sensing (MW RS). The monitoring of variables, like SM, AGB, and FT, helps to validate climate models, understand the feedback mechanisms between human activity and environmental responses, and inform policies aimed at mitigating human impacts on the climate. MW RS utilizes active sensors (e.g., radar) and passive sensors (e.g., radiometers) for continuous monitoring, irrespective of weather or lighting conditions. Depending on the platform (i.e., airborne or spacebone), data collection is possible on regional and global scales [45].

The retrieval techniques in MW RS data are based on theoretical [11], semi-empirical [39], and empirical [46] models, increasingly enhanced by modern machine learning (ML) techniques using neural networks (NN). Theoretical models leverage classical electrodynamics theories, such as, e.g., scattering theory from rough surfaces [45], Radiative Transfer Theory (RTT) [14], Foldy-Distorted Born Approximation (DBA) [40]. Semi-empirical models, such as the Water Cloud Model (WCM) [10], links microwave signals to soil and vegetation parameters. Empirical models, instead, use direct relationships between observed signals and ECVs, and are therefore driven by the physical observable characterized by the microwave sensor. ML is applied to handle complex datasets for more accurate predictions. However, the data sets required for the training are often very large; therefore, significant effort is needed for data annotation. In addition, data and ancillary data are often not continuously available due to acquisition methods, the type of sensor, the spatial and temporal resolution, as well as some practical conditions (e.g., the satellite orbit and the presence of disturbances like clouds).

Very recently, a new research project that aims to propose new methods for the retrieval of the ECVs has been funded, mixing both deterministic and nondeterministic procedures. This endeavour *RETINA* (see Fig. 1, in which the official logo of the *RETINA* project is shown), is funded by the European Union within the framework of the Italian National Recovery and Resilience Plan (NRRP) of the NextGenerationEU program, under the Italian Ministry of University and Research. *RETINA* proposes, for the first time, the application of direct and inverse analytical methods of Approximation Theory based on the theory of the so-called multivariate neural network (NN) operators (see, e.g., [16, 18, 19]) for the modeling and estimation of

SM, AGB, and FT, using data from space missions. In the context of analytical methods in Approximation Theory, we can mention the following significant contributions [1]-[4].

The fundamental idea in RETINA is to combine analytic inversion techniques (based on functional analysis tools, such as series expansions [33]) and Bayesian approaches performed in conjunction with Monte Carlo methods. The main two (complementary) strategies of *RETINA* can be summarized as follows:

(1) Data Modeling with well-known Multivariate NN Operators: Through functional analysis techniques, theoretical inversion of these operators is achieved, resulting in an approximate analytical model for the target geophysical variables that is useful for their estimation. To address potential data disturbances, as well as take into account the uncertainty of the model, the NN operators will be extended to have the possibility of representing interval-valued fuzzy sets (IVFS), which allow for the representation of (uncertain variables) situations that are more coherent with real-world situations.
(2) Bayesian Inversion with Monte Carlo Markov Chain (MCMC): These methods are used to sample from the posterior distribution, a robust technique for Bayesian inversion. Through MCMC techniques, Bayesian inversion complemented the NN operator approach. *RETINA* targets the introduction of a specific type of Markov Chain, Probabilistic Cellular Automata (PCA), characterized by a parallel updating rule, which is expected to be particularly effective for retrieving multi-component physical quantities, such as matrix-formatted data.

To set the state of the art of the retrieval techniques, a part of the *RETINA* project explores and summarizes the main algorithms available in the literature, focusing on the retrieval methods applied to the characterization of the bio-geophysical variables of interest. Providing an overview of these algorithms is the main motivation for this paper.

The remainder of this paper is organized as follows. Section 2 present a description of existing approaches for monitoring bio-geophysical variables by using Microwave Remote Sensing. Section 3 presents a freely accessible dataset of remote sensing data that can be used for training, testing, and benchmarking retrieval procedures for ECVs. Sections 4, 5, 6 describe recent techniques for the retrieval of vegetation biomass, freeze-thaw, and soil moisture, respectively. Section 7 discusses potential future developments of the theory. Finally, conclusions are drawn in Section 8.

## 2. Microwave Remote Sensing for Monitoring Bio-geophysical Variables

MW RS is a powerful tool for monitoring bio-geophysical variables such as SM, AGB, and FT. This is given by capability to operate under all weather conditions, during day and night, and by the possibility of penetrating clouds, rain, and vegetation. Although often characterized by low spatial resolution (i.e., with respect to optical sensors), MW RS can generally offer relatively high temporal resolution with revisit time in the order of days, and up to hours in certain specific conditions (e.g., in the presence of satellite constellations).

Microwave signals are sensitive to the dielectric properties of soil, water, and vegetation, and are capable of penetrating surface layers (in the order of centimeters, depending on the wavelength and on the structure of the media), making them suitable for observing both superficial and shallow phenomena.

Active sensors, including Synthetic Aperture Radar (SAR) systems, emit microwave radiation and collect the reflected or scattered signals. Passive sensors, such as radiometers, measure naturally emitted thermal radiation. MW radiation wavelengths range from 1 mm to 1 meter and are divided into different frequency bands (e.g., L-band, C-band, X-band and P-band),

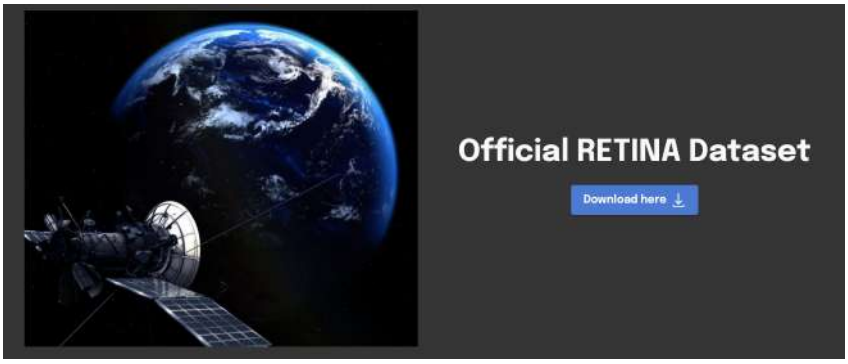


FIGURE 2. A screenshot of the *RETINA* dataset web page, available at the following link: `https://retina.sites.dmi.unipg.it/dataset.html`

which correspond to decreasing wavelength. In MW RS, L-band (15–30 cm wavelength) and C-band (4–8 cm wavelength) are commonly used thanks to their balance between penetration depth, spatial resolution, and the technological maturity of the system. Longer wavelengths (such as L-band or P-band, the latter planned to be used in the future ESA's Biomass mission [50]) are preferred for dense vegetation and high-biomass areas, even though they may suffer signal saturation. On the other hand, shorter wavelengths (e.g., C-band) are effective for less dense biomass or canopy surface and agricultural observations. Missions like SMAP [25] and SMOS [35] use both active and passive microwave sensors to monitor SM globally. Moreover, microwave sensors like AMSR-E [49] and CryoSat [60] were pivotal in polar studies and provided important data to help tracking sea surface temperatures, sea ice extent, and thickness.

Even though MWRS is a fundamental tool for measuring and monitoring many ECVs, performing these tasks is not straightforward. Among others, the following challenges have to be addressed: signal saturation that occurs in dense forests or very high biomass regions where backscatter no longer increases with increasing biomass; spatial resolution that can be improved by combining microwave data with optical, thermal, and in-situ observations for a more comprehensive understanding of Earth's systems; difficulty in separating contributions from soil, vegetation, and atmospheric layers in mixed environments. To overcome these limitations, satellite missions like ESA's Biomass [50], NISAR [22], and HydroGNSS [58] aim to improve data resolution, and extend observational capabilities. Moreover, alongside model-based approaches, AI techniques are increasingly used to analyze complex microwave datasets to obtain more accurate predictions and insights.

## 3. RETINA DATASET OF MICROWAVE REMOTE SENSING DATA

One of the main goals of the *RETINA* project is to release a freely accessible dataset of remote sensing data that can be used for training, testing, and benchmarking retrieval procedures for the considered ECVs. All the selected images have been collected in a dedicated open-access repository available from the *RETINA* website (see Fig. 2) at the following link: `https://retina.sites.dmi.unipg.it/dataset.html`.

The source of all the images (which can be downloaded for free, as stated in the copyright section of this paper) is the Sentinel-1 ([43, 56]) satellite constellation, the first of the ESA's Copernicus Program. The images are available in GeoTIFF format with VV and VH bands. To handle these images, it is possible to use any GIS software, such as QGIS or SNAP. Examples of RS images that can be found in the *RETINA* dataset are shown in Fig. 3.

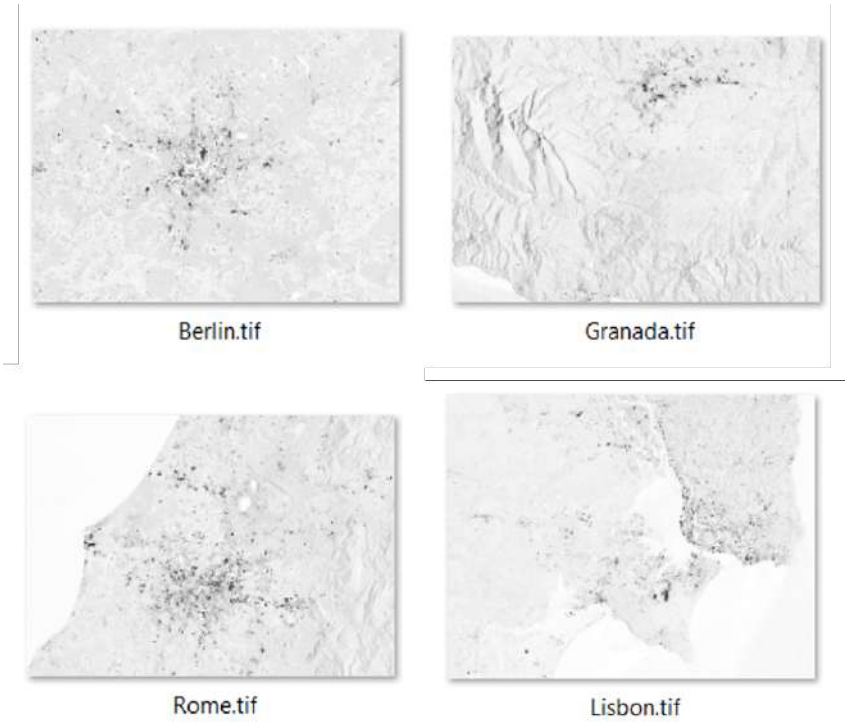


FIGURE 3. Four RS images extracted from the *RETINA* dataset. On the top (from left to right): Berlin (Germany) and Granada (Spain) areas. On the bottom: (from left to right): Rome (Italy) and Lisbon (Portugal) areas

To help users of the dataset, a Python script for reading the images is available on the *RETINA* website. The script allows the user to clip the pixel intensities between user-chosen minimum and maximum threshold values to customize how images are displayed.

Fig. 4 shows the functional scheme of the Python script for reading the .tif files in the Official RETINA Dataset. The input file is processed to produce two new images, one for Band 1 (clipped and normalized) and one for Band 2 (clipped and normalized). In addition to the Python script, also a MATLAB script is available on the *RETINA* website.

Below, we provide a short review of the main algorithms available in the literature for each one of the target ECVs considered in *RETINA*.

## 4. VEGETATION BIOMASS RETRIEVAL: AN OVERWIEW

Retrieving vegetation biomass, specifically AGB, is crucial for understanding carbon cycles, forest dynamics, and the role of vegetation in climate change mitigation. MW RS, particularly radar systems, has emerged as a key method for estimating AGB due to its ability to penetrate vegetation layers and provide detailed structural information [30].

Longer Wavelengths (L-band, P-band) penetrate deeper into vegetation, interacting with trunks and larger branches, making them ideal for high-biomass regions. Shorter Wavelengths (e.g.,C-band) are sensitive to canopy features like leaves and smaller branches.

Techniques for Biomass Retrieval include:

- Theoretical approaches, such as the Radiative Transfer Theory (RTT) and the Wave Theory. RTT focuses on the principles of energy conservation and the transfer of energy through disordered media (i.e., layers of vegetation, see, e.g., [15, 38]). A notable example is the Michigan Microwave Canopy Scattering model (MIMICS) [57], which

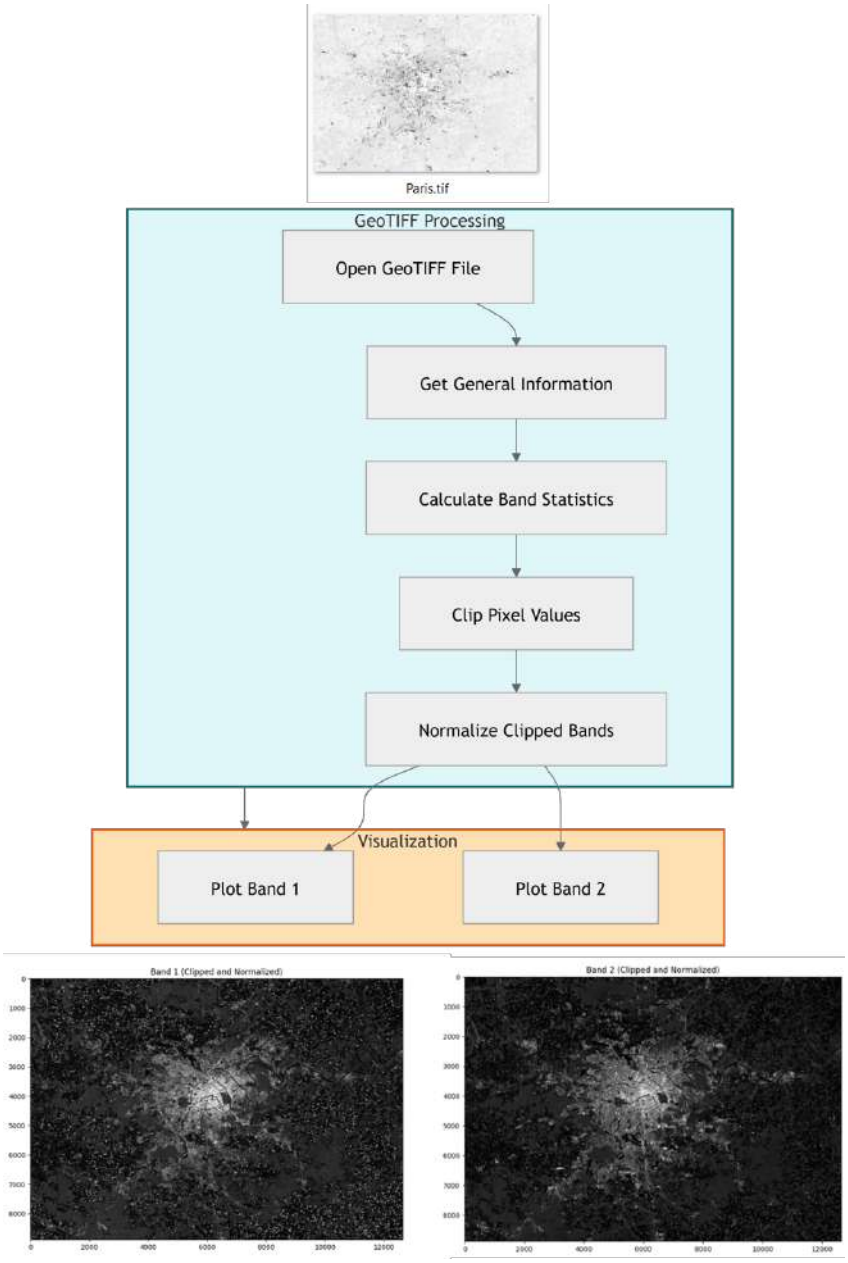


FIGURE 4. The functional scheme of the Python script for reading the Official RETINA Dataset

divides the vegetation into crown, trunk, and ground components, and solves the vector radiative transfer equation (VRTE) iteratively. The Wave Theory, on the other hand, approximates solutions to Maxwell's equations to describe scattered electromagnetic fields. An example of this class is the Distorted Born Approximation (DBA), which uses electromagnetic wave theories of scattering to simulate how microwaves interact with vegetation structures. DBA represents the vegetation layer as a random collection of individual scatterers (see, e.g., [41]). These models are highly accurate but computationally intensive, also requiring a large collection of input data.

- Semi-empirical, such as the Water Cloud Model (WCM), models are simplified models that relate the measured radar backscattering coefficient to biomass using empirical relationships. WCM combines contributions from vegetation and ground backscatter (see, e.g., [42, 47]). These methods are adjusted for vegetation density and structure

using parameters like canopy height and attenuation and are often used with regression techniques to estimate AGB from radar data. As the backscattering coefficient easily saturates at very high biomass densities, limiting the retrieval of AGB beyond a certain threshold, interferometric techniques are also often adopted as coherence appears to be effective to identify forested/nonforested areas and the height of the canopy [8]. The main observable is the complex degree of coherence (CDC), determining the potential of two electromagnetic signals to interfere [12], which can be extracted from the data and compared with models (based on the WCM) accounting for stem volumes, tree height, and fill-factor, i.e., the fraction of ground covered by trees [53].

- ML techniques utilize algorithms like NN and support vector machines to model complex relationships between radar signals and biomass (see [6] for a complete review on this topic). They can integrate multi-source data (e.g., optical and microwave) to improve AGB retrieval accuracy.

SM, surface roughness, and vegetation water content can complicate signal interpretation. In general, retrieval models must account for different vegetation types, structures, and climates to ensure accuracy.

Current and upcoming radar missions for vegetation biomass studies are: ESA Biomass Mission (P-band SAR), designed specifically for global forest AGB mapping; NISAR (L- and S-band SAR), which targets forest structure and biomass dynamics; Sentinel-1 (C-band SAR) that provides data for biomass monitoring with limited penetration depth; GNSS Reflectometry, that is a more recent technique using reflected GPS signals (e.g., CYGNSS, TDS-1) to assess the AGB distribution at global scale ([51]).

## 5. Freeze-Thaw Retrieval: An Overview

FT retrieval focuses on monitoring seasonal transitions in the soil's thermal state, particularly between frozen and unfrozen conditions. These transitions are critical for understanding the water cycle, energy balance, and greenhouse gas dynamics, especially in high-latitude regions where permafrost melting can release significant amounts of methane [29, 48].

MW RS has proven particularly well-suited for FT detection due to its high sensitivity to phase changes of water in soil, as the permittivity of water decreases dramatically between liquid and solid states [45]. Both active and passive MW remote sensing techniques are currently employed in the field, addressing the problem from different perspectives. SAR and radiometer systems (e.g., Sentinel 1 and SMAP) provide high-resolution spatial data by capturing the dynamic temporal variations of the corresponding physical observable, which are associated with seasonal changes in FT states [23, 37]. However, these systems are constrained by limited temporal resolution, typically on the order of days [13]. Passive MW RS instruments, like SMAP, measure naturally emitted microwave radiation, wherein the measured power is expressed in terms of blackbody equivalent radiometric temperature (or radiometer brightness temperature). Radiometers provide FT-related data with high temporal resolution compared to radar systems, $\sim$ 3 days, at the expense of a low spatial resolution [23]. This limitation arises from the relatively weak electromagnetic signal emitted by the natural media, thus requiring a vast observation region to detect meaningful signals. Very recently, GNSS Reflectometry (GNSS-R) has been successfully adopted in the field, addressing the limitations of traditional active and passive measurement techniques and providing sensible data with both high temporal and spatial resolution ([52]). This method involves detecting signals transmitted by a constellation of GNSS satellites and scattered off Earth's surface. Since it exploits signals originally designed for GPS purposes, it is often referred to as a signal of opportunity technique. In

this framework, the measured observable is the surface reflectivity for left-hand circularly polarized fields, related to the dielectric function, in turn connected to climate variables, through the Fresnel's coefficients.

FT retrieval methods include:

- Theoretical models, mostly developed for passive radiometric measurements. This approach aims to estimate the brightness temperature from the physical properties of the medium by means of radiative transport theories. A notable example is the Helsinki University of Technology (HUT) model [44], which solves the scalar radiative transfer equation for multilayered systems.
- Empirical approaches, wherein thresholding is a commonly adopted method. For active measurements, the collected backscattering coefficient is compared with its reference values for the thaw and freeze state, with the goal of determining the present water phase [23]. For GNSS systems the same method is adopted wherein the seasonal threshold algorithm now involves reference reflectivity values and measurements. Radiometric systems also employ a thresholding technique, but in this case, freeze/thaw discrimination is based on the difference between the vertical and horizontal polarizations of the brightness temperature.
- ML algorithms like Random Forest or NN process large datasets to detect patterns in FT transitions. These techniques are useful for integrating multi-sensor and auxiliary data (e.g., temperature and vegetation cover).

Problematics related to FT retrieval can include: vegetation and snow layers can obscure FT signals; rapid transitions may be missed without frequent observations; factors like soil roughness, composition, and moisture can complicate FT retrieval accuracy. New systems like ESA Biomass will enhance FT monitoring capabilities.

## 6. Soil Moisture Retrieval: An Overview

SM is a key bio-geophysical variable influencing global water cycles via evapotranspiration processes, exchange of heat between land and near-surface atmosphere, energy balance, and biochemical/carbon cycles [55]. Recently, MW RS has emerged as a valuable tool for real-time SM monitoring due to its sensitivity to the soil-water ratio, which alters both the medium's emissivity and the backscattering properties of signals, opening the possibility for active and passive RS measurements. Key missions and instruments for SM study are: SMAP that combines SAR and radiometer data for global soil moisture monitoring; SMOS that uses passive L-band radiometer for large-scale SM and salinity observations; Sentinel-1 C-band SAR that provides high-resolution SM data, especially useful for agricultural applications.

Techniques for SM retrieval include:

- Theoretical models are particularly relevant in radar active measurements. Among them, the Integral Equation Method (IEM), based on electromagnetic scattering theory, has gained prominence over time and is now the most widely used [26]. As active measurements are particularly affected by surface roughness conditions, the theory statistically accounts for the random variations of scattering surfaces, estimating the average values of scattered power as a function of surface roughness and the soil dielectric function.
- Semi-empirical models, which are widely preferred due to their effectiveness and relatively simple implementation. For active systems, we distinguish two types of semi-empirical approaches. The first is based on the experimental calibration of the IEM,

which corrects deviation of the theory from measurements correcting roughness effects [9]. The second exploits knowledge of scattering behavior in specific limiting cases, combined with experimental observations, to create ready-to-use formulas derived through data fitting. The most used frameworks are the Oh and the Dubois models [24, 54], both directly relating radar backscattering with volimetric soil moisture and roughness. For passive measurements, semi-empirical models are based on the scalar radiative transfer equation, solved at zeroth order and adjusted using experimental parameters to account for surface roughness, the mixing of different polarization components, and the influence of vegetation and atmospheric layers [36, 59]. Semi-empirical models have also been used for GNSS-based SM retrieval, revealing the potential of this novel technology to provide excellent results even at global scale [7].

- ML and Data Assimilation: ML algorithms analyze complex, multi-sensor datasets to improve accuracy; data assimilation integrates satellite observations with hydrological models for comprehensive SM monitoring. For instance, Convolutional Neural Networks (CNNs) are useful for processing SAR data [32], as they can extract spatial features from radar images, providing enhanced accuracy. More complex NN architecture can describe spatial features along with also temporal evolution.

Dense vegetation and uneven surfaces can obscure SM signals, so ground-truth data are often requested to ensure accuracy in various terrains and climates.

## 7. Future Developments: Neural Network Operators and Bayesian Inversion

NNs have become highly popular due to their utility across numerous fields, including Artificial Intelligence (AI), ML, and Approximation Theory (ATh). Within the *RETINA* project, NNs will be applied in the context of ATh to develop approximate analytical models and their inversions for specific ECVs. In relation to the theory of NNs, in [19], the authors explored the functional properties of the NN operators. This research highlights the potential of these operators in modeling general two-dimensional structures, such as SAR satellite images ([27]) or, more in general, RS data.

For the sake of completeness, we recall the definition of such operators in both their classical and Kantorovich form. The multivariate discrete NN operators can be defined as follows:

$$F_n^d(f,\underline{x}) := \frac{\sum_{\underline{k}\in\mathcal{J}_n} f\left(\frac{\underline{k}}{n}\right)\Psi_\sigma(n\underline{x}-\underline{k})}{\sum_{\underline{j}\in\mathcal{J}_n}\Psi_\sigma(n\underline{x}-\underline{k})}, \quad \underline{x}\in Q^d := [a_1,b_1]\times\cdots\times[a_d,b_d], \tag{7.1}$$

where the function

$$\Psi_\sigma(\underline{x}) := \phi_\sigma(x_1)\cdot\phi_\sigma(x_2)\cdots\phi_\sigma(x_d), \qquad \underline{x} := (x_1,...,x_d)\in\mathbb{R}^s \tag{7.2}$$

is the multivariate (tensor-product) density function defined by means of suitable sigmoidal function $\sigma:\mathbb{R}\to\mathbb{R}$ ([20]), the set of indexes

$$\mathcal{J}_n := \left\{\underline{k}\in\mathbb{Z}^d:\ \lfloor na_i\rfloor\le k_i\le\lceil nb_i\rceil\right\}$$

and

$$\phi_\sigma(x) := \frac{1}{2}\left[\sigma(x+1)-\sigma(x-1)\right], \quad x\in\mathbb{R}.$$

We recall that, by Cybenko's definition in [20], $\sigma:\mathbb{R}\to\mathbb{R}$ is called a *sigmoidal function* if $\lim_{x\to-\infty}\sigma(x)=0$ and $\lim_{x\to+\infty}\sigma(x)=1$.

While, the Kantorovich NN operators are:

$$K_n^d(f,\underline{x}) := \frac{\sum_{\underline{k}\in\mathcal{J}_n}\left[n^d\int_{R_{\underline{k}}^n} f(\underline{u})\,d\underline{u}\right]\Psi_\sigma(n\underline{x}-\underline{k})}{\sum_{\underline{j}\in\mathcal{J}_n}\Psi_\sigma(n\underline{x}-\underline{k})}, \quad \underline{x}\in Q^d, \tag{7.3}$$

where

$$R_{\underline{k}}^n := \left[\frac{k_1}{n},\frac{k_1+1}{n}\right]\times\cdots\times\left[\frac{k_d}{n},\frac{k_d+1}{n}\right] \tag{7.4}$$

are suitable multidimensional rectangles in which we will compute certain mean values of the considered function $f : Q^d \to \mathbb{R}$.

We stress that, with respect to the classical (non-deterministic) theory of shallow and deep NNs, the NN operators are instead widely studied (see. e.g., [16, 17]) mathematical operators that are suitable to pursue a deterministic modeling approach, and also an enhancement/rescaling one.

The task of inverting the above NN operators will be based on an analytical strategy, such as the possibility of exploiting Laurent's series, particularly when working with operators in Hilbert spaces, or methods of Approximation Theory. To address 2D data affected by measurement errors or other disturbances, interval-valued fuzzy sets (IVFS) have been proposed as a robust modeling tool (see [5]).

The NN operators approach will be enhanced through the integration of Bayesian inversion, which leverages advanced Monte Carlo Markov Chain (MCMC) techniques. These new techniques utilize a parallelized transition kernel to enable efficient sampling from the posterior distribution. In the Bayesian framework, the variable to retrieve is treated as a random variable. With this approach, the retrieval procedure aims to determine the probability distribution of this variable given the observed data (posterior probability distribution). Then, the outcome of the retrieval procedure is the value that maximizes the posterior probability distribution. A standard way to estimate such probability distribution is to simulate the evolution of a Markov chain designed so that its stationary distribution is the posterior probability distribution of the variable to retrieve. Since the empirical distribution of the Markov chain converges to its stationary distribution, if the chain is run for a sufficiently long time, its empirical distribution is an estimate of the posterior probability distribution of the variables to retrieve. In this context, algorithms commonly used to simulate the Markov chain include the Metropolis algorithm, the Metropolis-Hastings algorithm, and the Gibbs sampler. If the Markov chain has multi-component states, such as in the case of 2D data, the previous algorithms sample the next state of the chain from a set of neighbors of the current state differing in only one component. This strategy is referred to as single-flip dynamics [31]. An alternative to single-flip dynamics is the use of Probabilistic Cellular Automata (PCA) [28]. In PCA, all components of the state are updated simultaneously and independently at each step. This approach expands the set of neighbors to include the entire state space, resulting in higher motility and potentially faster convergence to equilibrium.

More formally, a PCA, is a Markov chain $(X_n)_{n\in\mathbb{N}}$ defined on $\mathcal{X} = \{1,\dots,k\}^N$, where $N$ is the number of components of the system, and whose transition matrix is such that

$$\mathbb{P}\{X_n = \tau | X_{n-1} = \sigma\} = \prod_{i=1}^{N}\mathbb{P}\{(X_n)_i = \tau_i | X_{n-1} = \sigma\}. \tag{7.5}$$

A transition matrix of this type is obtained by defining a function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of type $H(\sigma, \tau) = -\sum_{1 \le i \le N} h_i(\sigma)\tau_i$ and transition probabilities as

$$\mathrm{P}(\sigma, \tau) = \frac{e^{-\beta H(\sigma,\tau)}}{Z_\sigma} = \prod_{1 \le i \le N} \frac{e^{\beta h_i(\sigma)\tau_i}}{(Z_\sigma)_i} \tag{7.6}$$

where $\beta$ is a positive parameter called the inverse temperature and $Z_\sigma$ is a normalizing constant whose knowledge is not known to simulate the evolution of the chain. Then, if $h(\cdot)$ satisfies certain suitable conditions, the stationary measure of the chain can be proven to be $\pi(\sigma) = \frac{\sum_\tau \mathrm{P}(\sigma,\tau)}{\sum_{\sigma,\tau} \mathrm{P}(\sigma,\tau)}$ [21, 34]. The knowledge of the stationary measure of the chain allows to tune the algorithm so to favor the sampling of the more useful configurations [21, 34].

PCA's inherent parallelism offers significant computational advantages, particularly when leveraging massively parallel processing hardware such as GPUs or TPUs. These processors enable simultaneous updates of all components at each step, greatly enhancing efficiency.

## 8. Conclusions

Climate change is one of the biggest challenges Mankind is called to face. The precise estimation and monitoring of the ECVs on a global scale is a fundamental step to describing and understanding the rapid changes the Earth's climate is experiencing and, consequently, to determine the more appropriate actions to mitigate the adverse effects of these changes.

Microwave remote sensing is a cornerstone of modern Earth observation, enabling critical insights into climate dynamics, environmental monitoring, and resource management at local, regional, and global scales. This paper highlights advancements in microwave sensing technologies and their integration with ML to enhance the monitoring of Earth's bio-geophysical processes.

Particular emphasis is given to those retrieval techniques that will be exploited to estimate the target ECVs in the context *RETINA* project.

## Data Availability Statement

The data made available in the framework of the *RETINA* project can be obtained from the "RETINA dataset" at `https://RETINA.sites.dmi.unipg.it/dataset.html`. *RETINA* Code and documentation are available through the website `https://RETINA.sites.dmi.unipg.it`. Permission is granted to use, copy, or modify the documentation for educational and research purposes without fee.

## Acknowledgments

The authors have been supported within the PRIN 2022 PNRR: "*RETINA*: REmote sensing daTa INversion with multivariate functional modeling for essential climAte variables characterization", funded by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, under the Italian Ministry of University and Research (Project Code: P20229SH29, CUP: J53D23015950001).

## References

[1] U. Abel, A. M. Acu, M. Heilmann and I. Raşa: *On some Cauchy problems and positive linear operators*, Mediterr. J. Math., **22** (1) (2024), 1–19.

[2] A. M. Acu, M. Heilmann, I. Raşa and A. Seserman: *Poisson approximation to the binomial distribution: extensions to the convergence of positive operators*, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math., **117** (4) (2023), 1–13.

[3] A. M. Acu, I. Raşa and A. Seserman: *Positive linear operators and exponential functions*, Math. Found. Comput., **6** (3) (2024), 313–319.
[4] A. M. Acu, I. Rasa and F. Sofonea: *Composition of some positive linear integral operators*, Demonstr. Math., **57** (1) (2024), Article ID: 20240018.
[5] G. Alefeld, G. Mayer: *Interval analysis: theory and applications*, J. Comput. Appl. Math., **121** (1-2) (2000), 421–464.
[6] I. Ali, F. Greifeneder, J. Stamenkovic, M. Neumann and C. Notarnicola: *Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data*, Remote Sens., **7** (12) (2015), 16398–16421.
[7] M. M. Al-Khaldi, J. T. Johnson, A. J. O'Brien, A. Balenzano and F. Mattia: *Time-Series Retrieval of Soil Moisture Using CYGNSS*, IEEE Trans. Geosci. Remote Sensing, **57** (7) (2019), 4322–4331.
[8] J. Askne, P. Dammert, L. M. H. Ulander and G. Smith: *C-band repeat-pass interferometric SAR observations of the forest*, IEEE Trans. Geosci. Remote Sensing, **35** (1997), 25–35.
[9] N. Baghdadi, C. King, A. Chanzy and J. P. Wigneron: *An empirical calibration of the integral equation model based on SAR data, soil moisture, and surface roughness measurement over bare soils*, Int. J. Remote Sens., **23** (20) (2002), 4325–4340.
[10] E. B'eriaux, F. Waldner, F. Collienne, P. Bogaert and P. Defourny: *Maize leaf area index retrieval from synthetic quad pol sar time series using the water cloud model*, Remote Sens., **7** (12) (2015), 16204–16225.
[11] M. Burgin, D. Clewley, R. M. Lucas and M. Moghaddam: *A generalized radar backscattering model based on wave theory for multilayer multispecies vegetation*, IEEE Trans. Geosc. Remote Sens., **49** (12) (2011), 4832–4845.
[12] R. Carminati, J. C. Schotland: *Principles of Scattering and Transport of Light*, Cambridge University Press, (2021). DOI: 10.1017/9781316544693.
[13] H. Carreno-Luengo, G. Luzi and M. Crosetto: *Sensitivity of CyGNSS Bistatic Reflectivity and SMAP Microwave Radiometry Brightness Temperature to Geophysical Parameters Over Land Surfaces*, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., **12** (1) (2019), 107–122.
[14] S. Chandrasekhar: *Radiative Transfer*, Dover Books on Physics, Dover Publications, New York (2013).
[15] K. Chen, S. Tan: *A Multiple-Scattering Microwave Radiative Transfer Model for Land Emission with Vertically Heterogeneous Vegetation Coverage*, IEEE Trans. Geosc. Remote Sens., **62** (2024), 1–14.
[16] L. Coroianu, D. Costarelli, M. Natale and A. Pantis: *The approximation capabilities of Durrmeyer-type neural network operators*, J. Appl. Math. Comput., **70** (2024), 4581–4599.
[17] D. Costarelli: *Density results by deep neural network operators with integer weights*, Math. Modell. Anal., **27** (2022), 547–560.
[18] D. Costarelli, M. Piconi: *Asymptotic analysis for neural network operators employing the Hardy-Littlewood maximal inequality*, Mediterr. J. Math., **21** (2024), Article ID: 199.
[19] D. Costarelli, R. Spigler: *Multivariate neural network operators with sigmoidal activation functions*, Neural Networks, **48** (2013), 72–77.
[20] G. Cybenko: *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Systems, **2** (1989), 303–314.
[21] P. Dai Pra, B. Scoppola and E. Scoppola: *Sampling from a Gibbs measure with pair interaction by means of PCA*, J. Stat. Phys., **149** (2012), 722–737.
[22] A. Das, R. Kumar and P. Rosen: *NISAR mission overview and updates on ISRO science plan*, IEEE Inter. India Geosc. Remote Sens. Symposium (InGARSS), (2021), 269–272.
[23] C. Derksen, X. Xu, R. S. Dunbar, A. Colliander, Y. Kim, J. S. Kimball, T. A. Black, E. Euskirchen, A. Langlois, M. M. Loranty, P. Marsh, K. Rautiainen, A. Roy, A. Royer and J. Stephens: *Retrieving landscape freeze/thaw state from Soil Moisture Active Passive (SMAP) radar and radiometer measurements*, Remote Sens. Environ., **194** (2017), 48–62.
[24] P. C. Dubois, J. van Zyl and T. Engman: *Measuring soil moisture with imaging radars*, IEEE Trans. Geosc. Remote Sens., **33** (4) (1995), 915–926.
[25] D. Entekhabi, E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, S. D. Goodman, T. J. Jackson, J. Johnson et al.: *The soil moisture active passive (SMAP) mission*, Proc. IEEE, **98** (5) (2010), 704–716.
[26] A. K. Fung, Z. Li and K. S. Chen: *Backscattering from a randomly rough dielectric surface*, IEEE Trans. Geosc. Remote Sens., **30** (2) (1992), 356–369.
[27] L. Giustarini, R. Hostache, P. Matgen, G. J. P. Schumann, P. D. Bates and D. C. Mason: *A change detection approach to flood mapping in urban areas using TerraSAR-X*, IEEE Trans. Geosc. Remote Sens., **51** (4) (2012), 2417–2430.
[28] S. Goldstein, R. Kuik, J. L. Lebowitz and C. Maes: *From PCA's to equilibrium systems and back*, Comm. Math. Phys., **125** (1) (1989), 71–79.
[29] M. Goulden, S. Wofsy, J. Harden, S. Trumbore, P. Crill, S. Gower, T. Fries, B. Daube, S.-M. Fan, D. J. Sutton, A. Bazzaz and J. Munger: *Sensitivity of Boreal forest carbon balance to soil thaw*, Science, **279** (1998), 214–217.

[30] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice and J. R. G. Townshend: *High-resolution global maps of 21st-century forest cover change*, Science, **342** (6160) (2013), 850–853.

[31] W. K. Hastings: *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, **57** (1) (1970), 97–109.

[32] E. Hegazi, L. Yang and J. Huang: *A convolutional neural network algorithm for soil moisture prediction from Sentinel-1 SAR images*, Remote Sens., **13** (24) (2021), Article ID: 4964.

[33] P. Howlett, K. Avrachenkov, C. Pearce and V. Ejov: *Inversion of analytically perturbed linear operators that are singular at the origin*, J. Math. Anal. Appl., **353** (1) (2009), 68–84.

[34] M. Isopi, B. Scoppola and A. Troiani: *On some features of quadratic unconstrained binary optimization with random coefficients*, Boll. Unione Mat. Ital., **18** (2024), 615–635.

[35] Y. H. Kerr, P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M.-J. Escorihuela, J. Font, N. Reul, C. Gruhier, S. E. Juglea, M. R. Drinkwater, A. Hahne, M. Martín-Neira and S. Mecklenburg: *The SMOS mission: New tool for monitoring key elements of the global water cycle*, Proc. IEEE, **98** (5) (2010), 666–687.

[36] L. Karthikeyan, M. Pan, N. Wanders, D. N. Kumar and E. F. Wood: *Four decades of microwave satellite soil moisture observations: Part 1. A review of retrieval algorithms*, Adv. Water Resour., **109** (2017), 106–120.

[37] J. S. Kimball, K. C. McDonald, S. Frolking and S. W. Running: *Radar remote sensing of the spring thaw transition across a boreal landscape*, Remote Sens. Environ., **89** (2) (2004), 163–175.

[38] M. Kurum, R. H. Lang, P. E. O'Neill, A. T. Joseph, T. J. Jackson and M. H. Cosh: *A first-order radiative transfer model for microwave radiometry of forest canopies at L-Band*, IEEE Trans. Geosci. Remote Sens., **49** (9) (2011), 3167–3179.

[39] L. Kurvonen, J. Pulliainen and M. Hallikainen: *Retrieval of biomass in boreal forests from multitemporal ERS-1 and JERS-1 SAR images*, IEEE Trans. Geosci. Remote Sens., **37** (1) (1999), 198–205.

[40] R. H. Lang: *Electromagnetic backscattering from a sparse distribution of lossy dielectric scatterers*, Radio Sci., **16** (1) (1981), 15–30.

[41] R. H. Lang, A. Sharma and M. Cosh: *Scattering from a layer of vegetation: Enhancement effects*, 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, (2017), 1419–1421.

[42] J. Li, S. Wang: *Using SAR-derived vegetation descriptors in a water cloud model to improve soil moisture retrieval*, Remote Sens., **10** (9) (2018), Article ID: 1370.

[43] J. Liang, D. Liu: *A local thresholding approach to flood water delineation using Sentinel-1 SAR imagery*, ISPRS J. Photogrammetry and Remote Sens., **159** (2020), 53–62.

[44] J. Lemmetyinen, J. Pulliainen, A. Rees, A. Kontu, Y. Qiu and C. Derksen: *Multiple-Layer Adaptation of HUT Snow Emission Model: Comparison With Experimental Data*, IEEE Trans. Geosci. Remote Sens., **48** (7) (2010), 2781–2794.

[45] D. Long, F. Ulaby: *Microwave radar and radiometric remote sensing*, Artech House, (2015).

[46] R. M. Lucas, N. Cronin, A. Lee, M. Moghaddam, C. Witte and P. Tickle: *Empirical relationships between AIRSAR backscatter and LiDAR-derived forest biomass, Queensland, Australia*, Remote Sens. Environ., **100** (3) (2006), 407–425.

[47] G. S. Nijaguna, D. R. Manjunath, M. Abouhawwash, S. S. Askar, D. K. Basha and J. Sengupta: *Deep learning-based improved WCM technique for soil moisture retrieval with satellite images*, Remote Sens., **15** (8) (2023).

[48] B. Panneer Selvam, H. Laudon, F. Guillemette and M. Berggren: *Influence of soil frost on the character and degradability of dissolved organic carbon in boreal forest soils*, Journal of Geophysical Research: Biogeosciences, **121** (3) (2016), 829–840.

[49] C. L. Parkinson: *Aqua: An Earth-observing satellite mission to examine water and other climate variables*, IEEE Trans. Geosci. Remote Sens., **41** (2) (2003), 173–183.

[50] S. Quegan, T. Le Toan, J. Chave, J. Dall, J.-F. Exbrayat, D. H. T. Minh, M. Lomas, M. M. D'alessandro, P. Paillou, K. Papathanassiou et al.: *The European Space Agency Biomass Mission: Measuring forest above-ground biomass from space*, Remote Sens. Environ., **227** (2019), 44–60.

[51] C. S. Ruf, C. Chew, T. Lang, M. G. Morris, K. Nave, A. Ridley and R. Balasubramaniam: *A new paradigm in Earth environmental monitoring with the CYGNSS small satellite constellation*, Sci. Rep., **8** (1) (2018), 1–13, Article ID: 8782.

[52] N. Rodriguez-Alvarez, A. Camps, M. Vall-llossera, X. Bosch-Lluis, A. Monerris, I. Ramos-Perez, E. Valencia, J. F. Marchan-Hernandez, J. Martinez-Fernandez, G. Baroncini-Turricchia, C. Pérez-Gutiérrez and N. Sánchez: *Land geophysical parameters retrieval using the interference pattern GNSS-R technique*, IEEE Trans. Geosci. Remote Sens., **49** (1) (2010), 71–84.

[53] M. Santoro, J. Askne, G. Smith and J. E. S. Fransson: *Stem volume retrieval in boreal forests from ERS-1/2 interferometry*, Remote Sens. Environ., **81** (1) (2002), 19–35.

[54] Y. Oh, K. Sarabandi and F. T. Ulaby: *An empirical model and an inversion technique for radar scattering from bare soil surfaces*, IEEE Trans. Geosci. Remote Sens., **30** (2) (1992), 370–381.

[55] S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky and A. J. Teuling: *Investigating soil moisture–climate interactions in a changing climate: A review*, Earth-Sci. Rev., **99** (3) (2010), 125–161.
[56] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown et al.: *GMES Sentinel-1 Mission*, Remote Sens. Environ., **120** (2012), 9–24.
[57] F. T. Ulaby, K. Sarabandi, K. McDonald, M. W. Whitt and M. C. Dobson: *Michigan microwave canopy scattering model*, Int. J. Remote Sens., **11** (1990), 1223–1253.
[58] M. J. Unwin, N. Pierdicca, E. Cardellach, K. Rautiainen, G. Foti, P. Blunt, L. Guerriero, E. Santi and M. Tossaint: *An introduction to the HydroGNSS GNSS reflectometry remote sensing mission*, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., **14** (2021), 6987–6999.
[59] J. R. Wang, B. J. Choudhury: *Remote sensing of soil moisture content, over bare field at 1.4 GHz frequency*, Journal of Geophysical Research: Oceans, **86** (C6) (1981), 5277–5282.
[60] D. J. Wingham: *CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields*, Adv. Space Research, **37** (4) (2006), 841–871.

Laura Angeloni
University of Perugia
Department of Mathematics and Computer Science
Via Vanvitelli, 1, Perugia, Italy
*Email address*: laura.angeloni@unipg.it

Domenico Daniele Bloisi
International University of Rome UNINT
Department of International Humanities ans Social Sciences
Via Cristoforo Colombo, 200, Rome, Italy
*Email address*: domenico.bloisi@unint.eu

Paolo Burghignoli
Sapienza University of Rome
Department of Information Engineering, Electronic and Communications
Via Eudossiana, 18, 00184, Roma, Italy
*Email address*: paolo.burghignoli@uniroma1.it

Davide Comite
Sapienza University of Rome
Department of Information Engineering, Electronic and Communications
Via Eudossiana, 18, 00184, Roma, Italy
*Email address*: davide.comite@uniroma1.it

Danilo Costarelli
University of Perugia
Department of Mathematics and Computer Science
Via Vanvitelli, 1, Perugia, Italy
*Email address*: danilo.costarelli@unipg.it

Michele Piconi
University of Perugia
Department of Mathematics and Computer Science
Via Vanvitelli, 1, Perugia, Italy
*Email address*: michele.piconi@unipg.it

Anna Rita Sambucini
University of Perugia
Department of Mathematics and Computer Science
Via Vanvitelli, 1, Perugia, Italy
*Email address*: anna.sambucini@unipg.it

ALESSIO TROIANI
UNIVERSITY OF PERUGIA
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
VIA VANVITELLI, 1, PERUGIA, ITALY
*Email address*: alessio.troiani@unipg.it

ALESSANDRO VENERI
SAPIENZA UNIVERSITY OF ROME
DEPARTMENT OF INFORMATION ENGINEERING, ELECTRONIC AND COMMUNICATIONS
VIA EUDOSSIANA, 18, 00184, ROMA, ITALY
*Email address*: alessandro.veneri@uniroma1.it

MODERN MATHEMATICAL METHODS
**3** (2025), No. 2, pp. 72-84
https://modernmathmeth.com/
ISSN 3023 - 5294

*Research Article*

# Approximation by symmetrized and perturbed hyperbolic tangent activated convolutions as positive linear operators

GEORGE A. ANASTASSIOU*

ABSTRACT. In this work, we studied further the univariate symmetrized and perturbed hyperbolic tangent activated convolution type operators of three kinds. Here, this is done with the method of positive linear operators. Their new approximation properties are established by the quantitative convergence to the unit operator using the modulus of continuity. It is also studied the related simultaneous approximation, as well as the iterated approximation.

**Keywords:** Symmetrized and perturbed hyperbolic tangent, convolution operator, quantitative approximation, simultaneous approximation, iterated approximation.

**2020 Mathematics Subject Classification:** 41A17, 41A25, 41A35.

## 1. INTRODUCTION

The author studied extensively the quantitative approximation of positive linear operators to the unit since 1985, see for example [1]-[3], [8]. He originated from the quantitative weak convergence of finite positive measures to the unit Dirac measure, having as a method the geometric moment theory, see [2], and he produced best upper bounds, leading to attained (i.e. sharp Jackson type inequalities), e.g. see [1], [2]. These studies have been gone to all possible directions, univariate and multivariate, though in this work, we stay only on the univariate approach over an infinite domain.

Our convolution operators here have as a kernel the symmetrized and perturbed hyperbolic tangent activation function, which is used very commonly in the study of neural networks, and they can be interpreted as positive linear operators. So here our proving methods come from the theory of positive linear operators. Thus, in Section 2, we discuss about the symmetrized and perturbed hyperbolic tangent activation function. In Section 3, we describe our activated convolution type operators and we present their properties, such as differentiation and iteration, along with positive linear operators results to be applied. In Section 4, we derive some auxiliary results which are estimates to our operators, when applied to polynomial type functions and to be used into our main results. In Section 5, we present our main explicit results under the lens of positive linear operators theory. We treat also the simultaneous and iterated approximation cases under the same spirit. We are greatly inspired by our earlier works [4, 5, 7]. Furthermore, general motivation comes from the great works [12] and [13].

Received: 11.07.2025; Accepted: 11.08.2025; Published Online: 20.08.2025
*Corresponding author: George A. Anastassiou; ganastss@memphis.edu
DOI: 10.64700/mmm.73

## 2. About $q$-Deformed and $\lambda$-Parametrized Hyperbolic Tangent Function $g_{q,\lambda}$

Here, all this initial background comes from [6, Chapter 18]. We use $g_{q,\lambda}$, see (2.1), and exhibit that it is a sigmoid function and we will present several of its properties related to the approximation by neural network operators. So, let us consider the hyperbolic tangent activation function

$$g_{q,\lambda}(x) := \frac{e^{\lambda x} - qe^{-\lambda x}}{e^{\lambda x} + qe^{-\lambda x}}, \quad \lambda, q > 0, \ x \in \mathbb{R}. \tag{2.1}$$

We have that

$$g_{q,\lambda}(0) = \frac{1-q}{1+q}.$$

We notice also that

$$g_{q,\lambda}(-x) = \frac{e^{-\lambda x} - qe^{\lambda x}}{e^{-\lambda x} + qe^{\lambda x}} = \frac{\frac{1}{q}e^{-\lambda x} - e^{\lambda x}}{\frac{1}{q}e^{-\lambda x} + e^{\lambda x}} = -\frac{\left(e^{\lambda x} - \frac{1}{q}e^{-\lambda x}\right)}{e^{\lambda x} + \frac{1}{q}e^{-\lambda x}} = -g_{\frac{1}{q},\lambda}(x). \tag{2.2}$$

That is

$$g_{q,\lambda}(-x) = -g_{\frac{1}{q},\lambda}(x), \ \forall\, x \in \mathbb{R}, \tag{2.3}$$

and

$$g_{\frac{1}{q},\lambda}(x) = -g_{q,\lambda}(-x),$$

hence

$$g'_{\frac{1}{q},\lambda}(x) = g'_{q,\lambda}(-x). \tag{2.4}$$

It is

$$g_{q,\lambda}(x) = \frac{e^{2\lambda x} - q}{e^{2\lambda x} + q} = \frac{1 - \frac{q}{e^{2lx}}}{1 + \frac{q}{e^{2\lambda x}}} \underset{(x\to+\infty)}{\to} 1,$$

i.e.

$$g_{q,\lambda}(+\infty) = 1. \tag{2.5}$$

Furthermore

$$g_{q,\lambda}(x) = \frac{e^{2\lambda x} - q}{e^{2\lambda x} + q} \underset{(x\to-\infty)}{\to} \frac{-q}{q} = -1,$$

i.e.

$$g_{q,\lambda}(-\infty) = -1. \tag{2.6}$$

We find that

$$g'_{q,\lambda}(x) = \frac{4q\lambda e^{2\lambda x}}{\left(e^{2\lambda x} + q\right)^2} > 0, \tag{2.7}$$

therefore $g_{q,\lambda}$ is strictly increasing. Next, we obtain ($x \in \mathbb{R}$)

$$g''_{q,\lambda}(x) = 8q\lambda^2 e^{2\lambda x}\left(\frac{q - e^{2\lambda x}}{\left(e^{2\lambda x} + q\right)^3}\right) \in C(\mathbb{R}). \tag{2.8}$$

We observe that

$$q - e^{2\lambda x} \gtrless 0 \Leftrightarrow q \gtrless e^{2\lambda x} \Leftrightarrow \ln q \gtrless 2\lambda x \Leftrightarrow x \lessgtr \frac{\ln q}{2\lambda}.$$

So, in case of $x < \frac{\ln q}{2\lambda}$, we have that $g_{q,\lambda}$ is strictly concave up, with $g''_{q,\lambda}\left(\frac{\ln q}{2\lambda}\right) = 0$. In case of $x > \frac{\ln q}{2\lambda}$, we have that $g_{q,\lambda}$ is strictly concave down. Clearly, $g_{q,\lambda}$ is a shifted sigmoid function

with $g_{q,\lambda}(0)=\frac{1-q}{1+q}$, and $g_{q,\lambda}(-x)=-g_{q^{-1},\lambda}(x)$, (a semi-odd function). By $1>-1$, $x+1>x-1$, we consider the function

$$M_{q.\lambda}(x):=\frac{1}{4}\left(g_{q,\lambda}(x+1)-g_{q,\lambda}(x-1)\right)>0, \tag{2.9}$$

$\forall x\in\mathbb{R}$; $q,\lambda>0$. Notice that $M_{q,\lambda}(\pm\infty)=0$, so the $x$-axis is horizontal asymptote. We have that

$$\begin{aligned} M_{q,\lambda}(-x) &= \frac{1}{4}\left(g_{q,\lambda}(-x+1)-g_{q,\lambda}(-x-1)\right)=\frac{1}{4}\left(g_{q,\lambda}(-(x-1))-g_{q,\lambda}(-(x+1))\right) \\ &= \frac{1}{4}\left(-g_{\frac{1}{q},\lambda}(x-1)+g_{\frac{1}{q},\lambda}(x+1)\right)=\frac{1}{4}\left(g_{\frac{1}{q},\lambda}(x+1)-g_{\frac{1}{q},\lambda}(x-1)\right) \\ &= M_{\frac{1}{q},\lambda}(x),\ \ \forall\, x\in\mathbb{R}. \end{aligned} \tag{2.10}$$

Thus

$$M_{q,\lambda}(-x)=M_{\frac{1}{q},\lambda}(x),\ \forall\, x\in\mathbb{R};\ q,\lambda>0, \tag{2.11}$$

a deformed symmetry. Next, we have that

$$M'_{q,\lambda}(x)=\frac{1}{4}\left(g'_{q,\lambda}(x+1)-g'_{q,\lambda}(x-1)\right),\ \ \forall\, x\in\mathbb{R}. \tag{2.12}$$

Let $x<\frac{\ln q}{2\lambda}-1$, then $x-1<x+1<\frac{\ln q}{2\lambda}$ and $g'_{q,\lambda}(x+1)>g'_{q,\lambda}(x-1)$ (by $g_{q,\lambda}$ being strictly concave up for $x<\frac{\ln q}{2\lambda}$), that is $M'_{q,\lambda}(x)>0$. Hence $M_{q,\lambda}$ is strictly increasing over $\left(-\infty,\frac{\ln q}{2\lambda}-1\right)$. Let now $x-1>\frac{\ln q}{2\lambda}$, then $x+1>x-1>\frac{\ln q}{2\lambda}$, and $g'_{q,\lambda}(x+1)<g'_{q,\lambda}(x-1)$, that is $M'_{q,\lambda}(x)<0$. Therefore $M_{q,\lambda}$ is strictly decreasing over $\left(\frac{\ln q}{2\lambda}+1,+\infty\right)$. Let us next consider, $\frac{\ln q}{2\lambda}-1\leq x\leq\frac{\ln q}{2\lambda}+1$. We have that

$$\begin{aligned} M''_{q,\lambda}(x) &= \frac{1}{4}\left(g''_{q,\lambda}(x+1)-g''_{q,\lambda}(x-1)\right) \\ &= 2q\lambda^2\left[e^{2\lambda(x+1)}\left(\frac{q-e^{2\lambda(x+1)}}{\left(e^{2\lambda(x+1)}+q\right)^3}\right)-e^{2\lambda(x-1)}\left(\frac{q-e^{2\lambda(x-1)}}{\left(e^{2\lambda(x-1)}+q\right)^3}\right)\right]. \end{aligned} \tag{2.13}$$

By $\frac{\ln q}{2\lambda}-1\leq x\Leftrightarrow\frac{\ln q}{2\lambda}\leq x+1\Leftrightarrow\ln q\leq 2\lambda(x+1)\Leftrightarrow q\leq e^{2\lambda(x+1)}\Leftrightarrow q-e^{2\lambda(x+1)}\leq 0$. By $x\leq\frac{\ln q}{2\lambda}+1\Leftrightarrow x-1\leq\frac{\ln q}{2\lambda}\Leftrightarrow 2\lambda(x-1)\leq\ln q\Leftrightarrow e^{2\lambda(x-1)}\leq q\Leftrightarrow q-e^{2\lambda\beta(x-1)}\geq 0$. Clearly by (2.13), we get that $M''_{q,\lambda}(x)\leq 0$, for $x\in\left[\frac{\ln q}{2\lambda}-1,\frac{\ln q}{2\lambda}+1\right]$. More precisely $M_{q,\lambda}$ is concave down over $\left[\frac{\ln q}{2\lambda}-1,\frac{\ln q}{2\lambda}+1\right]$, and strictly concave down over $\left(\frac{\ln q}{2\lambda}-1,\frac{\ln q}{2\lambda}+1\right)$.

Consequently, $M_{q,\lambda}$ has a bell-type shape over $\mathbb{R}$. Of course it holds $M''_{q,\lambda}\left(\frac{\ln q}{2\lambda}\right)<0$. At $x=\frac{\ln q}{2\lambda}$, we have

$$M'_{q,\lambda}(x)=\frac{1}{4}\left(g'_{q,\lambda}(x+1)-g'_{q,\lambda}(x-1)\right)=q\lambda\left(\frac{e^{2\lambda(x+1)}}{\left(e^{2\lambda(x+1)}+q\right)^2}-\frac{e^{2\lambda(x-1)}}{\left(e^{2\lambda(x-1)}+q\right)^2}\right). \tag{2.14}$$

Thus

$$M'_{q,\lambda}\left(\frac{\ln q}{2\lambda}\right)=q\lambda\left(\frac{e^{2\lambda\left(\frac{\ln q}{2\lambda}+1\right)}}{\left(e^{2\lambda\left(\frac{\ln q}{2\lambda}+1\right)}+q\right)^2}-\frac{e^{2\lambda\left(\frac{\ln q}{2\lambda}-1\right)}}{\left(e^{2\lambda\left(\frac{\ln q}{2\lambda}-1\right)}+q\right)^2}\right)$$

$$= \lambda \left( \frac{e^{2\lambda} \left(e^{-2\lambda}+1\right)^2 - e^{-2\lambda}\left(e^{2\lambda}+1\right)^2}{\left(e^{2\lambda}+1\right)^2 \left(e^{-2\lambda}+1\right)^2} \right) = 0. \tag{2.15}$$

That is, $\frac{\ln q}{2\lambda}$ is the only critical number of $M_{q,\lambda}$ over $\mathbb{R}$. Hence at $x = \frac{\ln q}{2\lambda}$, $M_{q,\lambda}$ achieves its global maximum, which is

$$M_{q,\lambda}\left(\frac{\ln q}{2\lambda}\right) = \frac{1}{4}\left[g_{q,\lambda}\left(\frac{\ln q}{2\lambda}+1\right) - g_{q,\lambda}\left(\frac{\ln q}{2\lambda}-1\right)\right] = \frac{1}{4}\left[\left(\frac{e^{\lambda}-e^{-\lambda}}{e^{\lambda}+e^{-\lambda}}\right) - \left(\frac{e^{-\lambda}-e^{\lambda}}{e^{-\lambda}+e^{\lambda}}\right)\right]$$

$$= \frac{1}{4}\left[\frac{2\left(e^{\lambda}-e^{-\lambda}\right)}{e^{\lambda}+e^{-\lambda}}\right] = \frac{1}{2}\left(\frac{e^{\lambda}-e^{-\lambda}}{e^{\lambda}+e^{-\lambda}}\right) = \frac{\tanh(\lambda)}{2}. \tag{2.16}$$

**Conclusion 2.1.** *The maximum value of $M_{q,\lambda}$ is*

$$M_{q,\lambda}\left(\frac{\ln q}{2\lambda}\right) = \frac{\tanh(\lambda)}{2}, \ \ \lambda > 0. \tag{2.17}$$

We mention the following:

**Theorem 2.1** ([6, Ch. 18, p. 458]). *We have that*

$$\sum_{i=-\infty}^{\infty} M_{q,\lambda}(x-i) = 1, \ \forall\, x \in \mathbb{R}, \forall\, \lambda, q > 0. \tag{2.18}$$

Also, it holds the following:

**Theorem 2.2** ([6, Ch. 18, p. 459]). *It holds*

$$\int_{-\infty}^{\infty} M_{q,\lambda}(x)\,dx = 1, \ \ \lambda, q > 0. \tag{2.19}$$

So that $M_{q,\lambda}$ is a density function on $\mathbb{R}$; $\lambda, q > 0$. Similarly, we get that

$$\int_{-\infty}^{\infty} M_{\frac{1}{q},\lambda}(x)\,dx = 1, \ \ \lambda, q > 0, \tag{2.20}$$

so that $M_{\frac{1}{q},\lambda}$ is a density function. Furthermore, we observe the symmetry

$$\left(M_{q,\lambda} + M_{\frac{1}{q},\lambda}\right)(-x) = \left(M_{q,\lambda} + M_{\frac{1}{q},\lambda}\right)(x), \ \ \forall\, x \in \mathbb{R}. \tag{2.21}$$

Furthermore

$$\varphi = \frac{M_{q,\lambda} + M_{\frac{1}{q},\lambda}}{2} > 0 \tag{2.22}$$

is a new density function over $\mathbb{R}$, i.e.

$$\int_{-\infty}^{\infty} \varphi(x)\,dx = 1,$$

and $\varphi$ is an even function. Clearly, then

$$\int_{-\infty}^{\infty} \varphi(nx-u)\,du = 1, \ \ \forall\, n \in \mathbb{N}, x \in \mathbb{R}. \tag{2.23}$$

## 3. Basics

We give the following:

**Definition 3.1.** *Let $f \in C_B(\mathbb{R})$ (continuous and bounded functions on $\mathbb{R}$), $n \in \mathbb{N}$. We define the following basic activated hyperbolic tangent perturbed convolution type operators*

$$A_n(f)(x) := \int_{-\infty}^{\infty} f\left(\frac{u}{n}\right) \varphi(nx-u)\, du, \quad \forall\, x \in \mathbb{R}. \tag{3.24}$$

In this work, we examine the quantitative convergence of $A_n$ to the unit operator. We study similarly the activated Kantorovich type operators,

$$\begin{aligned} A_n^*(f)(x) &:= n \int_{-\infty}^{\infty} \left( \int_{\frac{u}{n}}^{\frac{u+1}{n}} f(t)\, dt \right) \varphi(nx-u)\, du \\ &= n \int_{-\infty}^{\infty} \left( \int_0^{\frac{1}{n}} f\left(t+\frac{u}{n}\right) dt \right) \varphi(nx-u)\, du, \end{aligned} \tag{3.25}$$

where $f \in C_B(\mathbb{R})$, $n \in \mathbb{N}$, $x \in \mathbb{R}$, and the activated quadrature operators

$$\overline{A_n}(f)(x) := \int_{-\infty}^{\infty} \left( \sum_{i=1}^{r} w_i f\left(\frac{u}{n} + \frac{i}{nr}\right) \right) \varphi(nx-u)\, du, \tag{3.26}$$

where $w_i \geq 0$, $\sum_{i=1}^{r} w_i = 1$; $f \in C_B(\mathbb{R})$, $n \in \mathbb{N}$, $x \in \mathbb{R}$. An essential property follows:

**Theorem 3.3** ([7]). *Let $0 < \alpha < 1$, $n \in \mathbb{N} : n^{1-\alpha} > 2$. Then*

$$\int_{\{u \in \mathbb{R}:\, |nx-u| \geq n^{1-\alpha}\}} \varphi(nx-u)\, du < \frac{\left(q+\frac{1}{q}\right)}{e^{2\lambda(n^{1-\alpha}-1)}}, \quad q, \lambda > 0. \tag{3.27}$$

The first modulus of continuity here is

$$\omega_1(f,\delta) := \sup_{x,y \in \mathbb{R}, |x-y| \leq \delta} |f(x) - f(y)|, \quad \delta > 0. \tag{3.28}$$

We need the following:

**Proposition 3.1** ([7]). *It holds ($k \in \mathbb{N}$)*

$$\int_{-\infty}^{\infty} |z|^k \varphi(z)\, dz \leq \left[ \frac{\tanh(\lambda)}{(k+1)} + \frac{\left(q+\frac{1}{q}\right) e^{2\lambda} k!}{(2\lambda)^k} \right] < \infty. \tag{3.29}$$

We make the following:

**Remark 3.1.** *Given $f \in C_B(\mathbb{R})$, by [7], we obtain that $A_n(f)$, $A_n^*(f)$, $\overline{A}_n(f) \in C_B(\mathbb{R})$. Clearly, here $A_n$, $A_n^*$, $\overline{A}_n$ are positive linear operators from $C_B(\mathbb{R})$ into itself, with the property $A_n(1) = A_n^*(1) = \overline{A}_n(1) = 1$, $n \in \mathbb{N}$. Let $i \in \mathbb{N}$ be fixed. Assume that $f \in C^{(i)}(\mathbb{R})$, with $f^{(j)} \in C_B(\mathbb{R})$, for $j = 0, 1, ..., i$. We derive from [7] that*

$$\begin{aligned} \left(A_n(f)\right)^{(j)}(x) &= A_n\left(f^{(j)}\right)(x), \\ \left(A_n^*(f)\right)^{(j)}(x) &= A_n^*\left(f^{(j)}\right)(x), \\ \left(\overline{A_n}(f)\right)^{(j)}(x) &= \overline{A_n}\left(f^{(j)}\right)(x), \quad \forall\, x \in \mathbb{R}, \end{aligned} \tag{3.30}$$

*for all* $j = 1, ..., i$.

Call $\Psi_n$ any of the $A_n$, $A_n^*$ and $\overline{A}_n$, $n \in \mathbb{N}$. We also make the following:

**Remark 3.2.** *Furthermore, it holds*

$$|A_n(f)(x)| \leq \|f\|_\infty \int_{-\infty}^{\infty} \varphi(z)\, dz = \|f\|_\infty,$$

*i.e.*

$$\|A_n(f)\|_\infty \leq \|f\|_\infty \tag{3.31}$$

*so* $A_n$ *is a bounded positive linear operator. Clearly, it holds*

$$\left\|A_n^2(f)\right\|_\infty = \|A_n(A_n(f))\|_\infty \leq \|A_n(f)\|_\infty \leq \|f\|_\infty. \tag{3.32}$$

*And for* $k \in \mathbb{N}$ *we obtain*

$$\left\|A_n^k(f)\right\|_\infty \leq \left\|A_n^{k-1}(f)\right\|_\infty \leq \left\|A_n^{k-2}(f)\right\|_\infty \leq ... \leq \|f\|_\infty, \tag{3.33}$$

*so the contraction property valid and* $A_n^k$ *is a bounded linear operator. Let* $r \in \mathbb{N}$. *We observe that*

$$\begin{aligned} A_n^r f - f = &\left(A_n^r f - A_n^{r-1} f\right) + \left(A_n^{r-1} f - A_n^{r-2} f\right) + \left(A_n^{r-2} f - A_n^{r-3} f\right) \\ &+ ... + \left(A_n^2 f - A_n f\right) + \left(A_n f - f\right). \end{aligned} \tag{3.34}$$

*Then*

$$\begin{aligned} \|A_n^r f - f\|_\infty \leq &\left\|A_n^r f - A_n^{r-1} f\right\|_\infty + \left\|A_n^{r-1} f - A_n^{r-2} f\right\|_\infty + \left\|A_n^{r-2} f - A_n^{r-3} f\right\|_\infty \\ &+ ... + \left\|A_n^2 f - A_n f\right\|_\infty + \|A_n f - f\|_\infty \\ = &\left\|A_n^{r-1}(A_n f - f)\right\|_\infty + \left\|A_n^{r-2}(A_n f - f)\right\|_\infty + ... + \|A_n(A_n f - f)\|_\infty \\ &+ \|A_n f - f\|_\infty \leq r \|A_n f - f\|_\infty. \end{aligned}$$

*Therefore*

$$\|A_n^r f - f\|_\infty \leq r \|A_n f - f\|_\infty. \tag{3.35}$$

*Let now* $m_1, m_2, ..., m_r \in \mathbb{N} : m_1 \leq m_2 \leq ... \leq m_r$, *and* $A_{m_i}$ *as above, then*

$$\begin{aligned} &A_{m_r}\left(A_{m_{r-1}}(...A_{m_2}(A_{m_1} f))\right) - f \\ &= ... = A_{m_r}\left(A_{m_{r-1}}(...A_{m_2})\right)(A_{m_1} f - f) + A_{m_r}\left(A_{m_{r-1}}(...A_{m_3})\right)(A_{m_2} f - f) \\ &+ A_{m_r}\left(A_{m_{r-1}}(...A_{m_4})\right)(A_{m_3} f - f) + ... + A_{m_r}\left(A_{m_{r-1}} f - f\right) + A_{m_r} f - f. \end{aligned} \tag{3.36}$$

*Consequently it holds, as in* [6, Chapter 2],

$$\left\|A_{m_r}\left(A_{m_{r-1}}(...A_{m_2}(A_{m_1} f))\right) - f\right\|_\infty \leq \sum_{i=1}^{r} \|A_{m_i} f - f\|_\infty. \tag{3.37}$$

*All of (3.31)-(3.37) are also true for* $A_n^*$ *and* $\overline{A}_n$, *and* $n \in \mathbb{N}$.

We need the following Hölder's type inequality for positive linear operators.

**Theorem 3.4** ([9]). *Let* $L$ *be a positive linear operator from* $C(\mathbb{R})$ *into* $C_B(\mathbb{R})$, *and* $f, g \in C(\mathbb{R})$, *furthermore let* $p, q > 1 : \frac{1}{p} + \frac{1}{q} = 1$. *Assume that* $L\left(\left(|f(\cdot)|^p\right)\right)(s_*)$, $L\left(\left(|g(\cdot)|^q\right)\right)(s_*) > 0$ *for some* $s_* \in \mathbb{R}$. *Then*

$$L(|f(\cdot)\, g(\cdot)|)(s_*) \leq \left(L\left(\left(|f(\cdot)^p|\right)\right)(s_*)\right)^{\frac{1}{p}} \left(L\left(\left(|g(\cdot)^q|\right)\right)(s_*)\right)^{\frac{1}{q}}. \tag{3.38}$$

We also need the following

**Theorem 3.5** ([9]). *Let $N \in \mathbb{N}$ and $f, f^{(N)} \in C_B(\mathbb{R})$, $x \in \mathbb{R}$. Consider $L_n$ a sequence of positive linear operators from $C_B(\mathbb{R})$ into itself, $n \in \mathbb{N}$, such that $L_n(1) = 1$. Assume $L_n\left(|\cdot - x|^{N+1}\right)(x) > 0$, and $f^{(i)}(x) = 0$, for $i = 1, ..., N$. Then*

$$|L_n(f)(x) - f(x)| \leq \omega_1\left(f^{(N)}, \left(\left(L_n\left(|\cdot - x|^{N+1}\right)\right)(x)\right)^{\frac{1}{N+1}}\right)$$

$$\times \left[L_n\left(|\cdot - x|^{N}\right)(x) + \frac{\left(L_n\left(|\cdot - x|^{N+1}\right)(x)\right)^{\frac{N}{N+1}}}{(N+1)}\right] < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{3.39}$$

*Given that $\lim_{N\to+\infty} L_n\left(|\cdot - x|^{N+1}\right)(x) = 0$, then $\lim_{n\to+\infty} L_n(f)(x) = f(x)$. If $N = 1$, we derive*

$$|L_n(f)(x) - f(x)| \leq \omega_1\left(f', \left(\left(L_n\left((\cdot - x)^2\right)\right)(x)\right)^{\frac{1}{2}}\right)$$

$$\times \left[L_n(|\cdot - x|)(x) + \frac{\left(L_n\left((\cdot - x)^2\right)(x)\right)^{\frac{1}{2}}}{2}\right] < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{3.40}$$

*Given that $\lim_{n\to+\infty} L_n\left((\cdot - x)^2\right)(x) = 0$, then $\lim_{n\to+\infty} L_n(f)(x) = f(x)$.*

**Note 3.1.** *Assuming $L_n\left(|\cdot - x|^{N+1}\right)(x) > 0$, by Theorem 3.4, for $g = 1$, and $L_n$ such that $L_n(1) = 1$, we obtain*

$$L_n\left(|\cdot - x|^{N}\right)(x) \leq \left(L_n\left(|\cdot - x|^{N+1}\right)(x)\right)^{\frac{N}{N+1}}. \tag{3.41}$$

*Proof.* In case of $N = 1$, we derive

$$L_n(|\cdot - x|)(x) \leq \sqrt{\left(L_n\left((\cdot - x)^2\right)(x)\right)}. \tag{3.42}$$

□

We also need the following:

**Theorem 3.6** ([11]). *Let $f \in C_B(\mathbb{R})$, $x \in \mathbb{R}$. Consider $L_n$ a sequence of positive linear operators from $C_B(\mathbb{R})$ into itself, $n \in \mathbb{N}$, such that $L_n(1) = 1$. Assume that $L_n(|\cdot - x|)(x) > 0$. Then*

$$|L_n(f)(x) - f(x)| \leq 2\omega_1\left(f, L_n(|\cdot - x|)(x)\right), \quad \forall\, n \in \mathbb{N}. \tag{3.43}$$

*Given that $\lim_{n\to+\infty} L_n(|\cdot - x|)(x) = 0$, and $f$ is also uniformly continuous, we obtain $\lim_{n\to+\infty} L_n(f)(x) = f(x)$.*

## 4. Auxiliary Results

We have the following result.

**Lemma 4.1.** *Let $m \in \mathbb{N}$. Then*

$$0 < A_N\left(|\cdot - x|^m\right)(x) \leq \frac{1}{n^m}\left[\frac{\tanh(\lambda)}{(m+1)} + \frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right](<+\infty) \to 0, \text{ as } n \to +\infty, \tag{4.44}$$

$\forall\, x \in \mathbb{R}$. *And, it is*

$$0 < \left\|A_N\left(|\cdot - x|^m\right)(x)\right\|_\infty \leq \frac{1}{n^m}\left[\frac{\tanh(\lambda)}{(m+1)} + \frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right](<+\infty) \to 0, \tag{4.45}$$

*as* $n \to +\infty$.

*Proof.* We have that ($m \in \mathbb{N}$, $x \in \mathbb{R}$)

$$\begin{aligned} 0 < A_N\left(|\cdot - x|^m\right)(x) &= \int_{-\infty}^{\infty}\left|\frac{u}{n} - x\right|^m \varphi(nx-u)\,du \\ &= \frac{1}{n^m}\int_{-\infty}^{\infty}|nx-u|^m\,\varphi(nx-u)\,du \\ &\overset{(z:=nx-u)}{=} \frac{1}{n^m}\int_{-\infty}^{\infty}|z|^m\,\varphi(z)\,dz \\ &\leq \frac{1}{n^m}\left[\frac{\tanh(\lambda)}{(m+1)} + \frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right](<+\infty) \to 0, \text{ as } n \to +\infty. \end{aligned} \tag{4.46}$$

□

It follows:

**Lemma 4.2.** *Let* $m \in \mathbb{N}$. *Then*

$$0 < A_N^*\left(|\cdot - x|^m\right)(x) \leq \frac{2^{m-1}}{n^m}\left[1 + \left(\frac{\tanh(\lambda)}{(m+1)} + \frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right)\right](<+\infty) \to 0, \tag{4.47}$$

*as* $n \to +\infty$ *and* $\forall\, x \in \mathbb{R}$. *And, it is*

$$0 < \left\|A_N^*\left(|\cdot - x|^m\right)(x)\right\|_\infty \leq \frac{2^{m-1}}{n^m}\left[1 + \left(\frac{\tanh(\lambda)}{(m+1)} + \frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right)\right](<+\infty) \to 0, \tag{4.48}$$

*as* $n \to +\infty$.

*Proof.* We have that ($m \in \mathbb{N}$, $x \in \mathbb{R}$)

$$\begin{aligned} 0 < A_N^*\left(|\cdot - x|^m\right)(x) &= n\int_{-\infty}^{\infty}\left(\int_0^{\frac{1}{n}}\left|t + \frac{u}{n} - x\right|^m dt\right)\varphi(nx-u)\,du \\ &\leq n\int_{-\infty}^{\infty}\left(\int_0^{\frac{1}{n}}\left(|t| + \left|\frac{u}{n} - x\right|\right)^m dt\right)\varphi(nx-u)\,du \\ &\leq \int_{-\infty}^{\infty}\left(\frac{1}{n} + \left|\frac{u}{n} - x\right|\right)^m \varphi(nx-u)\,du \\ &= \frac{1}{n^m}\int_{-\infty}^{\infty}(1+|nx-u|)^m\,\varphi(nx-u)\,du \end{aligned} \tag{4.49}$$

$$\leq \frac{2^{m-1}}{n^m}\left[1+\int_{-\infty}^{\infty}\left|nx-u\right|^m\varphi\left(nx-u\right)du\right]$$
$$=\frac{2^{m-1}}{n^m}\left[1+\int_{-\infty}^{\infty}\left|z\right|^m\varphi\left(z\right)dz\right]$$
$$\leq \frac{2^{m-1}}{n^m}\left[1+\left(\frac{\tanh\left(\lambda\right)}{\left(m+1\right)}+\frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{\left(2\lambda\right)^m}\right)\right]\to 0,\ \text{as } n\to+\infty,$$

and it is finite. □

At last, we obtain the following:

**Lemma 4.3.** *Let $m\in\mathbb{N}$. Then*

$$0<\overline{A}_N\left(\left|\cdot-x\right|^m\right)(x)\leq\frac{2^{m-1}}{n^m}\left[1+\left(\frac{\tanh\left(\lambda\right)}{\left(m+1\right)}+\frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{\left(2\lambda\right)^m}\right)\right](<+\infty)\to 0, \tag{4.50}$$

*as $n\to+\infty$ and $\forall\, x\in\mathbb{R}$. And, it is*

$$0<\left\|\overline{A}_N\left(\left|\cdot-x\right|^m\right)(x)\right\|_\infty\leq\frac{2^{m-1}}{n^m}\left[1+\left(\frac{\tanh\left(\lambda\right)}{\left(m+1\right)}+\frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{\left(2\lambda\right)^m}\right)\right](<+\infty)\to 0, \tag{4.51}$$

*as $n\to+\infty$.*

*Proof.* We observe that

$$\begin{aligned}0<\overline{A}_N\left(\left|\cdot-x\right|^m\right)(x)&=\int_{-\infty}^{\infty}\left(\sum_{i=1}^{r}w_i\left|\frac{u}{n}+\frac{i}{nr}-x\right|^m\right)\varphi\left(nx-u\right)du\\ &\leq\int_{-\infty}^{\infty}\left(\sum_{i=1}^{r}w_i\left(\left|\frac{u}{n}-x\right|+\frac{i}{nr}\right)^m\right)\varphi\left(nx-u\right)du\\ &\leq\int_{-\infty}^{\infty}\left(\frac{1}{n}+\left|\frac{u}{n}-x\right|\right)^m\varphi\left(nx-u\right)du\\ &=\frac{1}{n^m}\int_{-\infty}^{\infty}\left(1+\left|nx-u\right|\right)^m\varphi\left(nx-u\right)du.\end{aligned} \tag{4.52}$$

The proof finishes as in the proof of Lemma 4.2. □

## 5. Main Results

We present the following results.

**Theorem 5.7.** *Let $N\in\mathbb{N}$ and $f, f^{(N)}\in C_B\left(\mathbb{R}\right)$, $x\in\mathbb{R}$. Assume $f^{(i)}\left(x\right)=0$, $i=1,...,N$. Then*

$$\left|\Psi_n\left(f\right)(x)-f\left(x\right)\right|\leq\omega_1\left(f^{(N)},\left(\left(\Psi_n\left(\left|\cdot-x\right|^{N+1}\right)\right)(x)\right)^{\frac{1}{N+1}}\right)$$
$$\times\left[\Psi_n\left(\left|\cdot-x\right|^N\right)(x)+\frac{\left(\Psi_n\left(\left|\cdot-x\right|^{N+1}\right)(x)\right)^{\frac{N}{N+1}}}{\left(N+1\right)}\right]<+\infty,\ \ \forall\, n\in\mathbb{N}. \tag{5.53}$$

*Hence,* $\lim_{n\to+\infty}\Psi_n(f)(x)=f(x)$. *If* $N=1$, *we obtain*

$$|\Psi_n(f)(x)-f(x)|\leq\omega_1\left(f',\left(\left(\Psi_n\left((\cdot-x)^2\right)\right)(x)\right)^{\frac{1}{2}}\right)$$

$$\times\left[\Psi_n(|\cdot-x|)(x)+\frac{\left(\Psi_n\left((\cdot-x)^2\right)(x)\right)^{\frac{1}{2}}}{2}\right]<+\infty,\ \ \forall\, n\in\mathbb{N}. \tag{5.54}$$

*Again, it holds* $\lim_{n\to+\infty}\Psi_n(f)(x)=f(x)$.

*Proof.* By Theorem 3.5 and Lemmas 4.1-4.3. □

**Theorem 5.8.** *Let* $f\in C_B(\mathbb{R})$, $x\in\mathbb{R}$. *Then*

$$|\Psi_n(f)(x)-f(x)|\leq 2\omega_1(f,\Psi_n(|\cdot-x|)(x))<+\infty,\ \ \forall\, n\in\mathbb{N}. \tag{5.55}$$

*If* $f$ *is also uniformly continuous, we derive* $\lim_{n\to+\infty}\Psi_n(f)(x)=f(x)$.

*Proof.* Direct application of Theorem 3.6 and Lemmas 4.1-4.3. □

We make the following:

**Remark 5.3.** *By (3.41), (3.42), we derive that*

$$\Psi_n\left(|\cdot-x|^N\right)(x)\leq\left(\Psi_n\left(|\cdot-x|^{N+1}\right)(x)\right)^{\frac{N}{N+1}}, \tag{5.56}$$

*and*

$$\Psi_n(|\cdot-x|)(x)\leq\sqrt{\left(\Psi_n\left((\cdot-x)^2\right)(x)\right)},\ \ \forall\, n\in\mathbb{N}. \tag{5.57}$$

**Notation 5.1.** *Let* $m,n\in\mathbb{N}$. *Denote by*

$$\rho_{1n}(m):=\frac{1}{n^m}\left[\frac{\tanh(\lambda)}{(m+1)}+\frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right](<+\infty)\to 0,\ \ as\ n\to+\infty, \tag{5.58}$$

*and*

$$\rho_{2n}(m):=\frac{2^{m-1}}{n^m}\left[1+\left(\frac{\tanh(\lambda)}{(m+1)}+\frac{\left(q+\frac{1}{q}\right)e^{2\lambda}m!}{(2\lambda)^m}\right)\right](<+\infty)\to 0,\ \ as\ n\to+\infty. \tag{5.59}$$

We give the following explicit results.

**Corollary 5.1.** *Let* $N\in\mathbb{N}$ *and* $f,f^{(N)}\in C_B(\mathbb{R})$, $x\in\mathbb{R}$. *Assume* $f^{(i)}(x)=0$, $i=1,...,N$. *Then*

$$|A_n(f)(x)-f(x)|\leq\omega_1\left(f^{(N)},(\rho_{1n}(N+1))^{\frac{1}{N+1}}\right)$$

$$\times\left[\rho_{1n}(N)+\frac{(\rho_{1n}(N+1))^{\frac{N}{N+1}}}{(N+1)}\right]<+\infty,\ \ \forall\, n\in\mathbb{N}. \tag{5.60}$$

*Hence* $\lim_{n\to+\infty}A_n(f)(x)=f(x)$. *If* $N=1$, *we obtain*

$$|A_n(f)(x)-f(x)|\leq\omega_1\left(f',(\rho_{1n}(2))^{\frac{1}{2}}\right)$$

$$\times \left[ \rho_{1n}(1) + \frac{(\rho_{1n}(2))^{\frac{1}{2}}}{2} \right] < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.61}$$

*Again, it holds* $\lim_{n\to+\infty} A_n(f)(x) = f(x)$.

*Proof.* By Theorem 5.7 and Lemma 4.1, see also (5.58). □

**Corollary 5.2.** *Let* $N \in \mathbb{N}$ *and* $f, f^{(N)} \in C_B(\mathbb{R})$, $x \in \mathbb{R}$. *Assume* $f^{(i)}(x) = 0$, $i = 1, ..., N$. *Then*

$$\left\{ \begin{array}{l} |A_n^*(f)(x) - f(x)| \\ |\overline{A}_n(f)(x) - f(x)| \end{array} \right\} \leq \omega_1\left(f^{(N)}, (\rho_{2n}(N+1))^{\frac{1}{N+1}}\right)$$
$$\times \left[ \rho_{2n}(N) + \frac{(\rho_{2n}(N+1))^{\frac{N}{N+1}}}{(N+1)} \right] < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.62}$$

*Hence* $\lim_{n\to+\infty} A_n^*(f)(x) = \lim_{n\to+\infty} \overline{A}_n(f)(x) = f(x)$. *If* $N = 1$, *we obtain*

$$\left\{ \begin{array}{l} |A_n^*(f)(x) - f(x)| \\ |\overline{A}_n(f)(x) - f(x)| \end{array} \right\} \leq \omega_1\left(f', (\rho_{2n}(2))^{\frac{1}{2}}\right)$$
$$\times \left[ \rho_{2n}(1) + \frac{(\rho_{2n}(2))^{\frac{1}{2}}}{2} \right] < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.63}$$

*Again, it holds* $\lim_{n\to+\infty} A_n^*(f)(x) = \lim_{n\to+\infty} \overline{A}_n(f)(x) = f(x)$.

*Proof.* By Theorem 5.7 and Lemmas 4.2, 4.3, see also (5.59). □

**Corollary 5.3.** *Let* $f \in C_B(\mathbb{R})$. *Then*

$$\|A_n(f) - f\|_\infty \leq 2\omega_1(f, \rho_{1n}(1)) < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.64}$$

*If* $f$ *is also uniformly continuous, we get that* $\lim_{n\to+\infty} A_n(f) = f$, *pointwise and uniformly.*

*Proof.* By Theorem 5.8, Lemma 4.1, see also (5.58). □

**Corollary 5.4.** *Let* $f \in C_B(\mathbb{R})$. *Then*

$$\left\{ \begin{array}{l} \|A_n^*(f) - f\|_\infty \\ \|\overline{A}_n(f) - f\|_\infty \end{array} \right\} \leq 2\omega_1(f, \rho_{2n}(1)) < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.65}$$

*If* $f$ *is also uniformly continuous, we obtain that* $\lim_{n\to+\infty} A_n^*(f) = \lim_{n\to+\infty} \overline{A}_n(f) = f$, *pointwise and uniformly.*

*Proof.* By Theorem 5.8, Lemmas 4.2, 4.3 see also (5.59). □

We continue with simultaneous approximations.

**Theorem 5.9.** *Let* $f^{(j)} \in C_B(\mathbb{R})$, *for* $j = 0, 1, ..., N \in \mathbb{N}$. *Then*

$$\left\| (A_n(f))^{(j)} - f^{(j)} \right\|_\infty \leq 2\omega_1\left(f^{(j)}, \rho_{1n}(1)\right) < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.66}$$

*If* $f^{(j)}$ *is uniformly continuous, we get that* $\lim_{n\to+\infty} (A_n(f))^{(j)} = f^{(j)}$, *pointwise and uniformly.*

*Proof.* By (3.30) and Corollary 5.3. □

**Theorem 5.10.** *Let $f^{(j)} \in C_B(\mathbb{R})$, for $j = 0, 1, ..., N \in \mathbb{N}$. Then*

$$\left\{ \begin{array}{l} \left\| \left(A_n^*(f)\right)^{(j)} - f^{(j)} \right\|_\infty \\ \left\| \left(\overline{A}_n(f)\right)^{(j)} - f^{(j)} \right\|_\infty \end{array} \right\} \leq 2\omega_1\left(f^{(j)}, \rho_{2n}(1)\right) < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.67}$$

*If $f^{(j)}$ is uniformly continuous, we obtain that $\lim_{n\to+\infty} \left(A_n^*(f)\right)^{(j)} = \lim_{n\to+\infty} \left(\overline{A}_n(f)\right)^{(j)} = f^{(j)}$, pointwise and uniformly.*

*Proof.* By (3.30) and Corollary 5.4. □

Next, we talk about iterated approximation.

**Remark 5.4.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$. Here $\Psi_n$ is any of $A_n$, $A_n^*$, $\overline{A}_n$, $\forall\, n \in \mathbb{N}$. By Remark 3.2, see also (3.35), we have that*

$$\left\|\Psi_n^r(f) - f\right\|_\infty \leq r \left\|\Psi_n(f) - f\right\|_\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.68}$$

**Corollary 5.5.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$. Then*

$$\left\|A_n^r(f) - f\right\|_\infty \leq r \left\|A_n(f) - f\right\|_\infty \leq 2r\omega_1\left(f, \rho_{1n}(1)\right) < +\infty, \quad \forall\, n \in \mathbb{N}. \tag{5.69}$$

*If $f$ is also uniformly continuous, we get that $\lim_{n\to+\infty} A_n^r(f) = f$, pointwise and uniformly. And the speed of convergence of $A_n^r$ to the unit operator is not worse than of $A_n$ to the unit.*

*Proof.* By (5.64) and (5.68). □

**Corollary 5.6.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$. Then*

$$\begin{aligned} \left\{ \begin{array}{l} \left\|A_n^{*^r}(f) - f\right\|_\infty \\ \left\|\overline{A}_n^r(f) - f\right\|_\infty \end{array} \right\} &\leq r \left\{ \begin{array}{l} \left\|A_n^*(f) - f\right\|_\infty \\ \left\|\overline{A}_n(f) - f\right\|_\infty \end{array} \right\} \\ &\leq 2r\omega_1\left(f, \rho_{2n}(1)\right) < +\infty, \quad \forall\, n \in \mathbb{N}. \end{aligned} \tag{5.70}$$

*If $f$ is also uniformly continuous, we get that $\lim_{n\to+\infty} A_n^{*^r}(f) = \lim_{n\to+\infty} \overline{A}_n^r(f) = f$, pointwise and uniformly. And the speed of convergence of $A_n^{*^r}$, $\overline{A}_n^r$ to the unit operator is not worse than of $A_n^*$, $\overline{A}_n$ to the unit, respectively.*

*Proof.* By (5.65) and (5.68). □

We finish with more general iterated approximation results.

**Remark 5.5.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$, and $m_1, m_2, ..., m_r \in \mathbb{N} : m_1 \leq m_2 \leq ... \leq m_r$. Here $\Psi_{m_i}$ is any of $A_{m_i}$, $A_{m_i}^*$, $\overline{A}_{m_i}$. By Remark 3.2, see also (3.37), we have that*

$$\left\|\Psi_{m_r}\left(\Psi_{m_{r-1}}\left(...\Psi_{m_2}\left(\Psi_{m_1} f\right)\right)\right) - f\right\|_\infty \leq \sum_{i=1}^{r} \left\|\Psi_{m_i} f - f\right\|_\infty. \tag{5.71}$$

**Corollary 5.7.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$, and $m_1, m_2, ..., m_r \in \mathbb{N} : m_1 \leq m_2 \leq ... \leq m_r$. Then*

$$\begin{aligned} \left\|A_{m_r}\left(A_{m_{r-1}}\left(...A_{m_2}\left(A_{m_1} f\right)\right)\right) - f\right\|_\infty &\leq \sum_{i=1}^{r} \left\|A_{m_i} f - f\right\|_\infty \leq 2\left(\sum_{i=1}^{r} \omega_1\left(f, \rho_{1m_i}(1)\right)\right) \\ &\leq 2r\omega_1\left(f, \rho_{1m_1}(1)\right). \end{aligned}$$

*The speed of convergence to the unit operator of the above activated multiply iterated operators is not worse than the speed of convergence to the unit of $A_{m_1}$.*

*Proof.* By (5.71), (5.64) and (5.58). □

**Corollary 5.8.** *Let $f \in C_B(\mathbb{R})$, $r \in \mathbb{N}$, and $m_1, m_2, ..., m_r \in \mathbb{N} : m_1 \leq m_2 \leq ... \leq m_r$. Then*

$$\left\| A^*_{m_r}\left(A^*_{m_{r-1}}\left(...A^*_{m_2}\left(A^*_{m_1} f\right)\right)\right) - f \right\|_\infty \leq \sum_{i=1}^{r} \left\| A^*_{m_i} f - f \right\|_\infty \leq 2\left(\sum_{i=1}^{r} \omega_1\left(f, \rho_{2m_i}(1)\right)\right)$$
$$\leq 2r\omega_1\left(f, \rho_{2m_1}(1)\right),$$

*and*

$$\left\| \overline{A}_{m_r}\left(\overline{A}_{m_{r-1}}\left(...\overline{A}_{m_2}\left(\overline{A}_{m_1} f\right)\right)\right) - f \right\|_\infty \leq \sum_{i=1}^{r} \left\| \overline{A}_{m_i} f - f \right\|_\infty \leq 2\left(\sum_{i=1}^{r} \omega_1\left(f, \rho_{2m_i}(1)\right)\right)$$
$$\leq 2r\omega_1\left(f, \rho_{2m_1}(1)\right).$$

*The speed of convergence to the unit operator of the above activated multiply iterated operators is not worse than the speed of convergence to the unit of $A^*_{m_1}$, $\overline{A}_{m_1}$, respectively.*

*Proof.* By (5.71), (5.70) and (5.59). □

## References


[1] G. A. Anastassiou: *A "K-Attainable" inequality related to the convergence of positive linear operators*, J. Approx. Theory, **44** (1985), 380–383.

[2] G. A. Anastassiou: *Moments in probability and approximation theory*, Pitman Research Notes in Math., Longman Sci. & Tech., Harlow (1993).

[3] G. A. Anastassiou: *Quantitative approximations*, Chapmen & Hall/CRC, London, New York (2001).

[4] G. A. Anastassiou: *Intelligent computations: Abstract fractional calculus, inequalities, approximations*, Springer, New York (2018).

[5] G. A. Anastassiou, S. G. Gal: *Approximation theory*, Birkäuser, Berlin (2000).

[6] G. A. Anastassiou: *Parametrized, deformed and general neural networks*, Springer, New York (2023).

[7] G. A. Anastassiou: *Approximation by symmetrized and perturbed hyperbolic tangent activated convolution type operators*, Mathematics, **12** (20) (2024), Article ID: 3302.

[8] G. A. Anastassiou: *Trigonometric and hyperbolic generated approximation theory*, World Scientific, New York (2025).

[9] G. A. Anastassiou: *Neural networks in infinite domain as positive linear operastors*, Annales Univ. Sci. Budapest, Sect. Comp., **58** (2025), 15–29.

[10] R. Bojanic, O. Shisha: *On the precision of uniform approximation of continuous functions by certain linear operators of convolution type*, J. Approx. Theory, **8** (1973), 101–113.

[11] R. A. DeVore: *The approximation of continuous functions by positive linear operators*, Lecture Notes in Mathematics, **293**, Springer, New York (1972).

[12] H. S. Jung, R. Sakai: *Local saturation of a positive linear convolution operator*, J. Inequal. Appl., **2014** (2014), Article ID: 329.

[13] A. I. Kamzolov: *The order of approximation of functions of class $Z_2(E_n)$ by positive linear convolution operators*, Mat. Zametki, **7** (1970), 723–732 (Russian).

[14] G. Moldovan: *Discrete convolutions in connection with functions of several variables and positive linear operators*, Stud. Univ. Babeș-Bolyai, Math., **19** (1) (1974), 51–57 (Romanian).

[15] G. Moldovan: *Discrete convolutions and linear positive operators*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math., **15** (1972), 31–44.

[16] J. J. Swetits, B. Wood: *Local $L_p$-saturation of positive linear convolution operators*, J. Approx. Theory, **34** (4) (1982), 348–360.



George A. Anastassiou
University of Memphis
Department of Mathematical Sciences
TN 38152, Memphis, U.S.A.
*Email address*: ganastss@memphis.edu

MODERN MATHEMATICAL METHODS
**3** (2025), No. 2, pp. 85-91
https://modernmathmeth.com/
ISSN 3023 - 5294

*Research Article*

# On the general solution and stability of the functional equation $f(x-y)-f(x)f(y)=d\sin x\sin y$

JIRAPORN KUNRATTANAWORAWONG AND WUTIPHOL SINTUNAVARAT*

ABSTRACT. This paper is concerned with the investigation of the general solution to the following functional equation:

$$f(x-y)-f(x)f(y)=d\sin x\sin y$$

for all $x,y\in\mathbb{R}$, where $f:\mathbb{R}\to\mathbb{R}$ is an unknown function and $d\in\mathbb{R}\setminus\{0\}$ satisfying $d<1$. This nonlinear functional equation establishes an intriguing interplay between multiplicative and additive behaviors of the function $f$, perturbed by a bounded trigonometric term. We provide a complete characterization of all real-valued functions satisfying the equation, under minimal regularity assumptions. In addition, we analyze the Hyers–Ulam stability and the so-called superstability of the equation in the sense of functional equations, establishing that approximate solutions under certain bounded perturbations necessarily converge to exact solutions.

**Keywords:** Functional equations, stability, trigonometric identity.

**2020 Mathematics Subject Classification:** 65Q20, 97I70.

## 1. INTRODUCTION

Functional equations constitute an important branch of mathematics. The theory of functional equations is typically divided into two main areas. The first concerns the determination of exact solutions to a given functional equation, while the second focuses on the study of the stability of functional equations. The concept of stability arises from the following classical question: Under what conditions does a function that approximately satisfies a functional equation $\rho$ necessarily remain close to an exact solution of $\rho$? If the problem admits a solution, we say that the equation $\rho$ is stable.

The aforementioned question has given rise to various stability concepts, among which Ulam [1] formulated the first notable problem in 1940 in the context of group homomorphisms. Formally, Ulam's problem can be stated as follows: Let $G_1$ be a group and let $G_2$ be a metric group with a metric $d$. Given $\varepsilon>0$, does there exist a $\delta>0$ such that if a function $h:G_1\to G_2$ satisfies the inequality

$$d(h(xy),h(x)h(y))<\delta$$

for all $x,y\in G_1$, then there is a homomorphism $H:G_1\to G_2$ with

$$d(h(x),H(x))<\varepsilon$$

for all $x\in G_1$?

In 1941, Hyers [2] provided a partial solution to Ulam's problem for the additive Cauchy functional equation in Banach spaces, stated as follows:

Received: 31.05.2025; Accepted: 13.08.2025; Published Online: 20.08.2025
*Corresponding author: Wutiphol Sintunavarat; wutiphol@mathstat.sci.tu.ac.th
DOI: 10.64700/mmm.67

**Theorem 1.1** ([2]). *Let $E_1$ and $E_2$ be two Banach space and suppose that the mapping $f : E_1 \to E_2$ satisfies the inequality*

$$\|f(x+y) - f(x) - f(y)\| \leq \delta$$

*for all $x, y \in E_1$, where $\delta > 0$. Then there is a unique additive mapping $T : E_1 \to E_2$ satisfying*

$$\|f(x) - T(x)\| \leq \delta$$

*for all $x \in E_1$.*

The stability result of Hyers [2] was subsequently refined by Aoki [3], who established the case involving the existence of a unique additive Cauchy mapping. Building upon this foundation, Rassias [4] significantly broadened Hyers' theorem by introducing the concept of an unbounded Cauchy difference endowed with a $p$-order norm for $0 < p < 1$. In turn, the Rassias theorem was further generalized by Gajda [5] for the case $p > 1$, and by Rassias [6] for the case $p < 1$.

Over the past several decades, considerable effort has been devoted to investigating solutions and stability phenomena for a wide range of functional equations. Among these contributions, Butler [7] in 2003 posed the following intriguing problem concerning the solution of a specific functional equation: prove that, for $d < -1$, there exist exactly two functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$f(x+y) - f(x) f(y) = d \sin x \sin y \tag{1.1}$$

for all $x, y \in \mathbb{R}$. This problem was subsequently resolved by Rassias [8] in 2004, who proved that the general solution is given by

$$f(x) = \pm c \sin x + \cos x$$

for all $x \in \mathbb{R}$, where $c = \sqrt{-d-1}$. More recently, Jung et al. [9] have obtained notable results on the superstability and stability of the functional equation (1.1).

Motivated by the aforementioned perspective, this work introduces a new functional equation, closely related to (1.1), given by

$$f(x-y) - f(x) f(y) = d \sin x \sin y \tag{1.2}$$

for all $x, y \in \mathbb{R}$, where $f : \mathbb{R} \to \mathbb{R}$ is an unknown function and $d$ is a nonzero real constant with $d < 1$. The general solution of (1.2) is derived in the next section, followed by a detailed analysis of its superstability and stability properties in Section 3.

## 2. General Solutions of Functional Equation (1.2)

In this section, we establish the complete characterization of the solutions to the functional equation (1.2), as formalized in the theorem below.

**Theorem 2.2.** *Let $d$ be a nonzero constant with $d < 1$. The functional equation (1.2) has exactly two solutions in the class of functions $f : \mathbb{R} \to \mathbb{R}$. More precisely, if a function $f : \mathbb{R} \to \mathbb{R}$ is a solution of functional equation (1.2), then $f$ has one of the forms*

$$f(x) = c \sin x + \cos x \quad \text{and} \quad f(x) = -c \sin x + \cos x,$$

*where $c = \sqrt{1-d}$.*

*Proof.* It follows from (1.2) that

$$\begin{aligned} &df(x)\sin y\sin z+d\sin x\sin(y-z)-d\sin z\sin(y-x)-df(z)\sin y\sin x\\ &=f(x)[f(y-z)-f(y)f(z)]+[f(x-y+z)-f(x)f(y-z)]\\ &-[f(z-y+x)-f(z)f(y-x)]-f(z)[f(y-x)-f(y)f(x)]\\ &=0 \end{aligned} \tag{2.3}$$

for all $x,y,z\in\mathbb{R}$. If we set $y=z=\frac{\pi}{2}$ in the above equality, then

$$\begin{aligned} df(x)-d\sin\left(\frac{\pi}{2}-x\right)-df\left(\frac{\pi}{2}\right)\sin x&=0\\ df(x)-d\cos(x)-df\left(\frac{\pi}{2}\right)\sin x&=0\\ f(x)-f\left(\frac{\pi}{2}\right)\sin x-\cos x&=0 \end{aligned} \tag{2.4}$$

for all $x\in\mathbb{R}$. Substituting $0$ for $x$ in (2.4) yields $f(0)=1$. If we put $x=y=\frac{\pi}{2}$ in (1.2), then we obtain

$$\left[f\left(\frac{\pi}{2}\right)\right]^2=1-d \tag{2.5}$$

and hence

$$f\left(\frac{\pi}{2}\right)=c\quad\text{or}\quad f\left(\frac{\pi}{2}\right)=-c, \tag{2.6}$$

where $c:=\sqrt{1-d}$. Consequently, by (2.4), we have

$$f(x)=c\sin x+\cos x\quad\text{or}\quad f(x)=-c\sin x+\cos x \tag{2.7}$$

for all $x\in\mathbb{R}$, where each expression satisfies (1.2). □

## 3. Stability and Superstability Results for (1.2)

First, we prove a theorem concerning the superstability of the functional equation

$$f(x-y)=f(x)f(y)$$

for all $x,y\in\mathbb{R}$, where $f:\mathbb{R}\to\mathbb{C}$ is an unknown function.

**Theorem 3.3.** *Let an arbitrary constant $\varepsilon>0$ be fixed. If a function $f:\mathbb{R}\to\mathbb{C}$ satisfies the inequality*

$$|f(x-y)-f(x)f(y)|\le\varepsilon \tag{3.8}$$

*for all $x,y\in\mathbb{R}$, then either for each $x\in\mathbb{R}$, $|f(x)|\le\frac{1+\sqrt{1+4\varepsilon}}{2}$ or $|f(-x)|\le\frac{1+\sqrt{1+4\varepsilon}}{2}$ or $f(x-y)=f(x)f(y)$ for all $x,y\in\mathbb{R}$.*

*Proof.* If we set $\delta:=\frac{1+\sqrt{1+4\varepsilon}}{2}$, then $\delta^2-\delta=\varepsilon$ and $\delta>1$. Suppose that there exists an $a\in\mathbb{R}$ such that $|f(a)|>\delta$ and $|f(-a)|>\delta$, say $|f(a)|=\delta+p$ and $|f(-a)|=\delta+q$ for some $p,q>0$. It follows from (3.8) that

$$\begin{aligned} |f(2a)|&=|f(a)f(-a)-f(a)f(-a)+f(2a)|\\ &\ge|f(a)f(-a)|-|f(a)f(-a)-f(2a)|\\ &\ge(\delta+p)(\delta+q)-\varepsilon\\ &=\delta+(p+q)\delta+pq\\ &>\delta+(p+q). \end{aligned}$$

By using the same process, we get $|f(-2a)|>\delta+(p+q)$.

Now, we will show that

$$|f(2^n a)| > \delta + 2^{n-1}(p+q) \quad \text{and} \quad |f(-2^n a)| > \delta + 2^{n-1}(p+q) \tag{3.9}$$

for all $n \in \mathbb{N}$. To claim the above fact by using mathematical induction, we suppose that (3.9) holds for some $n \in \mathbb{N}$. From (3.8), we get

$$\begin{aligned} \left|f\left(2^{n+1}a\right)\right| &= \left|f(2^n a) f(-2^n a) - f(2^n a) f(-2^n a) + f\left(2^{n+1}a\right)\right| \\ &\geq |f(2^n a) f(-2^n a)| - \varepsilon \\ &> \left(\delta + 2^{n-1}(p+q)\right)^2 - \left(\delta^2 - \delta\right) \\ &> \delta + 2^n (p+q). \end{aligned}$$

Similarly, we get

$$\left|f\left(-2^{n+1}a\right)\right| > \delta + 2^n (p+q).$$

Hence, (3.9) is established for all $n \in \mathbb{N}$ via mathematical induction. For every $x, y, z \in \mathbb{R}$, it follows from (3.8) that

$$|f(z-x+y) - f(z) f(x-y)| \leq \varepsilon$$

and

$$|f(y-x+z) - f(y) f(x-z)| \leq \varepsilon.$$

This yields that

$$|f(z) f(x-y) - f(y) f(x-z)| \leq 2\varepsilon$$

for all $x, y, z \in \mathbb{R}$. Hence,

$$\begin{aligned} |f(z) f(x-y) - f(x) f(y) f(z)| &\leq |f(z) f(x-y) - f(y) f(x-z)| \\ &\quad + |f(y) f(x-z) - f(x) f(y) f(z)| \\ &\leq 2\varepsilon + |f(y)|\,\varepsilon, \end{aligned}$$

that is,

$$|f(x-y) - f(x) f(y)|\,|f(z)| \leq 2\varepsilon + |f(y)|\,\varepsilon$$

for all $x, y, z \in \mathbb{R}$. In particular,

$$|f(x-y) - f(x) f(y)| \leq \frac{2\varepsilon + |f(y)|\,\varepsilon}{|f(2^n a)|} \tag{3.10}$$

for all $x, y \in \mathbb{R}$ and any $n \in \mathbb{N}$. Letting $n \to \infty$ in (3.10) and considering (3.9), we conclude that $f(x-y) = f(x) f(y)$ for all $x, y \in \mathbb{R}$. □

From the above result, we obtain the following theorem.

**Theorem 3.4.** *Let $d$ be a nonzero constant with $d < 1$ and $\varepsilon$ be a real constant with $0 < \varepsilon < |d|$. If a function $f : \mathbb{R} \to \mathbb{R}$ satisfies the following inequality*

$$|f(x-y) - f(x) f(y) - d \sin x \sin y| \leq \varepsilon \tag{3.11}$$

*for all $x, y \in \mathbb{R}$, then for each $x \in \mathbb{R}$, we have*

$$|f(x)| \leq \frac{1+\sqrt{1+8|d|}}{2} \quad \text{or} \quad |f(-x)| \leq \frac{1+\sqrt{1+8|d|}}{2}. \tag{3.12}$$

*Proof.* As $0<\varepsilon<|d|$, it follows from (3.11) that

$$|f(x-y)-f(x)f(y)|\leq 2|d|$$

for all $x,y\in\mathbb{R}$. According to Theorem 3.3, $f(x-y)=f(x)f(y)$ for all $x,y\in\mathbb{R}$ or (3.12) holds. If (3.12) does not hold, then it would follow from (3.11) that $|d\sin x\sin y|\leq\varepsilon$ for all $x,y\in\mathbb{R}$, which is contrary to our hypothesis, $\varepsilon<|d|$. Therefore, for each $x\in\mathbb{R}$, we get

$$|f(x)|\leq\frac{1+\sqrt{1+8|d|}}{2}\quad\text{or}\quad|f(-x)|\leq\frac{1+\sqrt{1+8|d|}}{2}.$$

□

Next, we shall prove the stability of Equation (1.2). We start with the following lemma.

**Lemma 3.1.** *Let $d$ be a nonzero constant with $d<1$ and $\varepsilon$ be a real constant with $0<\varepsilon<|d|$. If a function $f:\mathbb{R}\to\mathbb{R}$ satisfies the inequality (3.11) for all $x,y\in\mathbb{R}$, then $f$ is bounded.*

*Proof.* If we replace $x,y$ by $z$ and $y-x$ in (3.11) respectively, then we have

$$|f(z-y+x)-f(z)f(y-x)-d\sin z\sin(y-x)|\leq\varepsilon \tag{3.13}$$

for all $x,y,z\in\mathbb{R}$. If we replace $y$ by $y-z$ in (3.11), then we get

$$|f(x-y+z)-f(x)f(y-z)-d\sin x\sin(y-z)|\leq\varepsilon \tag{3.14}$$

for all $x,y,z\in\mathbb{R}$. Using (3.13) and (3.14), we obtain

$$\begin{aligned}
&|f(x)f(y-z)-f(z)f(y-x)+d\sin x\sin(y-z)-d\sin z\sin(y-x)|\\
&=|[f(z-y+x)-f(z)f(y-x)-d\sin z\sin(y-x)]\\
&-[f(x-y+z)-f(x)f(y-z)-d\sin x\sin(y-z)]|\\
&=|f(z-y+x)-f(z)f(y-x)-d\sin z\sin(y-x)|\\
&+|f(x-y+z)-f(x)f(y-z)-d\sin x\sin(y-z)|\\
&\leq 2\varepsilon
\end{aligned} \tag{3.15}$$

for all $x,y,z\in\mathbb{R}$. It follows from (3.15) that

$$\begin{aligned}
&|f(x)[f(y-z)-f(y)f(z)-d\sin y\sin z]+f(x)f(y)f(z)+df(x)\sin y\sin z\\
&-f(z)[f(y-x)-f(x)f(y)-d\sin x\sin y]-f(x)f(y)f(z)-df(z)\sin x\sin y\\
&+d\sin x\sin(y-z)-d\sin z\sin(y-x)|\\
&=|f(x)f(y-z)-f(z)f(y-x)+d\sin x\sin(y-z)-d\sin z\sin(y-x)|\\
&\leq 2\varepsilon
\end{aligned} \tag{3.16}$$

for all $x, y, z \in \mathbb{R}$. Hence, in view of (3.11) and (3.16), we can now get

$$
\begin{aligned}
&|df(x)\sin y\sin z + d\sin x\sin(y-z) - df(z)\sin x\sin y - d\sin z\sin(y-x)| \\
&= |f(x)[f(y-z) - f(y)f(z) - d\sin y\sin z] + f(x)f(y)f(z) + df(x)\sin y\sin z \\
&- [f(y-x) - f(x)f(y) - d\sin x\sin y]f(z) - f(x)f(y)f(z) - df(z)\sin x\sin y \\
&+ d\sin x\sin(y-z) - d\sin z\sin(y-x) - f(x)[f(y-z) - f(y)f(z) - d\sin y\sin z] \\
&+ [f(y-x) - f(x)f(y) - d\sin x\sin y]f(z)| \\
&\leq |f(x)[f(y-z) - f(y)f(z) - d\sin y\sin z] + f(x)f(y)f(z) + df(x)\sin y\sin z \\
&- [f(y-x) - f(x)f(y) - d\sin x\sin y]f(z) - f(x)f(y)f(z) - df(z)\sin x\sin y \\
&+ d\sin x\sin(y-z) - d\sin z\sin(y-x)| + |f(x)|\,|[f(y-z) - f(y)f(z) - d\sin y\sin z]| \\
&+ |[f(y-x) - f(x)f(y) - d\sin x\sin y]|\,|f(z)| \\
&\leq |f(x)f(y-z) - f(z)f(y-x) + d\sin x\sin(y-z) - d\sin z\sin(y-x)| + |f(x)|\varepsilon + |f(z)|\varepsilon \\
&= |[f(x-y+z) - f(z)f(y-x) - d\sin z\sin(y-x)] \\
&- [f(x-y+z) - f(x)f(y-z) - d\sin x\sin(y-z)]| \\
&+ |f(x)|\varepsilon + |f(z)|\varepsilon \\
&\leq 2\varepsilon + |f(x)|\varepsilon + |f(z)|\varepsilon \\
&\leq (2 + |f(x)| + |f(z)|)\varepsilon
\end{aligned}
$$

for all $x, y, z \in \mathbb{R}$. If we set $y = z = \frac{\pi}{2}$ in the above inequality, then

$$\left|df(x) - df\left(\frac{\pi}{2}\right)\sin x - d\cos x\right| \leq \left(2 + |f(x)| + \left|f\left(\frac{\pi}{2}\right)\right|\right)\varepsilon \tag{3.17}$$

for all $x \in \mathbb{R}$.

Next, we assume that $f$ is unbounded. Then there exists a sequence $\{x_n\} \subseteq \mathbb{R}$ such that $f(x_n) \neq 0$ for every $n \in \mathbb{N}$ and $|f(x_n)| \to \infty$ as $n \to \infty$. Replacing $x$ by $\{x_n\}$ in (3.17), and dividing both sides of the resulting inequality by $|f(x_n)|$, and then let $n$ close to infinity, we obtain $|d| \leq \varepsilon$ which is contrary to our hypothesis, say $\varepsilon < |d|$. Therefore, $f$ must be bounded. □

**Theorem 3.5.** *Let $d$ be a nonzero constant with $d < 1$ and $\varepsilon$ be a real constant with $0 < \varepsilon < |d|$. If a function $f : \mathbb{R} \to \mathbb{R}$ satisfies the inequality (3.11) for all $x, y \in \mathbb{R}$, then*

$$\left|f(x) - f\left(\frac{\pi}{2}\right)\sin x - \cos x\right| \leq \frac{2(1 + M_f)}{|d|}\varepsilon$$

*for all $x \in \mathbb{R}$, where $M_f := \sup_{x\in\mathbb{R}} |f(x)|$.*

*Proof.* It follows from Lemma 3.1 that $f$ is bounded, and hence, $M_f := \sup_{x\in\mathbb{R}} |f(x)|$ has to be finite. From the proof of Lemma 3.1, we obtain (3.17) holds. It follows from (3.17) that

$$\left|f(x) - f\left(\frac{\pi}{2}\right)\sin x - \cos x\right| \leq \frac{2(1 + M_f)}{|d|}\varepsilon \tag{3.18}$$

for all $x \in \mathbb{R}$. □

**Remark 3.1.** *Based on Theorems 2.2 and 3.4, it has been established that the newly proposed functional equation (1.2) is stable.*

## 4. Conclusion

To summarize, this paper has determined the general solution $f:\mathbb{R}\to\mathbb{R}$ of the functional equation

$$f(x-y)-f(x)f(y)=d\sin x\sin y$$

for all $x,y\in\mathbb{R}$, where $d$ is a nonzero constant with $d<1$. The solution is given by

$$f(x)=c\sin x+\cos x \quad \text{or} \quad f(x)=-c\sin x+\cos x,$$

where $c=\sqrt{1-d}$. In addition, the stability of this functional equation has been examined through a theorem addressing its superstability. Specifically, it has been established that for a nonzero constant $d$ with $d<1$ and a real constant $\varepsilon$ satisfying $0<\varepsilon<|d|$, if a function $f:\mathbb{R}\to\mathbb{R}$ fulfills the functional inequality

$$|f(x-y)-f(x)f(y)-d\sin x\sin y|\leq\varepsilon$$

for all $x,y\in\mathbb{R}$, then for each $x\in\mathbb{R}$, we have

$$|f(x)|\leq\frac{1+\sqrt{1+8|d|}}{2} \quad \text{or} \quad |f(-x)|\leq\frac{1+\sqrt{1+8|d|}}{2},$$

and the following estimate holds

$$|f(x)-f\Big(\frac{\pi}{2}\Big)\sin x-\cos x|\leq\frac{2(1+M_f)}{|d|}\,\varepsilon,$$

where $M_f:=\sup_{x\in\mathbb{R}}|f(x)|$.

## Acknowledgments

The authors are thankful to the referees and the editor for their comments and suggestions, which substantially improve the initial version of the paper.

**Financial disclosure.** This work was supported by Thammasat University Research Unit in Fixed Points and Optimization.

## References

[1] S. M. Ulam: *Problems in Modern Mathematics*, Science Editions, Wiley, New York (1964), Chapter VI.
[2] D. H. Hyers: *On the stability of the linear functional equation*, Proc. Natl. Acad. Sci. USA, **27** (1941), 222–224.
[3] T. Aoki: *On the stability of the linear transformation in Banach spaces*, J. Math. Soc. Jpn., **2** (1950), 64–66.
[4] M. Th. Rassias: *On the stability of the linear mapping in Banach spaces*, Proc. Am. Math. Soc., **72** (1978), 297–300.
[5] Z. Gajda: *On stability of additive mappings*, Int. J. Math. Math. Sci., **14** (3) (1991), 431–434.
[6] M. Th. Rassias: *On a modified Hyers–Ulam sequence*, J. Math. Anal. Appl., **158** (1) (1991), 106–113.
[7] S. Butler: *Problem no. 11030*, Am. Math. Mon., **110** (2003), 637–639.
[8] M. Th. Rassias: *Solution of a functional equation problem of Steven Butler*, Octogon Math. Mag., **12** (2004), 152–153.
[9] S. M. Jung, M. Th. Rassias and C. Mortici: *On a functional equation of trigonometric type*, Appl. Math. Comput., **252** (2015), 293–303.

Jiraporn Kunrattanaworawong
Vajiravudh College
Ratchawithi Road, 10300, Bangkok, Thailand
*Email address*: jiraporn.kunr@gmail.com

Wutiphol Sintunavarat
Thammasat University
Faculty of Science and Technology
Department of Mathematics and Statistics
Phaholyothin Road, 12120, Pathum Thani, Thailand
*Email address*: wutiphol@mathstat.sci.tu.ac.th

MODERN MATHEMATICAL METHODS
**3** (2025), No. 2, pp. 92-100
https://modernmathmeth.com/
ISSN 3023 - 5294

*Research Article*

# Normalized solutions to the fractional Schrödinger equations with potential and saturable nonlinearity

WEN LIAO AND QIONGFEN ZHANG*

ABSTRACT. This paper is concerned with the existence of normalized solutions to a kind of fractional Schrödinger equations driven by a fractional operator with a parametric potential term and a saturable nonlinear term. We achieve the minimization of the energy functional and prove the existence of normalized solutions for the equation under specific conditions that we assume for the potential and saturable nonlinearity.

**Keywords:** Normalized solutions, Schrödinger equations, saturable nonlinearity, potential function.

**2020 Mathematics Subject Classification:** 34C37, 35A15, 37J45, 47J30.

## 1. INTRODUCTION

This paper studies the existence of normalized solutions for the following fractional Schrödinger equations

$$\imath\frac{\partial\varphi}{\partial t} = (-\Delta)^s\varphi + \mu V(x)\varphi - \mu\frac{\psi(x)+\varphi^2}{1+\psi(x)+\varphi^2}\varphi, \ \text{ in } \mathbb{R}^N, \tag{1.1}$$

where $(-\Delta)^s$ is the fractional Laplacian, $\varphi = \varphi(t,x)$, $s \in (0.25, 1)$, $2s < N < 4s$, $\mu > 0$, $\lambda \in \mathbb{R}$ is a parameter, the potential function $V : \mathbb{R}^N \to [0, +\infty)$ is continuous and bounded. The function $\psi$ is bounded in $\mathbb{R}^N$. A solution of the problem (1.1) is called a standing wave solution if it has the form $\varphi(t,x) = \ell^{i\lambda t}$. Indeed, $u$ is a time-independent and real-valued function that satisfies the following fractional Schrödinger equation:

$$(-\Delta)^s u + \mu V(x)u + \lambda u = \mu\frac{\psi(x)+u^2}{1+\psi(x)+u^2}u, \ \text{ in } \mathbb{R}^N, \tag{1.2}$$

is defined by

$$(-\Delta)^s u(x) = C(N,s)P.V.\int_{\mathbb{R}^N}\frac{u(x)-u(y)}{|x-y|^{N+2s}}dy,$$

here, the symbol $P.V.$ is the Cauchy principal value and $C(N,s)$ is a constant that depends on $N$ and $s$. For more information, please see [19] and the reference therein. $(-\Delta)^s$ originates from applied scientific fields such as obstacle problems, phase transition phenomena, fractional quantum mechanics, and Markov processes (see [10, 19, 20, 21]). We are examining the feasibility of solving the constraint minimization problem described as below:

$$I_a = \inf_{u\in S(a)} E(u),$$

Received: 07.02.2025; Accepted: 13.08.2025; Published Online: 20.08.2025
*Corresponding author: Qiongfen Zhang; qfzhangcsu@163.com
DOI: 10.64700/mmm.53

where $a$ is a positive constant and

$$S(a) = \{u \in W : \|u\|_2^2 = \int_{\mathbb{R}^N} |u|^2 dx = a^2\}. \tag{1.3}$$

Moreover, the weighted fractional Sobolev space $W$ is defined as

$$W = \{u \in H^s(\mathbb{R}^N) : \int_{\mathbb{R}^N} \mu V(x)|u|^2 dx < +\infty\}, \tag{1.4}$$

with norm

$$\|u\|_W = \left(\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + \int_{\mathbb{R}^N} \mu V(x)|u|^2 dx\right)^{\frac{1}{2}}. \tag{1.5}$$

We will study the energy functional $E : W \to \mathbb{R}$, which is defined by

$$E(u) = \frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + \frac{1}{2}\int_{\mathbb{R}^N} \mu V(x)|u|^2 dx - \frac{\mu}{2}\int_{\mathbb{R}^N} \left[u^2 - \ln\left(1 + \frac{u^2}{1+\psi(x)}\right)\right] dx. \tag{1.6}$$

In addition, the fractional Sobolev space $H^s(\mathbb{R}^N)$ is given by:

$$H^s(\mathbb{R}^N) = \{u \in L^2(\mathbb{R}^N) : \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx = \int_{\mathbb{R}^{2N}} \frac{|u(x) - u(y)|^2}{|x-y|^{N+2s}} dxdy < \infty\}, \tag{1.7}$$

with norm

$$\|u\|_{H^s} = \left(\int_{\mathbb{R}^N} |u|^2 dx + \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx\right)^{\frac{1}{2}}. \tag{1.8}$$

In recent years, the following fractional Schrödinger equation has been studied by many scholars:

$$(-\Delta)^s u + \mu u + \lambda V(x)u = h(u), \text{ in } \mathbb{R}^N. \tag{1.9}$$

When $h(u) = |u|^{p-2}u$, Laskin [11] obtained a couple of normalized solution by using a fiber map and concentration-compactness principle. Zuo [25] obtained minimization of the energy functional associated with the problem (1.9). When $h(u) = u \log u^2$, Alves [1, 2] employed minimization techniques and used the Lusternik-Schnirelmann category to prove existence of multiple normalized solutions. The solution to the problem (1.9) can be studied from two aspects. On the one hand, one can choose the fixed frequency $\mu \in \mathbb{R}$ and studied the existence of nontrivial solution of problem (1.9). On the other hand, taking $\mu$ as unknown, then we can look to prescribed $L^2$-norm solutions. A large number of people have conducted comprehensive studies utilizing variational and topological methods, see [5, 6, 12] and the references therein. As we all know, some previous works have explored this aspect, although they did not incorporate a potential term, see [3, 13, 17, 23].

We want to point out that the nonlinearity $f(x,u) = \mu\frac{\psi(x)+u^2}{1+\psi(x)+u^2}u$ for $\mu > 0$ is usually called saturable nonlinear term, which is used to describe photo refractive media [8]. Lin [14] firstly studied normalized solution for the Schrödinger equation with saturable nonlinearity. Later, Lin [15] used a convexity argument to expand the result when $g(x)$ becomes nonzero. Very recently, Sun [22] used variational methods to prove the existence of normalized solutions for a kind of quasilinear Schrödinger equations.

Inspired by the above paper, our objective is to expand the current findings to the fractional Schrödinger equations and investigate the existence of normalized solutions of fractional Schrödinger equations with a saturable nonlinear term $\mu\frac{\psi(x)+u^2}{1+\psi(x)+u^2}u$ and an external potential $V$. However, due to the presence of the saturable nonlinear term and the potential function,

we have to face several difficulties, such as the scaling result of the energy functional $E(\cdot)$ and the strong convergence of the minimizing sequence in the weighted fractional Sobolev space $W$. In view of this, it is necessary to introduce some inequality and new ideas.

In order to present our main result, we require the following assumptions on $V(x)$ and $\psi(x)$.

$(V_1)$ There exists a positive constant $C_0 > 0$ such that the following measure of the set $\Omega$ is finite, where

$$\Omega = \{x \in \mathbb{R}^N : V(x) < C_0\}. \tag{1.10}$$

$(V_2)$ $V \in L^\infty(\mathbb{R})$.

$(\Psi_1)$ The function $\psi(x)$ is radially symmetric with $-1 < \psi_1 \le \psi(|x|) \le \psi_2$ for $x \in \mathbb{R}^N$, where $\psi_1$, $\psi_2$ are two constants.

In conclusion, we will now present our primary result.

**Theorem 1.1.** *Assume that the condition $(V_1), (V_2)$ and $(\Psi_1)$ holds, and let $2s < N < 4s$, $p \in (2, \min\{\frac{N}{N-2s}, 2 + \frac{4s}{N}\})$. Then, for each $a > 0$, there exist a sufficiently large positive $\mu^* > 0$ and $\Upsilon = \Upsilon(\mu_*, a) > 0$ such that $I_a < 0$ for all $\mu > \mu^*$ satisfying $\mu\|V\|_\infty < \Upsilon$. Moreover, the infimum $I_a$ is attained by a function $u \in S(a)$, which is a normalized solution of (1.1) with $\lambda = \lambda_a$ as a Lagrange multiplier.*

## 2. Auxiliary Lemmas and Proof of Theorem 1.1

In this section, we give out some auxiliary lemmas, and then we give the proof of Theorem 1.1.

**Lemma 2.1** ([16], Lemma 2.2). *For each $2 < p \le \min\{4, 2^*\}$ ($2^* = \infty$ if $N = 1, 2$; $2^* = \frac{2N}{N-2}$ if $N \ge 3$), there exists a positive constant*

$$A_p = \begin{cases} \frac{1}{2}, & \text{if } p = 4; \\ \frac{p^{\frac{p-2}{2}}(4-p)^{\frac{4-p}{2}}}{2p}, & \text{if } 2 < p \le \min\{4, 2^*\} \text{ and } p \ne 4, \end{cases}$$

*such that*

$$t^2 - \ln(1 + \frac{t^2}{1+\psi(x)}) \le \frac{\psi(x)}{1+\psi(x)} t^2 + \frac{A_p}{(1+\psi(x))^{\frac{p}{2}}} t^p \text{ for all } \ge 0. \tag{2.11}$$

**Lemma 2.2** ([9], Fractional Gagliardo-Nirenberg Inequality). *For each $p \in (2, \frac{2N}{N-2s})$, there exists a constant $C(s, N, p)$, such that*

$$\int_{\mathbb{R}^N} |u|^p dx \le C(s, N, p)\Big(\int_{\mathbb{R}^N} |u|^2 dx\Big)^{\frac{p}{2} - \frac{N(p-2)}{4s}} \Big(\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx\Big)^{\frac{N(p-2)}{4s}}. \tag{2.12}$$

**Lemma 2.3.** *Assuming that condition $(\Psi_1)$ is satisfied, it can be concluded that the energy functional $E(u)$ is both coercive and bounded from below on $S(a)$ for all $a > 0$.*

*Proof.* For any $u \in S(c)$, observe that condition $(\Psi_1)$ guarantees $1 + \psi(x) \ge 1 + \psi_1(x) > 0$ for all $x \in \mathbb{R}^N$, this make the logarithmic term well defined. And controlling the range of the $\psi(x)$ make the function $\frac{1}{1+\psi(x)}$ bounded. It follow from Lemmas 2.1 and 2.2 that

$$\begin{aligned} E(u) &= \frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + \frac{1}{2}\int_{\mathbb{R}^N} \mu V(x)|u|^2 dx - \frac{\mu}{2}\int_{\mathbb{R}^N} \left[u^2 - \ln\left(1 + \frac{u^2}{1+\psi(x)}\right)\right] dx \\ &\ge \frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx - C(s, N, p, A_p)\frac{\mu}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^{\frac{N(p-2)}{4s}} dx - \frac{\mu a^2}{2}. \end{aligned}$$

Since $p \in (2, 2+\frac{4s}{N})$, we infer that $0 < \frac{N(p-2)}{4s} < 1$. We can therefore conclude that the energy functional $E(u)$ is coercive and bounded from below on $S(a)$. We complete the proof. □

Even if it is not explicitly stated, in the following series of lemmas we always make the assumption that $(u_n)_{n\in\mathbb{N}} \subset S(a)$ is a minimizing sequence for $I_a$.

**Lemma 2.4.** *The sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $W$.*

*Proof.* It can be deduced from the definition of a minimizing sequence that

$$I_a = \lim_{n\to\infty} E(u_n). \tag{2.13}$$

According to Lemma 2.3, it is established that $(\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2)_{n\in\mathbb{N}}$ is a bounded sequence for $I_a$. As a result, the sequence $(\int_{\mathbb{R}^N} \mu V(x)|u_n|^2 dx)_{n\in\mathbb{N}}$ is also bounded. In a word, $(u_n)_{n\in\mathbb{N}}$ is a bounded sequence in $W$. This concludes the proof. □

**Lemma 2.5.** *Assuming that conditions $(V_1)$ and $(V_2)$ are satisfied, then for any $\mu_2 > \mu > 0$ there exists $\Upsilon = \Upsilon(\mu_2, a) > 0$ such that $I_a < 0$ if $\lambda\|V\|_\infty < \Upsilon$.*

*Proof.* Let $(u_n) \subset S(a)$ be a minimizing sequence for $I_a = \inf_{u\in S(a)} E(u)$. Then $(u_n)$ is bounded on $S(a)$ by Lemma 2.3. It is evident that the embedding $W(\mathbb{R}^N) \hookrightarrow L^h(\mathbb{R}^N)$ is compact for any $h \in [2, 2_s^*]$ (see Wang et al. [4]). Since there exists $\overline{u} \in H$ such that $u_n \rightharpoonup \overline{u}$ weakly in $W(\mathbb{R}^N)$, $u_n \rightharpoonup \overline{u}$ strongly in $L^h(\mathbb{R}^N)$. According to the Lagrange multipliers rule, there exists $\lambda_n \in \mathbb{R}^N$ such that

$$(-\Delta)^s u_n + \mu V(x)u_n + \lambda_n u_n = \mu \frac{\psi(x) + u_n^2}{1 + \psi(x) + u_n^2} u_n, \text{ in } \mathbb{R}^N. \tag{2.14}$$

In particular, we have

$$\begin{aligned}\lambda_n a^2 &= -\frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx - \frac{1}{2}\int_{\mathbb{R}^N} \mu V(x)|u_n|^2 dx + \int_{\mathbb{R}^N} \mu \frac{\psi(x) + u_n^2}{1 + \psi(x) + u_n^2} u_n^2 dx + o(1) \\ &\leq \int_{\mathbb{R}^N} \mu \frac{\psi(x) + u_n^2}{1 + \psi(x) + u_n^2} u_n^2 dx + o(1) < \mu a + o(1),\end{aligned}$$

which implies that $\{\lambda_n\}$ is bounded, thus allows us to use Bolzano-Weierstrass to extract a convergent subsequence of $\lambda_n$, then restrict all further analysis to this subsequence. And then we make the assumption that $\lambda_{n_r} \to \overline{\lambda}$ as $n \to \infty$.

For a fixed $u \in S(a)$, we choose a constant $\mu_1 > 0$ such that for $\mu > \mu_1$,

$$\frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + \frac{1}{2}\int_{\mathbb{R}^N} \mu V(x)|u|^2 dx - \frac{\mu}{2}\int_{\mathbb{R}^N} [u^2 - \ln(1 + \frac{u^2}{1+\psi(x)})]dx < 0, \tag{2.15}$$

which implies that $I_a < 0$. Furthermore, there exists $\mu_2 \geq \mu_1 > 0$ such that

$$\frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx - \frac{\mu_2}{2}\int_{\mathbb{R}^N} [u^2 - \ln(1 + \frac{u^2}{1+\psi(x)})]dx = B_{\mu_2} < 0, \tag{2.16}$$

hence, fixed the number $\Upsilon = \frac{-B_{\mu_2}}{a^2}$ and in light of the assumption $\lambda\|V\|_\infty < \Upsilon$, we infer that $E(u_n) < 0$, which implies that $I_a < 0$. Consequently, we complete the proof. □

**Lemma 2.6.** *Assuming that condition $(V_2)$ is satisfied, then for any $a > 0$, there exist $\Upsilon(a) > 0$ and $\eta > 0$ such that*

$$\liminf_{n\to\infty} \int_{\mathbb{R}^N} |u_n|^p dx \geq \eta \text{ if } \mu\|V\|_\infty < \Upsilon. \tag{2.17}$$

*Proof.* From Lemma 2.5, it is easy to find $\Upsilon(a) > 0$ and $\eta > 0$ such that $I_a < -\eta$ if $\mu\|V\|_\infty < \Upsilon$. Now, since$(u_n) \subset S(a)$ is a minimizing sequence for $I_a$, then we have

$$\begin{aligned} I_a + o(1) =& E(u_n) \\ =& \frac{1}{2}\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx + \frac{1}{2}\int_{\mathbb{R}^N} \mu V(x)|u_n|^2 dx \\ &- \int_{\mathbb{R}^N} \frac{\mu\psi(x)}{2(1+\psi(x))}|u_n|^2 dx - \int_{\mathbb{R}^N} \frac{\mu A_p}{2(1+\psi(x))^{\frac{p}{2}}}|u_n|^p dx. \end{aligned}$$

Furthermore, we obtain that

$$-\eta + o(1) \geq -\frac{\mu a^2}{2} - \int_{\mathbb{R}^N} \frac{\mu A_p}{2(1+\psi(x))^{\frac{p}{2}}}|u_n|^p dx. \tag{2.18}$$

Hence (2.17) holds. We complete the proof. □

In the following lemma, we establish a quantitative relationship between the two ordered values of $a$ and the corresponding values of $I_a$.

**Lemma 2.7.** *Assuming that condition $(V_2)$ is satisfied and if $0 < a_1 < a_2$ is achieved, then $\frac{I_{a_2}}{a_2^2} < \frac{I_{a_1}}{a_1^2}$.*

*Proof.* To prove this lemma, we borrow the ideas form Lv and Li [18]. Let $\xi > 1$ such that $a_2 = \xi a_1$, and let $(u_n) \subset S(a_1)$ be a minimizing sequence with respect to $I_{a_1}$, that is

$$E(u_n) \to I_{a_1}, \text{ as n} \to \infty. \tag{2.19}$$

Setting $v_n = \xi u_n$, we have that $(v_n) \subset S(a_2)$, and so

$$I_{a_2} \leq E(v_n) = \xi^2 E(u_n) - \frac{\mu\xi^2}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{u_n^2}{1+\psi(x)})dx + \frac{\mu}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{(\xi u_n)^2}{1+\psi(x)})dx. \tag{2.20}$$

Letting $n \to \infty$, it follows from $\xi > 1$ that

$$\frac{\mu}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{(\xi u_n)^2}{1+\psi(x)})dx - \frac{\mu\xi^2}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{u_n^2}{1+\psi(x)})dx < 0, \tag{2.21}$$

which implies that

$$I_{a_2} \leq \xi^2 E(u_n) - \frac{\mu\xi^2}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{u_n^2}{1+\psi(x)})dx + \frac{\mu}{2}\int_{\mathbb{R}^N} \ln(1 + \frac{(\xi u_n)^2}{1+\psi(x)})dx < \xi^2 E(u_n), \tag{2.22}$$

that is

$$I_{a_2} a_1^2 < I_{a_1} a_2^2. \tag{2.23}$$

The proof is completed. □

**Lemma 2.8.** *Assuming that condition $(V_2)$ is satisfied and if $u_n \rightharpoonup u$ in $W$, $u_n(x) \to u(x)$ a.e. in $\mathbb{R}^N$ and $u \neq 0$, then $u \in S(a)$, $E(u) = I_a$ and $u_n \to u$ in $W$.*

*Proof.* We note that if $\|u\|_2 = m \neq a$, in light of Fatou's lemma and the assumption $u \neq 0$, then we get that $m \in (0, a)$. From the continuity of embedding $W \hookrightarrow L^h(\mathbb{R}^N)$ for any $h \in [2, 2_s^*]$ and two kinds of Brézis-Lieb lemmas in [7, 24], we get that

$$\begin{aligned} \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx &= \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} (u_n - u)|^2 dx + \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + o(1), \\ \int_{\mathbb{R}^N} |u_n|^2 dx &= \int_{\mathbb{R}^N} |u_n - u|^2 dx + \int_{\mathbb{R}^N} |u|^2 dx + o(1). \end{aligned} \tag{2.24}$$

Let $v_n = u_n - u$, $\|v_n\|_2 = r_n$ and assume that $\|v_n\|_2 \to r$, by (2.14), we infer that $a^2 = m^2 + r^2$ and $r_n \in (0, a)$ for sufficiently big $n$, which implies that

$$\begin{aligned} I_a + o(1) &= E(u_n) \\ &= E(v_n) + E(u) + o(1) \\ &\geq I_{r_n} + I_m + o(1) \\ &\geq \frac{r_n^2}{a^2} I_a + I_m + o(1). \end{aligned}$$

Letting $n \to \infty$, we can derive the inequality

$$I_a \geq \frac{r^2}{a^2} I_a + I_m. \tag{2.25}$$

Using the fact that $m \in (0, a)$ and Lemma 2.7 in the above inequality, we can infer that

$$I_a > \frac{r^2}{a^2} I_a + \frac{m^2}{a^2} I_a = I_a. \tag{2.26}$$

This inequality presents a contradiction. Therefore $\|u\|_2 = a$, namely, $u \in S(a)$. It can be deduced from the premises $\|u_n\|_2 = \|u\|_2 = a$ and $u_n \rightharpoonup u$ in $L^2(\mathbb{R}^N)$ (since $W \hookrightarrow L^2(\mathbb{R}^N)$ is a continuous embedding) that $u_n \to u$ in $L^2(\mathbb{R}^N)$. Furthermore, the combination of this deduction with the interpolation inequality allows us to conclude that $u_n \to u$ in $L^p(\mathbb{R}^N)$. In addition, since $\int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx + \int_{\mathbb{R}^N} \mu V(x)|u_n|^2 dx$ is convex and continuous in $W$, it can be inferred that it is a weak lower semicontinuous, namely

$$\liminf_{n\to+\infty} \Big( \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx + \int_{\mathbb{R}^N} \mu V(x)|u_n|^2 dx \Big) \geq \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u|^2 dx + \int_{\mathbb{R}^N} \mu V(x)|u|^2 dx. \tag{2.27}$$

According to $I_a = \lim\limits_{n\to\infty} E(u_n)$, we infer that $I_a \geq E(u)$, reuse the definition of $I_a$ and $u \in S(a)$, we get that $I_a = E(u)$, and hence

$$\lim_{n\to\infty} E(u_n) = E(u). \tag{2.28}$$

Finally, based on the convergence $u_n \to u$ in $L^2(\mathbb{R}^N)$ and $u_n \to u$ in $L^p(\mathbb{R}^N)$, we can get that $u_n \to u$ in $W$. Thus, we have successfully completed the proof. □

**Lemma 2.9.** *Let $p \in [1, \frac{N}{N-2}]$. Assuming that conditions $(V_1)$ and $(V_2)$ are satisfied, and let $R > 0$ and $\mu^* > 0$ are given, for any $\mu > \mu^*$, we could get*

$$\limsup_{n\to+\infty} \int_{B_R^c(0)} |u_n|^p \leq \frac{\eta}{2},$$

*where $B_R^c(0) = \{x \in \mathbb{R}^N : |x| > R\}$, $a > 0$ is as given in Lemma 2.5.*

*Proof.* We now prove this lemma by a contradiction. We borrow the ideas from Bartsch and Wang [4], for $R > 0$, we consider two sets

$$\Phi(R) := \{x \in \mathbb{R}^N : |x| > R, V(x) \geq C_0\} \text{ and } \Theta(R) := \{x \in \mathbb{R}^N : |x| > R, V(x) < C_0\}. \tag{2.29}$$

Let $C > 0$ represent a constant with a variable value that may change from line to line. By Lemma 2.4, we know that $\|u_n\|_W^2 \leq C$ for all $n \in \mathbb{N}$, hence we get that

$$\begin{aligned}\int_{\Phi(R)} u_n^2 dx &\leq \frac{1}{\mu C_0 + 1} \int_{\mathbb{R}^N} (\mu V(x) + 1) u_n^2 dx \\ &\leq \frac{1}{\mu C_0 + 1} \Big( \int_{\mathbb{R}^N} (\mu V(x) + 1) u_n^2 + \int_{\mathbb{R}^{2N}} |(-\Delta)^{\frac{s}{2}} u_n|^2 \Big) dx \\ &= \frac{1}{\mu C_0 + 1} \|u_n\|_W^2 \\ &\leq \frac{C}{\mu C_0 + 1}. \end{aligned} \tag{2.30}$$

Clearly, the term on the right-hand side of the above inequality can be arbitrarily small when $\mu > \mu^*$ (large enough). According to the Hölder inequality, the continuous embedding $H^s(\mathbb{R}^N) \hookrightarrow L^{2p}(\mathbb{R}^N)$ for any $p \in [1, \frac{N}{N-2}]$ (see ([19], Theorem 6.5)) and condition $(V_1)$, we obtain

$$\begin{aligned}\int_{\Theta(R)} u_n^2 dx &\leq \left( \int_{\Theta(R)} |u_n|^{2p} dx \right)^{\frac{1}{p}} \left( \int_{\Theta(R)} dx \right)^{\frac{1}{q}} \\ &\leq \left( \int_{\mathbb{R}^N} |u_n|^{2p} dx \right)^{\frac{1}{p}} \left( \int_{\Theta(R)} dx \right)^{\frac{1}{q}} \\ &= C \|u_n\|_{H^s}^2 |\Theta(R)|^{\frac{1}{q}} \\ &\leq C |\Theta(R)|^{\frac{1}{q}}. \end{aligned} \tag{2.31}$$

With the help of condition $(V_1)$, the term on the right-hand side of the above inequality can also be arbitrarily small if $R$ is sufficiently large since $|\Theta(R)| \to 0$ as $R \to +\infty$, where $\frac{1}{p} + \frac{1}{q} = 1$. It follows from Lemma 2.2 that

$$\begin{aligned}\int_{B_R^c(0)} |u_n|^p dx &\leq C(s, N, p) \left( \int_{B_R^c(0)} |u_n|^2 dx \right)^{\frac{p}{2} - \frac{N(p-2)}{4s}} \left( \int_{B_R^c(0)} |(-\Delta)^{\frac{s}{2}} u_n|^2 dx \right)^{\frac{N(p-2)}{4s}} \\ &\leq C(s, N, p) C^{\frac{N(p-2)}{4s}} \|u_n\|_W^{\frac{N(p-2)}{4s}} \left( \int_{\Phi(R)} u_n^2 dx + \int_{\Theta(R)} u_n^2 dx \right)^{\frac{p}{2} - \frac{N(p-2)}{4s}} \\ &\leq C \left( \int_{\Phi(R)} u_n^2 dx + \int_{\Theta(R)} u_n^2 dx \right)^{\frac{p}{2} - \frac{N(p-2)}{4s}}. \end{aligned} \tag{2.32}$$

To sum up, we prove the lemma. □

**Lemma 2.10.** *Assuming the conditions $(V_1)$ and $(V_2)$ are satisfied, and let $\mu^{**} > 0$ is given, for any $\mu > \mu^{**}$ the sequence $(u_n)_{n\in\mathbb{N}}$ admits a nontrivial weak limit $u$ in $W$.*

*Proof.* By Lemma 2.4, it is know that there exists $u \in W$ and a subsequence of $(u_n)_{n\in\mathbb{N}}$, which is still denoted as itself, such that

$$u_n \rightharpoonup u \text{ in } W, u_n(x) \to u(x) \text{ a.e. in } \mathbb{R}^N. \tag{2.33}$$

Now, employing the method of argument by contradiction, we make the assumption that $u = 0$ for some $\mu > \mu^*$ as given in Lemma 2.9. Because the compactness of the embedding over the

bounded domain, we have $u_n \to 0$ in $L^p(B_R(0))$ for any $R > 0$. In light of Lemmas 2.6 and 2.9, we have

$$\begin{aligned}\eta &\leq \liminf_{n\to+\infty} \int_{\mathbb{R}^N} |u_n|^p dx \\ &= \liminf_{n\to+\infty} \int_{B_R^c(0)} |u_n|^p dx \\ &\leq \limsup_{n\to+\infty} \int_{B_R^c(0)} |u_n|^p dx \\ &\leq \frac{\eta}{2}, \end{aligned} \tag{2.34}$$

this inequality presents a contradiction. Therefore, we can conclude that there exists $\mu^* \leq \mu^{**}$ such that $u$ is nontrivial for any $\mu > \mu^{**}$. □

In the end, we present the proof of our main result.

*Proof of Theorem 1.1.* By Lemmas 2.4 and 2.10, we know that there exists a minimizing sequence $(u_n)_{n\in\mathbb{N}} \subset S(a)$ for $I_a$, which is bounded in $W$ and its weak limit $u$ is nontrivial. According to the Lemma 2.8, we have that $u \in S(a)$, $E(u) = I_a$ and $u_n \to u$ in $W$. Then, we use the Lagrange multiplier method, there exists $\lambda_a \in \mathbb{R}$ solving the equation

$$E'(u) + \lambda_a J'(u) = 0 \text{ in } \mathrm{W}^*, \tag{2.35}$$

where $W^*$ is the dual space of $W$ and $J : W \to \mathbb{R}$ is defined by

$$J(u) = \|u\|_2, u \in W. \tag{2.36}$$

By (2.13), we deduce that

$$(-\Delta)^s u + \lambda_a u + \mu V(x)u - \mu \frac{\psi(x) + u^2}{1 + \psi(x) + u^2} u = 0 \text{ in } \mathbb{R}^{\mathrm{N}}. \tag{2.37}$$

We complete the proof of Theorem 1.1. □

## Acknowledgments

The authors would like to thank the referees for their useful suggestions which have significantly improved the paper.

**Financial disclosure.** This work is supported by the National Natural Science Foundation of China (No. 11961014) and Guangxi Natural Science Foundation (2021GXNSFAA196040).

## References

[1] C. O. Alves, N. V. Thin: *On existence of multiple normalized solutions to a class of elliptic problems in whole $\mathbb{R}^N$ via Lusternik-Schnirelmann category*, SIAM J. Math. Anal., **55** (2023), 1264–1283.
[2] C. O. Alves, C. Ji: *Multiple normalized solutions to a logarithmic Schrödinger equation via Lusternik–Schnirelmann category*, J. Geom. Anal., **34** (2024), Article ID: 198.
[3] L. Appolloni, S. Secchi: *Normalized solutions for the fractional NLS with mass supercritical nonlinearity*, J. Differential Equations, **286** (2021), 248–283.
[4] T. Bartsch, Z. Q. Wang: *Existence and multiplicity results for some superlinear elliptic problems on $\mathbb{R}^n$*, Comm. Partial Differential Equations, **20** (1995), 1725–1741.
[5] H. Berestycki, P. L. Lions: *Nonlinear scalar field equations I: existence of a ground state*, Arch. Ration. Mech. Anal., **82** (1983), 313–345.
[6] H. Berestycki, P. L. Lions: *The concentration-compactness principle in the calculus of variations. The locally compact case, part 1*, Ann. Inst. Henri. Poincaré Anal. Non Linéaire, **11** (1984), 109–145.

[7] H. Brézis, E. Lieb: *A relation between pointwise convergence of functions and convergence of functionals*, Proc. Am. Math. Soc., **88** (3) (1983), 486–490.
[8] N. K. Efremidis, J. Hudock, D. N. Christodoulides, J. W. Fleischer, O. Cohen and M. Segev: *Two-dimensional optical lattice solitons*, Phys Rev Lett., **91** (2003), 213–906.
[9] R. L. Frank, E. Lenzmann and L. Silvestre: *Uniqueness of radial solutions for the fractional Laplacian*, Comm. Pure Appl. Math., **69** (2016), 1671–1726.
[10] Q. Y. Guan, Z. M. Ma: *Boundary problems for fractional Laplacians*, Stoch. Dyn., **5** (2005), 385–424.
[11] N. Laskin: *Fractional Schrödinger equation*, Phys. Rev. E, **66** (2002), Article ID: 56108.
[12] S. Liang, P. Pucci and B. Zhang: *Multiple solutions for critical Choquard-Kirchhoff type equations*, Adv. Nonlinear Anal., **10** (2021), 400–419.
[13] Q. Li, W. Zou: *The existence and multiplicity of the normalized solutions for fractional Schrödinger equations involving Sobolev critical exponent in the $L^2$-subcritical and $L^2$-supercritical cases*, Adv. Nonlinear Anal., **11** (2022), 1531–1551.
[14] T. C. Lin, M. R. Belić, M. S. Petrović and G. Chen: *Ground states of nonlinear Schrödinger systems with saturable nonlinearity in $\mathbb{R}^2$ for two counterpropagating beams*, J. Math. Phys., **55** (2014), Article ID: 011505.
[15] T. C. Lin, X. Wang and Z. Q. Wang: *Orbital stability and energy estimate of ground states of saturable nonlinear Schrödinger equations with intensity functions in $\mathbb{R}^2$*, J. Diff. Equ., **263** (2017), 2750–2786.
[16] T. C. Lin, T. F. Wu: *Multiple positive solutions of saturable nonlinear Schrödinger equations with intensity functions*, Discrete Contin. Dyn. Syst., **70** (2020), 2165–2187.
[17] H. Luo, Z. Zhang: *Normalized solutions to the fractional Schrödinger equations with combined nonlinearities*, Calc. Var. Partial Differ. Equ., **59** (2020), Article ID: 143.
[18] Y-C Lv, G-D Li: *Multiplicity of normalized solutions to a fractional logarithmic Schrödinger equation*,Fractal fract., **8** (7) (2024), Article ID: 391.
[19] E. D. Nezza, G. Palatucci and E. Valdinoci: *Hitchhiker's guide to the fractional Sobolev spaces*, Bull. Sci. Math., **136** (2012), 521–573.
[20] L. Silvestre: *Regularity of the obstacle problem for a fractional power of the Laplace operator*, Comm. Pure Appl. Math., **60** (2007), 67–112.
[21] Y. Sire, E. Valdinoci: *Fractional Laplacian phase transitions and boundary reactions: a geometric inequality and a symmetry result*, J. Funct. Anal., **256** (2009), 1842–1864.
[22] Y. Zhang, J. Sun: *Normalized solutions of quasilinear Schrödinger equations with saturable nonlinearity*, Appl. Math. Lett., **138** (2023), Article ID: 108531.
[23] M. Zhen, B. Zhang: *Normalized ground states for the critical fractional NLS equation with a perturbation*, Rev. Mat. Complut., **35** (2022), 89–132.
[24] J. Zuo, T. An and A. Fiscella: *A critical Kirchhoff-type problem driven by a $p(\cdot)$-fractional Laplace operator with variable $s(\cdot)$-order*, Math. Methods Appl. Sci., **44** (2021), 1071–1085.
[25] J. B. Zuo, C. G. Liu and C. Vetro: *Normalized solutions to the fractional Schrödinger equation with potential*, Mediterr. J. Math., **216** (2023), 1660–5446.

Wen Liao
Guilin University of Technology
School of Mathematics and Statistics
Guilin, Guangxi 541004, PR China

Guangxi Colleges and Universities
Key Laboratory of Applied Statistics
Guilin, Guangxi 541004, PR China
*Email address*: 1936692552@qq.com

Qiongfen Zhang
Guilin University of Technology
School of Mathematics and Statistics
Guilin, Guangxi 541004, PR China

Guangxi Colleges and Universities
Key Laboratory of Applied Statistics
Guilin, Guangxi 541004, PR China
*Email address*: qfzhangcsu@163.com

MODERN MATHEMATICAL METHODS
**3** (2025), No. 2, pp. 101-109
https://modernmathmeth.com/
ISSN 3023 - 5294

*Research Article*

# Critical issues with the Pearson's chi-square test

VLADIMIR GURVICH AND MARIYA NAUMOVA*

ABSTRACT. Pearson's chi-square tests are among the most commonly applied statistical tools across a wide range of scientific disciplines, including medicine, engineering, biology, sociology, marketing and business. However, its usage in some areas is not correct. For example, the chi-square test for homogeneity of proportions (that is, comparing proportions across groups in a contingency table) is frequently used to verify if the rows of a given nonnegative $m \times n$ (contingency) matrix $A$ are proportional. The null hypothesis $H_0$: "$m$ rows are proportional" (for the whole population) is rejected with confidence level $1-\alpha$ if and only if $\chi^2_{stat} > \chi^2_{crit}$, where the first term is given by Pearson's formula, while the second one depends only on $m, n$, and $\alpha$, but not on the entries of $A$.

It is immediate to notice that the Pearson's formula is not invariant. More precisely, whenever we multiply all entries of $A$ by a constant $c$, the value $\chi^2_{stat}(A)$ is multiplied by $c$, too, $\chi^2_{stat}(cA) = c\chi^2_{stat}(A)$. Thus, if all rows of $A$ are exactly proportional then $\chi^2_{stat}(cA) = c\chi^2_{stat}(A) = 0$ for any $c$ and any $\alpha$. Otherwise, $\chi^2_{stat}(cA)$ becomes arbitrary large or small, as positive $c$ is increasing or decreasing. Hence, at any fixed significance level $\alpha$, the null hypothesis $H_0$ will be rejected with confidence $1-\alpha$, when $c$ is sufficiently large and not rejected when $c$ is sufficiently small. Yet, obviously, the rows of $cA$ should be proportional or not for all $c$ simultaneously. For this reason, Pearson's test certainly cannot be applied to "physical data", which are obtained by measurements. Indeed, in this case matrix $A$ depends on the unit of measurement. The test can be applied only to categorical data and even then some further limitations are required, which we consider in this paper.

**Keywords:** Pearson chi-square test, difference between two proportions, goodness of fit, contingency tables.

**2020 Mathematics Subject Classification:** 62-00, 62-07XX, 62P10, 62P20, 62P25.

## 1. INTRODUCTION

The chi-square test for proportionality determines whether there is a statistically significant difference between the proportions of a certain outcome in two independent groups.

The original chi-square test, commonly referred to as Pearson's chi-square, originated from Karl Pearson's papers in the late 1800 - early 1900s [11, 12, 13, 14]. It is used both as a "goodness of fit test" - where data are classified along a single dimension - and as a test for contingency tables, where classification occurs across two or more dimensions. A historical overview of the development of the test is provided in [2, 15]. Pearson suggested but did not provide a proof that the test statistic follows the chi-square distribution [2]. The correction regarding the number of degrees of freedom was addressed in Fisher's papers published in 1922 and 1924 [5, 6].

Currently, the chi-square test is extensively utilized across a broad spectrum of research disciplines in analyzing categorical data. Its applications span fields such as social sciences, biomedical research, economics, education, and marketing, where it is employed to assess associations between variables and evaluate the goodness of fit of observed data to expected

Received: 21.07.2025; Accepted: 20.08.2025; Published Online: 26.08.2025
*Corresponding author: Mariya Naumova; mnaumova@business.rutgers.edu
DOI: 10.64700/mmm.75

distributions. This widespread adoption underscores its substantial role in empirical research and statistical inference. According to [1], the chi-square test has seen more use than any other statistical test beside possibly Student's $t$-test. However, despite its widespread application, multiple authors (see for ex., [2, 7, 9, 10]) mention that the chi-square test presents several methodological limitations, such as (not theoretically proven) lower and upper bounds on the sample size, the assumption that the data are derived from independent observations, etc.

In this paper, we demonstrate an extreme sensitivity of the test to the sample sizes and data representation. Specifically, we show that if the given contingency matrix is multiplied by a constant $c$, the test statistic is multiplied by $c$ as well. The results of the test can therefore be easily manipulated (to reject or not reject $H_0$).

**Remark 1.1.** *Similar problems were recently detected for the analysis of variance (ANOVA) and the Tukey-Kramer test. Tukey-Kramer's formula (see (4) in* [8]*) of the critical range for groups of observations $i$ and $j$ depends not only on these two groups but also on all other groups. This contradicts common sense, since these other groups may be not related to groups $i$ and $j$ at all, and immediately leads to some logical contradictions in ANOVA with more than two groups of observations; see* [8] *for definitions and more details.*

## 2. Pearson Chi-Square Test for Homogeneity of Proportions

All three omnibus tests in the Pearson family - goodness of fit, independence, and homogeneity of proportions - share basically the same underlying formula for the test statistic. We will concentrate on the homogeneity test.

Consider a scenario in which $T$ outcomes of multinomial trials are classified according to two distinct criteria, into one of the same $m$ categories, in one of $n$ groups. The outcomes can be presented in a two-way contingency table with $m$ rows and $n$ columns, as shown in Table 1. For example, in [4], a population-based case-control study of prostatic cancer in Alberta

| | Group 1 | Group 2 | $\cdots$ | Group $j$ | $\cdots$ | Group $n$ | Row Total |
|---|---|---|---|---|---|---|---|
| Category 1 | $O_{11}$ | $O_{12}$ | $\cdots$ | $O_{1j}$ | $\cdots$ | $O_{1n}$ | $O_{1\cdot}$ |
| Category 2 | $O_{21}$ | $O_{22}$ | $\cdots$ | $O_{2j}$ | $\cdots$ | $O_{2n}$ | $O_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| Category $i$ | $O_{i1}$ | $O_{i2}$ | $\cdots$ | $O_{ij}$ | $\cdots$ | $O_{in}$ | $O_{i\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| Category $m$ | $O_{m1}$ | $O_{m2}$ | $\cdots$ | $O_{mj}$ | $\cdots$ | $O_{mn}$ | $O_{m\cdot}$ |
| Column Total | $O_{\cdot 1}$ | $O_{\cdot 2}$ | $\cdots$ | $O_{\cdot j}$ | $\cdots$ | $O_{\cdot n}$ | $T$ |

TABLE 1. General structure of an $m \times n$ contingency table. Here, $O_{ij}$ denotes the observed (absolute) frequency in cell $(i, j)$, $O_{i\cdot}$, $O_{\cdot j}$ denote the marginal totals for row $i$ and column $j$ respectively, and $T$ is the grand total.

uses the contingency table related to 376 newly diagnosed prostatic cancer patients and 620 controls, group-matched based on their ethnicity (see Table 2). Let $O_{ij}$ denote the observed (absolute) frequency in the cell corresponding to row $i$ and column $j$, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

| | Cases | Controls | Row Total |
|---|---|---|---|
| British | 200 | 279 | 479 |
| French | 16 | 20 | 36 |
| German | 55 | 93 | 148 |
| Ukrainian | 31 | 79 | 110 |
| Others | 74 | 149 | 223 |
| Column Total | 376 | 620 | 996 |

TABLE 2. The contingency table for a population-based case-control study of prostatic cancer in Alberta, with 376 newly diagnosed prostatic cancer patients and 620 controls of different ethnicities [4].

Let

$$R_i = \sum_{j=1}^{n} O_{ij}, \quad i \in \{1, 2, ..., m\} \qquad \text{(row sum for population } i\text{)},$$

$$C_j = \sum_{i=1}^{m} O_{ij}, \quad j \in \{1, 2, ..., n\} \qquad \text{(column sum for category } j\text{)},$$

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} O_{ij}. \qquad \text{(total sample size)}.$$

The null hypothesis is:

$$H_0 : \pi_{1j} = \pi_{2j} = \cdots = \pi_{mj}, \quad \forall j \in \{1, \ldots, n\},$$

where $\pi_{ij} \in [0, 1]$ denotes the proportion of individuals in group $i$ falling into category $j$, and $\sum_{j=1}^{n} \pi_{ij} = 1$ for each $i$. Under $H_0$, the expected frequency $E_{ij}$ in cell $(i, j)$ is given by:

$$E_{ij} = \frac{R_i \cdot C_j}{N}, \quad \forall i = 1, \ldots, m; \quad j = 1, \ldots, n.$$

Define the test statistic:

$$\chi^2_{stat} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{2.1}$$

Under the test assumptions (see Section 3), the test statistic $\chi^2_{stat}$ approximately follows a chi-square distribution with $(m-1)(n-1)$ degrees of freedom under $H_0$. Let $\alpha \in (0, 1)$ be the significance level. Then the decision rule is:

$$\text{Reject } H_0 \quad \text{if} \quad \chi^2_{stat} > \chi^2_{1-\alpha,(m-1)(n-1)},$$

where $\chi^2_{1-\alpha,(m-1)(n-1)}$ is the $(1-\alpha)$-quantile of the chi-square distribution with $(m-1)(n-1)$ degrees of freedom.

## 3. ASSUMPTIONS OF THE CHI-SQUARE TEST

While the mathematical formulation of the chi-square test statistic is relatively simple - comparing observed frequencies with expected frequencies under a specified null hypothesis- the validity of the test's conclusions rests on a set of non-trivial assumptions. These assumptions

pertain to the way the data are generated, the structure of the contingency table, and the theoretical conditions under which the chi-square distribution serves as a valid approximation to the true sampling distribution of the test statistic. Specifically, the assumptions of the test are as follows [10]:

(i) **Random Sampling:** The data must be drawn via random sampling.

(ii) **Categorical Data:** The variables involved must be categorical. The full set of $m \times n$ cells represents all possible combinations of category levels, and each observation must belong to exactly one cell.

(iii) **Expected Frequency Condition:** The expected frequency in each cell should be sufficiently large to ensure the validity of the chi-square approximation. Specifically, it is recommended that for each cell $(i, j)$,

$$E_{ij} \geq 5^{1}, \quad \forall\, i \in \{1, \ldots, m\},\ j \in \{1, \ldots, n\}.$$

(iv) **Independence of Observations:** Each observation must contribute to exactly one cell, and observations must be independent. That is,

$$\mathrm{Cov}(O_{ij}, O_{kl}) = 0 \quad \text{for all } (i, j) \neq (k, l).$$

(v) **Fixed Margins (if applicable):** For the test of homogeneity or independence, it is often assumed that either the row totals, the column totals, or both are fixed by the sampling design or conditioning.

## 4. Only Invariant Tests Can Verify Proportionality

Obviously, $\chi^2_{stat}(cA) = c\chi^2_{stat}(A)$ for all real $c \geq 0$. The null hypothesis $H_0$ stated in Section 2 is rejected for any fixed significance level $\alpha$, if and only if $\chi^2_{stat}(A) > \chi^2_{crit}(A)$, where $\chi^2_{crit}(A) = \chi^2_{1-\alpha,(m-1)(n-1)}$ is the $(1-\alpha)$-quantile of the chi-square distribution with $(m-1)(n-1)$ degrees of freedom, which depends only on $m, n$, and $\alpha$, but not on the entries of $A$.

**Observation 4.1.** *The following five statements are equivalent:*

*(a) the rows of $A$ are exactly proportional;*
*(b) the columns of $A$ are exactly proportional;*
*(c) $O_{ij} = E_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$;*
*(d) $\chi^2_{stat}(A) = 0$;*
*(e) $\chi^2_{stat}(cA) = 0$ for all real positive $c$.*

Furthermore, if statements (*a* - *e*) fail, then $\chi^2_{stat}(A) > 0$ and for any fixed significance level $\alpha$, the null hypothesis $H_0$ will be rejected for $cA$ with confidence $1 - \alpha$, if $c$ is sufficiently large, and it will not be rejected if $c$ is sufficiently small.

*Proof.* It is enough to notice that in the numerator and denominator of (2.1) are, respectively, a quadratic and linear functions of the entries of $A$. Hence, $\chi^2_{stat}(cA) = c\chi^2_{stat}(A)$. □

Yet, obviously, the result of testing proportionality of the rows of $cA$ should not linearly depend on $c$.

[1] In some sources, $E_{ij} \geq 10$ [2], but neither requirement seems to be mathematically justified.

## 5. Examples

**Example 5.1.** *We consider the following* $2 \times 2$ *contingency table:*

| | *Group 1* | *Group 2* | *Row Total* |
|---|---|---|---|
| *Category A* | 1 | 1 | 2 |
| *Category B* | 1 | 11 | 12 |
| *Column Total* | 2 | 12 | 14 |

*We test the null hypothesis that the proportions of outcomes are the same across the two groups. The expected value for each cell under the null hypothesis are*

$$E_{11} = \frac{2 \cdot 2}{14} = \frac{4}{14} = \frac{2}{7}, \quad E_{12} = \frac{2 \cdot 12}{14} = \frac{24}{14} = \frac{12}{7},$$
$$E_{21} = \frac{12 \cdot 2}{14} = \frac{24}{14} = \frac{12}{7}, \quad E_{22} = \frac{12 \cdot 12}{14} = \frac{144}{14} = \frac{72}{7},$$

*and the chi-square statistic is found to be*

$$\chi^2_{stat} = \frac{(1-\frac{2}{7})^2}{\frac{2}{7}} + \frac{(1-\frac{12}{7})^2}{\frac{12}{7}} + \frac{(1-\frac{12}{7})^2}{\frac{12}{7}} + \frac{(11-\frac{72}{7})^2}{\frac{72}{7}} = \frac{25}{7}\left(\frac{1}{2} + \frac{1}{6} + \frac{1}{72}\right) = \frac{175}{72}.$$

*With* $1$ *degree of freedom, and using the critical value from the chi-square distribution table at significance level* $\alpha = 0.05$*, which is approximately* $3.841$*, we have that*

$$\chi^2_{stat} \approx 2.43 < 3.841.$$

*Thus, we fail to reject the null hypothesis. There is not enough evidence to suggest a significant difference in proportions between the two groups. Let's modify the original table by multiplying each entry by* $2$*:*

| | *Group 1* | *Group 2* | *Row Total* |
|---|---|---|---|
| *Category A* | 2 | 2 | 4 |
| *Category B* | 2 | 22 | 24 |
| *Column Total* | 4 | 24 | 28 |

*Performing the same operations, we find the expected counts under the null hypothesis:*

$$E_{11} = \frac{4 \cdot 4}{28} = \frac{16}{28} = \frac{4}{7}, \quad E_{12} = \frac{4 \cdot 24}{28} = \frac{96}{28} = \frac{24}{7},$$
$$E_{21} = \frac{24 \cdot 4}{28} = \frac{96}{28} = \frac{24}{7}, \quad E_{22} = \frac{24 \cdot 24}{28} = \frac{576}{28} = \frac{144}{7}.$$

*Then the chi-square statistic is computed as*

$$\chi^2_{stat} = \frac{(2-\frac{4}{7})^2}{\frac{4}{7}} + \frac{(2-\frac{24}{7})^2}{\frac{24}{7}} + \frac{(2-\frac{24}{7})^2}{\frac{24}{7}} + \frac{(22-\frac{144}{7})^2}{\frac{144}{7}} = \frac{100}{7}\left(\frac{1}{4} + \frac{1}{12} + \frac{1}{144}\right) = \frac{175}{36}.$$

*Comparing this value to the critical value* $\chi^2_{0.05,1} = 3.841$*, we see that*

$$\chi^2_{stat} \approx 4.86 > 3.841,$$

*and reject the null hypothesis. There is sufficient evidence to suggest that the proportions differ between the two groups.*

Vectors $(1, 1)$ and $(1, 11)$ do not look proportional at all. So it seems a bit strange that we cannot reject proportionality, for the population, with confidence 95%, based on such a sample.

Some researchers (see, for ex. [2, 3, 7]) recommend to apply the Pearson's test only to samples with "large enough" entries. In particular, the expected frequencies in each cell should

generally be at least ten, or in some sources, no less than 5 (both numbers seem to be arbitrarily chosen) [2], or otherwise, the test can result in inaccurate $p$-values and reduced statistical power. Indeed, doubling the values in the first sample, we obtain the second one, with row vectors (2, 2) and (2, 22), which already enables us to reject proportionality with confidence 95%. However, one should note that with very large samples, even trivial associations may yield statistically significant results, potentially leading to overinterpretation.

These restrictions are ambiguous and do not solve the problem in general, since, as we already mentioned, $\chi^2_{stat}(cA) \equiv c\chi^2_{stat}(A)$.

**Example 5.2.** *We illustrate the test for the following $2 \times 2$ contingency table:*

| | *Group A* | *Group B* | *Row Total* |
|---|---|---|---|
| *Category 1* | 22 | 18 | 40 |
| *Category 2* | 18 | 22 | 40 |
| *Column Total* | 40 | 40 | 80 |

*Under the null hypothesis of homogeneity, expected counts for all cells are*

$$E_{ij} = \frac{40 \cdot 40}{80} = 20, i, j \in \{1, 2\}.$$

*The chi-square statistic:*

$$\chi^2_{stat} = \frac{(22-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(22-20)^2}{20} = \frac{4}{5} = 0.8.$$

*Since $0.8 < 3.841$, we do not reject the null hypothesis. There is no significant difference in proportions between the groups.*

*If we modify the original contingency table by multiplying all elements by 1,000 (for example, assuming that the elements of the matrix represent the actual number of cases, while before they represented the number of cases in thousands). This yields the following scaled observed frequencies:*

| | *Group A* | *Group B* | *Row Total* |
|---|---|---|---|
| *Category 1* | 22000 | 18000 | 40000 |
| *Category 2* | 18000 | 22000 | 40000 |
| *Column Total* | 40000 | 40000 | 80000 |

*Under the null hypothesis of homogeneity, all expected counts are*

$$E_{ij} = \frac{40000 \cdot 40000}{80000} = 20000, i, j \in \{1, 2\},$$

*and the chi-square statistic:*

$$\chi^2_{stat} = \frac{(22000-20000)^2}{20000} + \frac{(18000-20000)^2}{20000} + \frac{(18000-20000)^2}{20000} + \frac{(22000-20000)^2}{20000} = 800.$$

*Note that the new $\chi^2_{stat}$ is $1,000$ times greater than the original one. Comparing $\chi^2_{stat}$ to the critical value ($800 > 3.841$), we reject the null hypothesis. There is a significant difference in proportions between the groups.*

*In this example, the null hypothesis $H_0$ claims that all entries of the population are equal. For the first (small) sample, deviations $\pm 2$ can appear with probability $> 5\%$, so we cannot reject $H_0$ with confidence $95\%$. Yet, the second sample is 1,000 times larger, so deviations $\pm 2,000$ have chances $< 5\%$, so we reject $H_0$ with confidence $95\%$. It is known that the standard deviation of entries in $cA$ is proportional to $\sqrt{c}$, rather than to $c$. And still, the value of $c$ must be limited, even for the categorical data.*

**Example 5.3.** *In this case, we work with the following* $3 \times 4$ *contingency matrix of observed frequencies:*

| | *Group 1* | *Group 2* | *Group 3* | *Group 4* | *Row Total* |
|---|---|---|---|---|---|
| *Category A* | 98 | 86 | 79 | 71 | 334 |
| *Category B* | 78 | 82 | 88 | 51 | 299 |
| *Category C* | 75 | 62 | 82 | 77 | 296 |
| *Column Total* | 251 | 230 | 249 | 199 | 929 |

*The matrix of the expected frequencies is given by*

$$E = \begin{bmatrix} 90.3 & 82.7 & 89.5 & 71.6 \\ 80.9 & 73.9 & 80.2 & 64.1 \\ 79.9 & 73.3 & 79.2 & 63.4 \end{bmatrix},$$

*and the chi-square statistic is found to be* $\chi^2_{stat} \approx 11.475$.

*For the given* $3 \times 4$ *contingency matrix, the number of degrees of freedom is* $(3-1)(4-1) = 6$. *So, the p-value is about* $0.07$ *and for* $\alpha = 0.05$ *we don't reject the null hypothesis. In this case, the critical value is* $12.59$. *It is easy to verify that multiplying all entries, for example, by* $2$ *yields the doubled test statistic* ($\chi^2_{stat} \approx 22.95$) *and the much lower p-value* ($\approx 0.0008$).

**Example 5.4.** *We now provide a construction of a matrix for which a proportionality test could be designed. Suppose that we want to analyze the joint probabilistic distribution of two random variables taking* $m$ *and* $n$ *distinct values, say,* $X = (x_1, \dots, x_m)$ *and* $Y = (y_1, \dots, y_n)$. *Our goal is to estimate the two corresponding probabilistic distributions* $(p_1, \dots, p_m)$ *and* $(q_1, \dots, q_n)$ *and decide if* $X$ *and* $Y$ *are independent. A given sample consists of* $T$ *trials, each of which gives us the values of* $X$ *and* $Y$, *simultaneously. Let* $Z_{ij}$ *be the number of trials realizing* $x_i$ *and* $y_j$. *Then,* $z_{ij} = Z_{ij}/T$ *are the corresponding observed frequencies. Set* $p_i = \sum_{j=1}^{n} z_{ij},\ q_j = \sum_{i=1}^{m} z_{ij}$, *and treat them as the observed frequencies for* $X$ *and* $Y$, *respectively. By the above definitions, we have*

$$\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ij} = 1.$$

*We define the expected frequencies by setting* $z^e_{ij} = p_i q_j$.

Although the last example is also a case of verifying proportionality, it is special. Since the sum of all entries equals 1, multiplication of all observed frequencies $z_{ij}$ by a constant $c \neq 1$ is impossible. More generally, the chi-square test will work when the sum of all entries must be constant, or almost constant.

## 6. Conclusions

The Pearson chi-square test is one of the most widely used tools for analyzing categorical data, particularly in contingency tables. As demonstrated through both theoretical formulation and numerical examples, even under its common assumptions, including sufficiently large expected cell frequencies and the independence of observations, the test is extremely sensitive to the input data, and its inaccurate application can lead to misleading conclusions. The interpretation of the results must therefore be approached with caution, as the heuristic nature of the test may introduce critical biases. Consequently, findings should be considered exploratory rather than definitive.

The Pearson's chi-square test cannot be applied for verifying proportionality for general contingency matrices. We show that the chi-square statistic increases linearly when the entire contingency table is multiplied by a constant factor, reflecting the proportional scaling of observed and expected frequencies. However, the test is applicable in cases where the total sum

of entries is a constant, or almost constant, for example, for testing the independence of two probabilistic distributions, given a sample of pairwise frequencies.

Pearson's test certainly cannot be applied to "physical data", which are obtained by measurements. Indeed, in this case, matrix $A$ depends on the units of measurement. For example, $A$ is replaced by $3,600A$, when we switch from hours to seconds. The test can be applied only to categorical data and even then further limitations are required; some of them are considered in this paper.

One could modify formula (2.1) making it invariant with respect to multiplication of $A$ by a constant, for example, by squaring all denominators in (2.1) or by dividing all entries of $A$ by their sum or maximum. However, then the distribution $\chi^2_{crit}$ should be modified accordingly. Both above modifications deserve a separate consideration.

## Acknowledgments

This research is performed within the framework of the HSE University Basic Research Program.

## References

[1] D. Baird: *The Fisher/Pearson Chi-squared controversy: a turning point for inductive inference*, Br. J. Philos. Sci., **34** (2) (1983), 105–118.
[2] W. G. Cochran: *The $\chi^2$ test of goodness of fit*, Ann. Math. Stat., **23** (3) (1952), 315–345.
[3] P. E. Greenwood, M. S. Nikulin: *A guide to chi-squared testing*, John Wiley & Sons., New York (1996).
[4] S. M. Fincham, G. B. Hill, J. Hanson and C. Wijayasinghe: *Epidemiology of prostatic cancer: a case-control study*, The Prostate, **17** (3) (1990), 189–206.
[5] R. A. Fisher: *On the interpretation of $\chi^2$ from contingency tables, and the calculation of P*, J. Roy. Statist. Soc., **85** (1) (1922), 87–94.
[6] R. A. Fisher: *The conditions under which $\chi^2$ measures the discrepancy between observation and hypothesis*, J. Roy. Statist. Soc., **87** (3) (1924), 442–450.
[7] T. M. Franke, T. Ho and C. A. Christie: *The chi-square test: Often used and more often misinterpreted*, Am. J. Eval., **33** (3) (2012), 448–458.
[8] V. Gurvich, M. Naumova: *Logical contradictions in the one-way ANOVA and Tukey-Kramer multiple comparisons tests with more than two groups of observations*, Symmetry, **13** (2021), Article ID: 1387.
[9] A. E. Kossovsky: *On the mistaken use of the chi-square test in Benford's law*, Stats., **4** (2) (2021), 419–453.
[10] M. L. McHugh: *The chi-square test of independence*, Biochem. Med., **23** (2) (2013), 143–149.
[11] K. Pearson: *Contributions to the mathematical theory of evolution*, Philos. Trans. R. Soc. A, **185** (1894), 71–110.
[12] K. Pearson: *X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material*, Philos. Trans. R. Soc. A, **186** (1895), 343–414.
[13] K. Pearson: *Mathematical contributions to the theory of evolution — On a form of spurious correlation which may arise when indices are used in the measurement of organs*, Proc. R. Soc. Lond., **60** (359-367) (1897), 489–498.
[14] K. Pearson: *X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, London Edinburgh Philos. Mag. & J. Sci., **50** (302) (1900), 157–175.
[15] R. L. Plackett: *Karl Pearson and the chi-squared test*, Int. Stat. Rev., **51** (1) (1983), 59–72.

Vladimir Gurvich
National Research University
Higher School of Economics
Moscow, Russia

Rutgers University
RUTCOR
Piscataway, NJ, USA
*Email address*: vgurvich@hse.ru, vladimir.gurvich@gmail.com

MARIYA NAUMOVA
RUTGERS UNIVERSITY
RUTGERS BUSINESS SCHOOL
PISCATAWAY, NJ, USA
*Email address*: mnaumova@business.rutgers.edu