European Mathematical Society Magazine

ISSN 2747-7894 / Issue 124 / June 2022

EMS Magazine

Preda Mihăilescu and Daniel Muzzulini
A Diophantine equation concerning epimoric ratios

Christos P. Kitsos
Sir David Cox (1924–2022)

Scott Mandelbrote
The Newton Project

Nuno Crato
Math curriculum matters

EUROPEAN MATHEMATICAL SOCIETY

3 A message from the president
Volker Mehrmann

3 Brief words from the editor-in-chief
Fernando Pestana da Costa

4 Statistical tools for anomaly detection as a part of predictive maintenance in the mining industry
Agnieszka Wyłomańska

16 A Diophantine equation concerning *epimoric* ratios
Preda Mihăilescu and Daniel Muzzulini

23 The Newton Project
Scott Mandelbrote

27 Sir David Cox: A wise and noble statistician (1924–2022)
Christos P. Kitsos

33 Mathematics at the Institute of Science and Technology Austria (ISTA)
Florian Schlederer

38 Live and remote: Nairobi workshops in algebraic geometry
Chiara Bonadiman, Diletta Martinelli, Jared Ongaro and Balázs Szendrői

42 Calculus: A new approach for schools that starts with simple algebra
R. Michael Range

49 Math curriculum matters: Statistical evidence and the Portuguese experience
Nuno Crato

57 ICMI column
Núria Planas

59 Book reviews

65 Letter of Ukrainian mathematicians
Yuriy Drozd, Rostyslav Hryniv, Olena Karlova and Sergiy Maksymenko

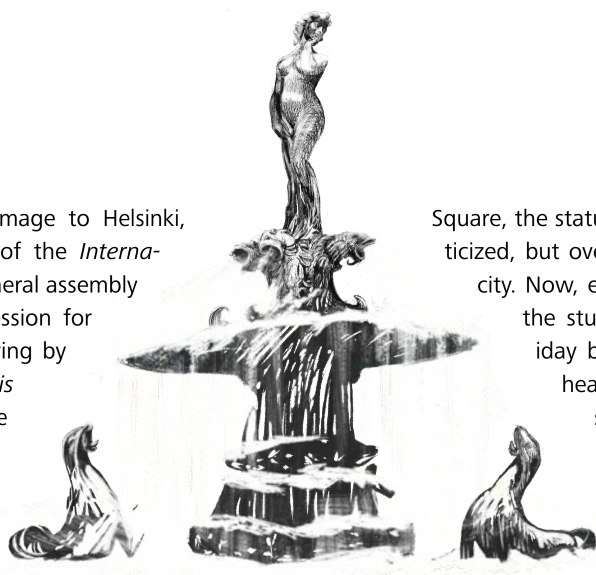
European Mathematical Society Magazine

Published by EMS Press, an imprint of the
European Mathematical Society – EMS – Publishing House GmbH
Technische Universität Berlin, Institut für Mathematik,
Straße des 17. Juni 136, 10623 Berlin, Germany

https://ems.press

Typesetting: WisSat Publishing + Consulting GmbH, Fürstenwalde, Germany
Printing: Beltz Bad Langensalza GmbH, Bad Langensalza, Germany

ISSN (print) 2747-7894; ISSN (online) 2747-7908



The cover illustration is a homage to Helsinki, which will be the 2022 site of the *International Mathematical Union* general assembly meeting and of the award session for the Fields medals. It is a drawing by António B. Araújo of the *Havis Amanda*, a work of sculptor Ville Vallgren (1855–1940), and one of Helsinki's best loved public statues. Upon its 1908 unveilling in Helsinki's Market Square, the statue was considered scandalous and much criticized, but over time it became a beloved symbol of the city. Now, every 1st of May the mermaid graduates as the students of Helsinki celebrate the Vappu holiday by placing the graduation white cap on its head. "Mermaid" was the original name of the statue. The present name, *Havis Amanda*, probably comes from the Swedish *Havets Amanda* ("Amanda of the Sea") and is fitting to the theme of the original commission: "Helsinki rising from the sea".

A message from the president

Dear EMS members,

In my editorial to the last issue, a lot of important developments were promised for 2022, but then the Russian invasion into Ukraine, the massive killing of civilians, the bombing of Ukrainian cities and the attempt of Russia to overtake Ukraine have completely shattered a lot of the values and beliefs of a Europe in peaceful cooperation. After strong petitions of the EMS and others, the ICM was moved to a virtual format and the IMU general assembly was taken away from St. Petersburg. But the war is going on, people are dying, and a large part of Ukraine is destroyed. Even though there are many voices in Russia despising the war, the government propaganda and the oppression of free speech and protest have not yet led to an end to the war. In the last months, also the EMS has been flooded with calls for boycotts of all kinds, see also the letter on page 65, and the EMS executive committee has discussed different ways on how to react. We have decided to stop all cooperation with Russian governmental institutions, but to keep relations with Russian colleagues and mathematical societies.

The executive committee of the EMS discussed also to again postpone the celebration of the 30th anniversary of the EMS, but in the end we decided to go on with the celebration and we had a wonderful event in Edinburgh which included memorial talks for Sir Michael Francis Atiyah and a session on the mathematics of the pandemic. We will also proceed with the EMS council June 25 and 26 in Bled, Slovenia, and we look forward to seeing many representatives there.

We truly hope that this war comes to an end soon and that the rest of Europe supports Ukraine to overcome this terrible hardship.

Volker Mehrmann
President of the EMS

Brief words from the editor-in-chief

Dear readers of the EMS Magazine,

On the 24th of February Russian troops invaded Ukraine and started a war of aggression that, at the time of writing, still rages on with no end in sight. More than three quarters of a century after the end of World War II in Europe, and about a quarter of a century after the siege and bombing of Sarajevo by Serbian forces, or NATO aerial bombing campaign of Serbia and Montenegro, we witness again in horror the bombing of European cities and an unlawful war between European nations.

In these dark times it is crucial that, in parallel to lending support to the Ukrainian people, we do not lose sight of the fact that there will be an afterwar (hopefully sooner rather than later) and all the countries that now oppose each other on the battlefield or on the diplomatic or economic fronts will still be there and will still need to talk with each other if a secure and long-term peaceful future for all is to be achieved. I believe this healing and coming together will be easier to achieve in the future if we, as individuals, do not drift too much apart in the present.

In this context, the EMS Magazine will continue its policy of not discriminating against any individuals, be they authors, editors, readers, or any other class of collaborator, based on any general characteristics such as ethnicity or nationality.

Fernando Pestana da Costa
Editor-in-chief

Statistical tools for anomaly detection as a part of predictive maintenance in the mining industry

Agnieszka Wyłomańska

We present new achievements in the area of anomaly detection related to predictive maintenance in the mining industry. The main focus is on the problem of local damage detection based on vibration signals analysis. The vibration signals acquired from machines usually have a complex spectral structure. As the signal of interest (SOI) is weak (especially at an early stage of damage) and covers some frequency range, it must be extracted from raw observations. Up to now, most the techniques assumed the presence of Gaussian noise. However, there are cases in which the non-informative part of the signal (considered as the noise) is non-Gaussian due to random disturbances or to the nature of the process executed by the machine. In such cases, the problem can be formulated as the extraction of the SOI from the non-Gaussian noise. Recently, the importance of this problem has been recognised by several authors, and some new ideas have been developed. We present here a comparison of the new techniques for benchmark signals. Our analysis will cover classical approaches and recently introduced algorithms based on the stochastic analysis of the vibration signals with non-Gaussian distribution.

1 Introduction

Vibration-based condition monitoring is commonly used for the maintenance of mechanical systems [35, 36]. The main focus is usually set on gearboxes and bearings, as these elements appear in most transmission systems and their failure is the most frequent reason for a machine breakdown; for recent reviews, see [16, 42].

From the mathematical/statistical point of view, the task is defined as the detection of the periodic impulsive behaviour in the vibration signal. The most popular approach is the envelope analysis (and its various modifications) and detection of fault frequencies in the spectrum of the envelope for a given signal; see e.g. [5, 6, 37, 40]. The filtration of a raw signal is used to select its informative part and avoid other spectral content not related to the local damage.

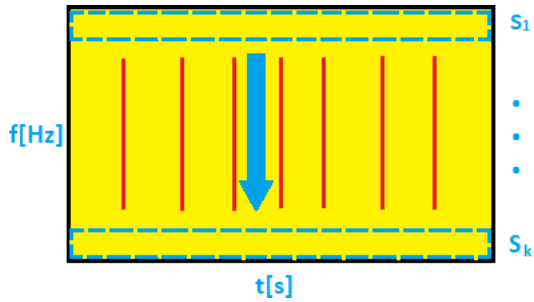
The most popular approach is based on spectral kurtosis as an informative frequency band (IFB) selector (filter characteristic). The kurtosis value is calculated for sub-signal at some narrow frequency band. As kurtosis is sensitive to outliers [41], one can select impulsive content at a given narrow frequency band and filter out other components. Kurtosis is the most intuitive statistic commonly used for machine diagnostics [36]. It has plenty of variations and extensions, e.g. the kurtogram [1] which is a coloured map in which the depth of the colour values is proportional to the kurtosis value. Moreover, other statistics are also used in such context [33].

In the literature, one can also find methods which are based on the cyclostationary approach. Recall that cyclostationary signals are considered when some of their characteristics are periodic in time. The most common characteristic used in this context is the autocovariance function, in which case we consider the cyclostationarity of the second order. The methods for the analysis of cyclostationary signals are dedicated to the cyclic behaviour identification. In the classical approach of cyclostationary-based techniques, the Gaussian distribution of the signal is usually assumed [2, 6, 7, 11, 29, 31].

One can also find other approaches for the informative frequency band selection based on artificial intelligence methods [26, 38, 50]. However, there is still a need for new approaches that would allow us to consistently handle restrictions linked to the amount of available data, specific type of noise, work specifications of the tested machine, etc.

Unfortunately, most of the standard statistical indicators used for local damage detection might be not appropriate in the presence of non-Gaussian noise. In the real environment, we observe signals with cyclic and impulsive behaviour. One of the examples is the crushing machine [47] used in mines. During the operation of such a machine (the crushing process), apart from the background noise (which is often assumed to be Gaussian white noise) in the vibration signal, large observations appear due to the nature of the machine's work. Moreover, in the case of local damage, the additional cyclic impulses are hidden in the signal. In this case, the detection of local damage is very difficult. The non-Gaussian noise could be also linked to other technological processes of the working machine and may correspond to milling, sieving, cutting, compressing, etc. In the literature, algorithms for signals with non-Gaussian distribution have been proposed [10, 19, 21, 27, 28, 48, 49]. A new definition of the cyclostationary non-Gaussian signal was introduced in [25].

We demonstrate here some recently proposed algorithms for local damage detection that take into consideration the possible non-Gaussian behaviour of the signal and utilise the advanced statistics that are robust for large observations not related to failure. We present two approaches. The first is related to new techniques that can replace the classical measure of impulsiveness, i.e. kurtosis used as the selector for the IFB. In the second approach, the local damage is detected using the cyclostationary analysis dedicated to non-Gaussian distributed signals. All these techniques were recently published and used for simulated and real signals [12–14, 25, 32]. Methods dedicated to non-Gaussian vibration signals were also developed during the OPMO project (EiT Raw Materials), which was carried out at the Wrocław University of Science and Technology (2019–2021), among others with a global mining industry leader KGHM. Those results were also the basis for two research projects currently implemented at the Faculty of Pure and Applied Mathematics at the Wrocław University of Science and Technology (the first in cooperation with the AMC Tech company and the second with the Tsinghua University, China).



(a)

$$
\begin{aligned}
S_1 &= \{|\mathrm{STFT}(t_1,f_1)|, |\mathrm{STFT}(t_2,f_1)|, \dots, |\mathrm{STFT}(t_l,f_1)|\} \\
S_2 &= \{|\mathrm{STFT}(t_1,f_2)|, |\mathrm{STFT}(t_2,f_2)|, \dots, |\mathrm{STFT}(t_l,f_2)|\} \\
&\vdots \\
S_k &= \{|\mathrm{STFT}(t_1,f_k)|, |\mathrm{STFT}(t_2,f_k)|, \dots, |\mathrm{STFT}(t_l,f_k)|\}
\end{aligned}
$$

(b)

Figure 1. (a) Simplified representation of the spectrogram, (b) matrix representation of the spectrogram [32]

2 Informative frequency band selectors for non-Gaussian signals

In the problem of IFB selection, the fist step is the decomposition of the raw signal into a set of narrow-band sub-signals using a time-frequency representation to obtain several dozen time series. To perform signal decomposition, one may use various techniques (wavelets, Wigner, EMD, etc.); see e.g. [9]. In this research, the Short-Time Fourier Transform (STFT) is used. The transform is defined as

$$\mathrm{STFT}(t,f) = \sum_{k=1}^{N} x_k w(t-k) e^{\frac{-2j\pi fk}{N}},$$

where $w(t-k)$ is the shift window, $x_1, x_2, \dots, x_N$ is the input signal, N is its length, $t \in T$ is the time point, and $f \in F$ is a frequency; see [3] for more details. Interpretation of the STFT is intuitive: the squared envelope of the STFT (spectrogram) describes the energy flow in time for some narrow frequency band, i.e. sub-signal. The simplified representation of the spectrogram is presented in Figure 1 (a).

The next step is the application of some statistics (called selectors) to time series obtained from the spectrogram $S_1, S_2, \dots, S_k$ (Figure 1 (b)) to identify if a given sub-signal fulfils expectations regarding information about faults and select similar sub-signals from the whole frequency range. Distribution of selectors along frequencies, after normalisation, may constitute the filter characteristic (frequency response). One should expect that IFB for the healthy machine will contain stationary noise. In case of damage, the energy flow in some frequency bins will reveal a non-stationary character, and distribution of such sub-signals will be far from Gaussian; see [34, 46, 47] for more details. In Figure 1, we present the general idea of preparing the signal to calculate the selectors for IFB using the time-frequency representation of the signal.

In the following subsections, we present three statistics that are considered in the literature as the IFB selectors.

2.1 Kurtosis

In probability theory and statistics, kurtosis is considered as the tail measure of the probability distribution [43]. For a given random variable X with finite fourth moment, its kurtosis is defined as

$$K = \frac{\mu_4}{\sigma^4},$$

where $\mu_4 = E(X-\mu)^4$ is the fourth central moment of X, μ is its expectation of X, and σ^2 is the variance of X. For a Gaussian random variable, the kurtosis is always equal to 3; sometimes the term excess kurtosis is used in reference to $K-3$. The empirical kurtosis is based on a scaled version of the fourth empirical moment of the data. Given a signal $x_1, x_2, \dots, x_N$, the empirical kurtosis is a statistic defined as

$$\hat{K} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{N\hat{\sigma}^4},$$

where $\hat{\sigma}^2$ is the empirical sample variance and $\bar{x}$ is the empirical mean; see [20] for details. In our case, the statistic $\hat{K}$ is calculated for the time series $S_1, S_2, \dots, S_k$ from time-frequency representation of the raw signal, see Figure 1 (b), and thus in this case it is called the spectral kurtosis [13].

2.2 *Alpha selector*

Let us first recall the class of α-stable distributions. It is considered as a natural extension (via the generalised limit theorem) of the classical Gaussian distribution; see [15, 39]. It is very useful in data modelling with impulsive behaviour (also for damage detection), since in general it contains heavy-tailed (power-law) distributions. In the problem under consideration, we do not use the α-stable distribution to describe the signal, but to analyse one of the parameters of this distribution (the stability index α) as the measure of impulsiveness, and consider its estimator as the selector for IFB.

The random variable X has a stable distribution with stability index $\alpha \in (0,2]$, scale parameter $\sigma > 0$, skewness parameter $\beta \in [-1,1]$, and shift parameter $\mu \in \mathbb{R}$, if its characteristic function $\phi_X(\theta) = \mathbb{E}[\exp\{i\theta X\}]$ is given by

$$\phi_X(\theta) = \begin{cases} e^{-\sigma^\alpha|\theta|^\alpha\{1-i\beta\operatorname{sign}(\theta)\tan(\pi\alpha/2)\}+i\mu\theta}, & \alpha \neq 1, \\ e^{-\sigma|\theta|\{1+i\beta\operatorname{sign}(\theta)2/\pi\log(|\theta|\}+i\mu\theta}, & \alpha = 1. \end{cases}$$

For $\beta = \mu = 0$, X has a symmetric stable distribution, and for $\alpha = 2$, the α-stable distribution reduces to the Gaussian.

In the literature, one can find different α estimation methods; see e.g. [4, 24]. We use the classical McCulloch method [30]. In further analysis, the statistic $1 - \hat{\alpha}$ obtained using this approach applied to the sub-signals $S_1, S_2, \ldots, S_k$ (Figure 1 (b)) is called the Alpha selector; see [13].

2.3 *Conditional variance-based selector*

The selector called conditional variance-based (CVB) was introduced in [13]. Below, we briefly describe this approach. Let us assume that X is a Gaussian random variable with mean μ and standard deviation σ. Let $\Phi_{\mu,\sigma}(\,\cdot\,)$ denote the distribution function (cdf) of X. For any level $0 < q < 0.5$, we define the left, right, and middle quantile partitioning of X by

$$\begin{aligned} L_q &\coloneqq (-\infty, \Phi^{-1}_{\mu,\sigma}(q)], \\ R_q &\coloneqq [\Phi^{-1}_{\mu,\sigma}(1-q), \infty), \\ M_q &\coloneqq (\Phi^{-1}_{\mu,\sigma}(q), \Phi^{-1}_{\mu,\sigma}(1-q)), \end{aligned}$$

where $\Phi^{-1}_{\mu,\sigma}$ is the inverse of $\Phi_{\mu,\sigma}$, i.e. $\Phi^{-1}_{\mu,\sigma}(d)$ denotes the d-quantile of X. Under the normality assumption and the partitioning ratio close to 20/60/20, i.e. for $q \approx 0.2$, we get that

$$\sigma^2_{L_q} = \sigma^2_{M_q} = \sigma^2_{R_q}, \tag{1}$$

where $\sigma^2_A \coloneqq \operatorname{Var}(X \mid X \in A)$ is the conditional variance of X on set A; see [17] for details. Moreover, the 20/60/20 ratio is the unique quantile (three set) partitioning satisfying property (1).

This property can be described as follows: if we split the large normal random sample into three sets, one corresponding to the worst (smallest) 20 % of outcomes, one corresponding to the middle 60 % of outcomes, and one corresponding to the best (largest) 20 % of outcomes, then the conditional variance on appropriate subsets is approximately the same.

As noted in [17], condition (1) creates a dispersion balance for the conditional populations. This might be linked to the statistical phenomenon commonly referred to as the 20/60/20 rule.

Since the ratio 20/60/20 is the unique ratio for which condition (1) is satisfied, it can be used to construct a goodness-of-fit test statistic. In other words, by comparing conditional variances with the conditional central variance, one can verify whether the sample comes from a Gaussian distribution. As the conditional tail variance might be seen as a measure of a tail heaviness, we can also use property (1) to benchmark any distribution tails with respect to normal tails without making any explicit assumptions about the distribution of X.

The statistic used in [18] for Gaussian distribution testing is defined as

$$\hat{C} \coloneqq \frac{1}{\rho}\left(\frac{\hat{\sigma}^2_{L_q} - \hat{\sigma}^2_{M_q}}{\hat{\sigma}^2} + \frac{\hat{\sigma}^2_{R_q} - \hat{\sigma}^2_{M_q}}{\hat{\sigma}^2}\right)\sqrt{N}, \tag{2}$$

where $q = 0.2$, $\rho \in \mathbb{R}$ is a normalisation constant, $\hat{\sigma}^2$ is the sample variance, and $\hat{\sigma}^2_A$ is set A conditional sample variance. Assuming that the sample is independent and identically distributed, the asymptotic distribution of $\hat{C}$ is standard normal. Moreover, if the sample under consideration comes from a (symmetric) heavy-tailed distribution, the values of the statistic C should be positive due to high values of conditional tail variances on sets L_q and R_q. Consequently, $\hat{C}$ could be considered as the measure of tail fatness, i.e. the bigger the value of $\hat{C}$, the fatter the tails. The value of $\hat{C}$ for real signals is based on the empirical conditional variance. See [13] for more details.

The statistic given in (2) can be extended, and a different number of partitioning sets could be considered. In [13], it was proposed to partition into seven quantile conditioning subsets and use the (unique) ratio guaranteeing conditional variance equality. The statistic used as the CVB selector is defined as

$$\hat{C}_1 \coloneqq \left(\frac{\hat{\sigma}^2_{A_3} - \hat{\sigma}^2_{A_4}}{\hat{\sigma}} + \frac{\hat{\sigma}^2_{A_5} - \hat{\sigma}^2_{A_4}}{\hat{\sigma}}\right)^2\sqrt{N},$$

where $\hat{\sigma}^2_A$ is the conditional sample variance on A and the subsets A_i come from the partitioning of the signal into seven appropriate subsets. In this approach, C_1 was used to measure the impact of the non-extreme (trimmed) tail variance on the central part of the distribution without bench-marking the model using Gaussian distribution. Similarly to what we saw above for kurtosis and the Alpha selector, $\hat{C}_1$ applied to the time series $S_1, S_2, \ldots, S_k$ (Figure 1 (b)) is used as the selector for IFB.

3 Comparative study of IFB selection – simulated signal analysis

In order to verify the efficiency of the presented selectors, we simulated four different types of signals: s_1, s_2, s_3 and s_4. The first signal s_1 is the Gaussian white noise, which corresponds to the bearing vibration in a healthy condition; see Figure 6 (a). In such a case, we expect any of the informative frequency band selectors to respond significantly. The signal's frequency is 25000 Hz and its length is 1 second. Signal s_2 corresponds to the locally damaged bearing vibration and is defined as

$$s_2 = \mathrm{ACI} \cdot \mathrm{gauspuls}(t_0, \mathrm{fc}_0, \mathrm{bw}_0) + s_1,$$

where $\mathrm{gauspuls}(t, \mathrm{fc}, \mathrm{bw})$ is a unity-amplitude Gaussian radio-frequency (RF) pulse at the times indicated in array t, with a centre frequency fc in hertz and a fractional bandwidth bw. ACI is the amplitude of the cyclic impulses (ACI = 3), fc_0 is set to 2500 Hz, and the frequency modulation of cyclic impulses is equal to 30 Hz. The simulated signal s_2 is presented in Figure 6 (b).

The signal s_3 imitates the bearing operation of the loaded machine in a healthy condition; it is defined as

$$s_3 = \mathrm{ANCI} \cdot \mathrm{gauspuls}(t_1, \mathrm{fc}_1, \mathrm{bw}_1),$$

where ANCI = 30 is the amplitude of the non-cyclic impulses and t_1 is the location of the non-cyclic impulses, with the uniform distribution. The simulated signal s_3 is depicted in Figure 6 (c).

The last of the signals is a mixture of the previous ones, namely, it is defined as $s_4 = s_2 + s_3$. It imitates the bearing vibrations in the case of a loaded machine operating in the unhealthy condition of bearing; see Figure 6 (d). In Figure 7, we present the spectrograms for the simulated signals presented in Figure 6. Panel (a) illustrates the spectrogram of the signal s_1. The signal is the Gaussian white noise, so it is neither impulsive nor periodic. In panel (b), we demonstrate the spectrogram for signal s_2. For this cyclic impulsive signal, we expect the techniques to point out the informative frequency band between 2 and 3 kHz (the centre frequency is set to 2500 Hz). In panel (c), we show the spectrogram for signal s_3. For this non-cyclic impulsive signal, any IFB is expected. Finally, in panel (d), the spectrogram for signal s_4 is demonstrated. This is the most complicated case. The signal contains large non-cyclic and cyclic impulses. Thus, we expect to find information about the cyclic impulses with possibly suppressed information about non-cyclic ones.

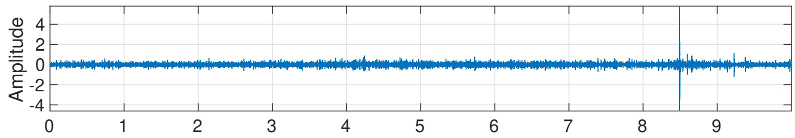
In Figures 8–10, we present the considered selectors for the four simulated signals s_1, s_2, s_3, s_4. As one can see, if the signal contains only the Gaussian noise (s_1), all considered techniques, kurtosis, Alpha selectors and CVB, have relatively small amplitudes and do not indicate the informative frequency band, as expected. There is no frequency band which significantly stands out from the others. For the signal s_2, all techniques work well, but the results for the CVB selector seem to be the most unequivocal. The value of the CVB selector in the range of the IFB is significantly higher than for other frequency bins. If the cyclic impulses have higher amplitudes, then their variance will increase and the value of the CVB will increase as well. For the non-cyclic impulsive signal s_3, all techniques properly select the frequency band where the non-cyclic impulses appear. Note that none of the methods take into consideration the cyclic behaviour of the signal, only its impulsiveness. For the most complicated case, i.e. for the signal with a large non-cyclic to cyclic impulse amplitude ratio (s_4), surprisingly only the CVB selector correctly identified the frequency band corresponding to the cyclic impulses, based on the distribution of their amplitudes. The kurtosis and Alpha selector indicate both cyclic and non-cyclic impulses frequency ranges. For more details and comparison with other selectors, see [14].

4 Real signal analysis

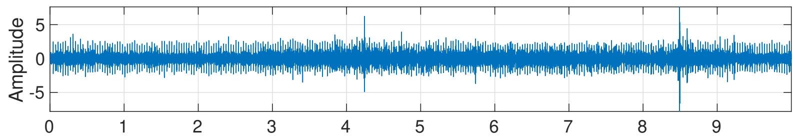
In this part, we present an application of the statistics we are considering to a real signal. The effectiveness of the different methods is verified on the data from the bearing of a crushing machine; see Figure 2.

However, due to the lack of local faults in the considered vibration data, we introduce an artificial component related to local damage. A similar approach was performed in [13, 44, 47]. The signal is presented in Figure 4. The length of the signal is 6 seconds and the sampling frequency is 25 kHz. The local fault was added with frequency modulation equal to 30 kHz and carrier frequency equal to 2.5 kHz (2–3 kHz). The spectrogram of the data is presented in Figure 11 (a). As we can see, the data reveals high-energy wide-band impulses around 0.25, 4.5 and 5.25 seconds, which correspond to falling rocks. The real vibration signal contains various components with a complex structure, and the component related to the damage is almost imperceptible above the noise. We apply the informative frequency band selectors to the real signal

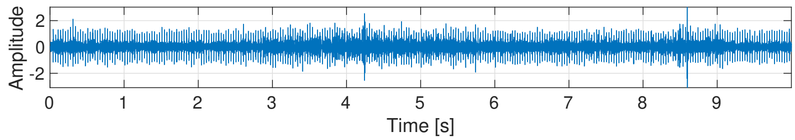
Figure 2. Crushing machine in a copper ore mine [45]



(a) Kurtosis selector



(b) Alpha selector



(c) CVB selector

Figure 3. Results of the copper ore crusher's signal filtration performed by three different selectors [13]

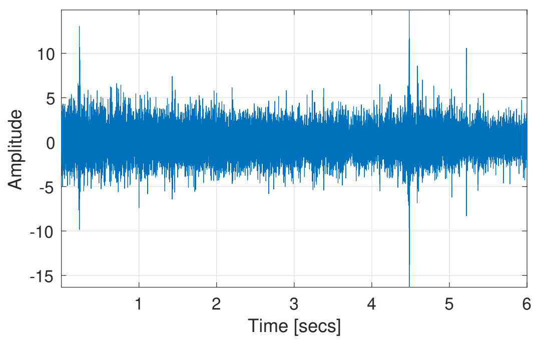


Figure 4. Analysed real signal [14]

with its added fault. The results are presented in Figure 11 (b)–(d). The amplitude of the non-cyclic impulses is smaller than in the simulated data, but one can see that the results of the kurtosis fail. The Alpha selector properly indicates the IFB, but with much less selectivity than the CVB selector.

The signal filtered using the analysed selectors is presented in Figure 3. Filtering driven by the kurtosis selector (panel (a)) provides a single impulsive component that is a non-informative signal. In contrast, the Alpha selector (panel (b)) allows extraction of cyclic impulsive components with 2 random impulses. Finally, results of the CVB selector-based filtering (panel (c)) show a similar effect to the Alpha selector with slightly smaller random impulses. To summarise, Figure 3 shows that the filtration with the kurtosis selector fails, while the results of Alpha and CVB-based methods are acceptable and similar; however, we note that the shape of the CVB selector is better.

5 Cyclostationary analysis for non-Gaussian signals

In this part, we discuss new research related to the cyclostationary analysis for non-Gaussian signals. More precisely, we propose to apply dependency measures expressed by the known correlation coefficients to detect the cyclic behaviour on the time-frequency map. The idea is as follows. First, we represent the signal as the time-frequency map (Figure 1 (a)), then we consider the corresponding sub-signals $S_1, S_2, \dots, S_k$ as the separate time series (Figure 1 (b)), and finally, we apply the dependency measure $m(\cdot,\cdot)$ (see examples below) to the time series extracted from the time-frequency representation of the signal. The general idea of this approach is illustrated in Figure 5.

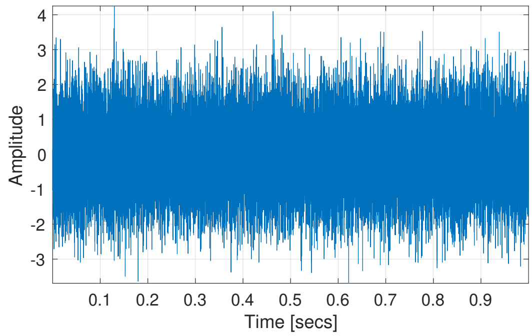
5.1 *Pearson correlation map*

Let $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ be a bi-dimensional sample of a random vector (X, Y), where N is the sample length. The Pearson correlation of (X, Y) is defined as follows [8]:

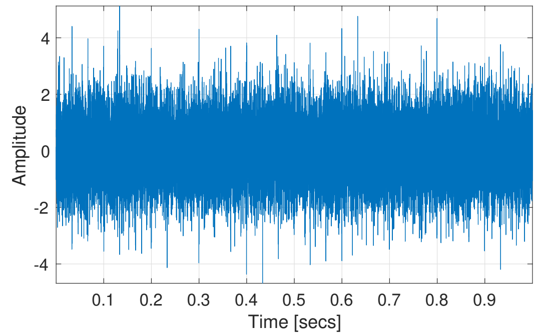
$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y},$$

$$\boxed{\begin{array}{l} m(s_1,s_1), m(s_1,s_2), \dots, m(s_1,s_l) \\ m(s_2,s_1), m(s_2,s_2), \dots, m(s_2,s_l) \\ \vdots \\ m(s_k,s_1), m(s_k,s_2), \dots, m(s_k,s_l) \end{array}}$$

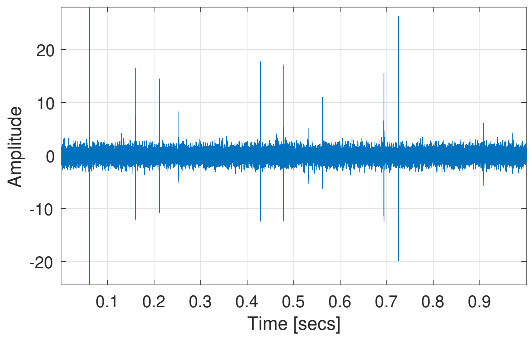
Figure 5. Dependency map structure for spectrogram sub-signals and a given dependency measure $m(\cdot,\cdot)$ [32]



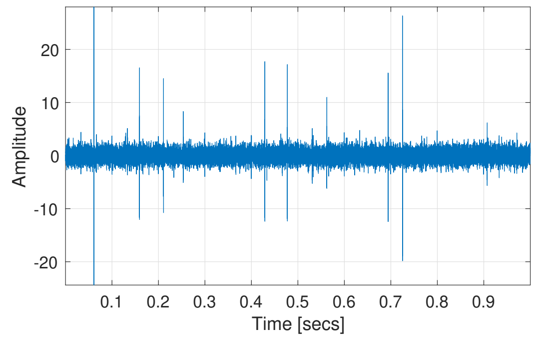
(a) s_1 – Gaussian noise



(b) s_2 – Gaussian noise with cyclic impulses

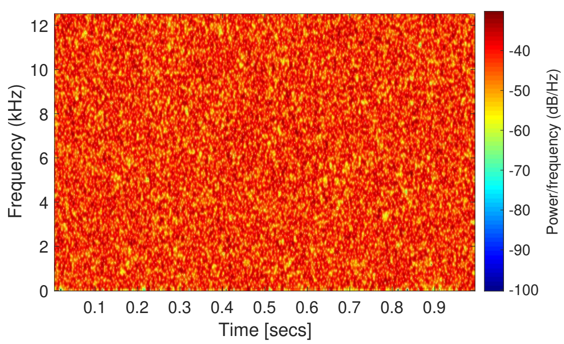


(c) s_3 – non-Gaussian noise

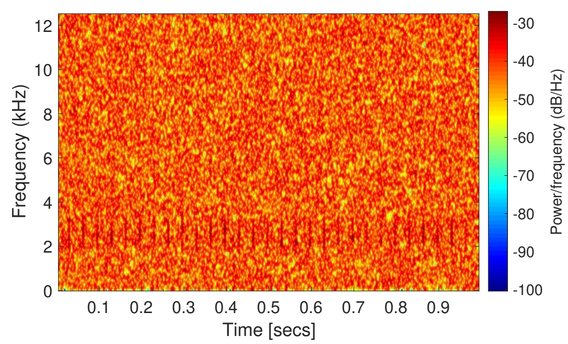


(d) s_4 – non-Gaussian noise with cyclic impulses

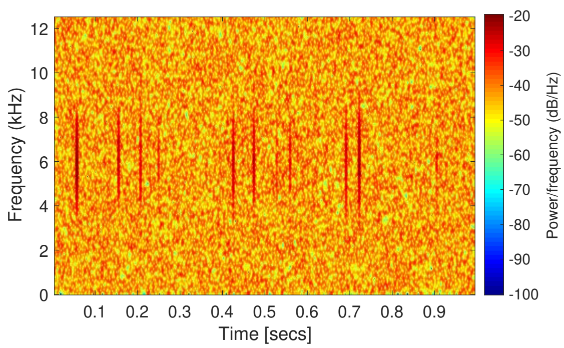
Figure 6. Simulations of the considered signals [14]



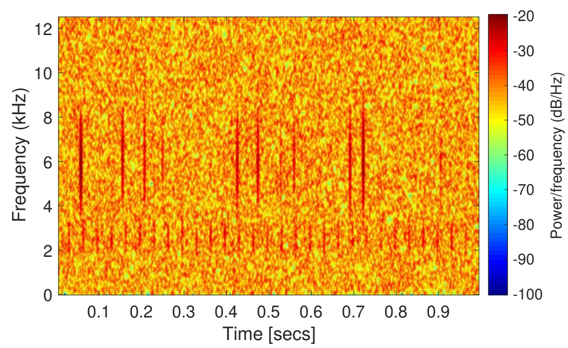
(a) s_1 – Gaussian noise



(b) s_2 – Gaussian noise with cyclic impulses

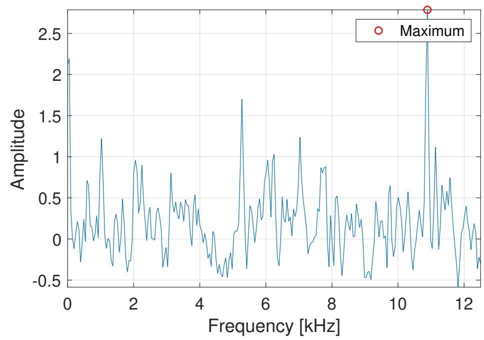


(c) s_3 – non-Gaussian noise

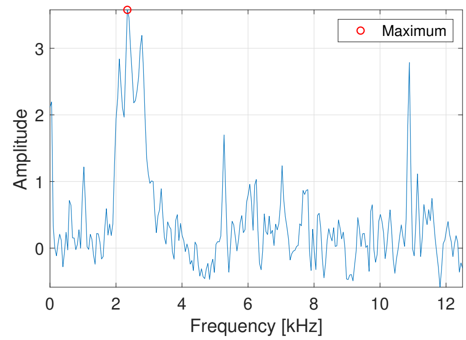


(d) s_4 – non-Gaussian noise with cyclic impulses

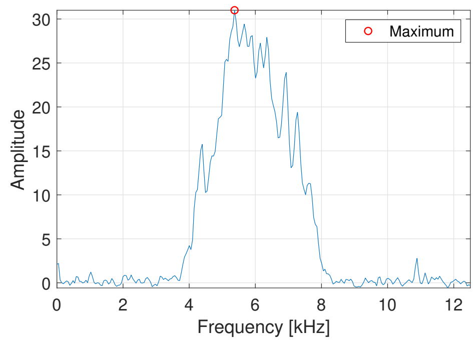
Figure 7. Spectrograms of the simulated signals [14]



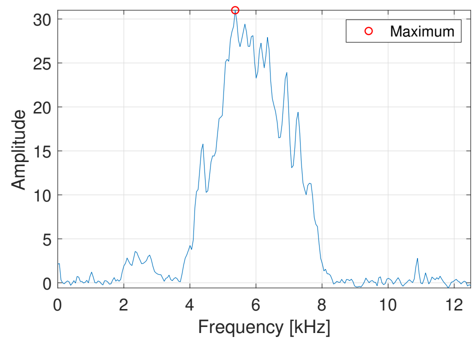
(a) s_1 – Gaussian noise



(b) s_2 – Gaussian noise with cyclic impulses

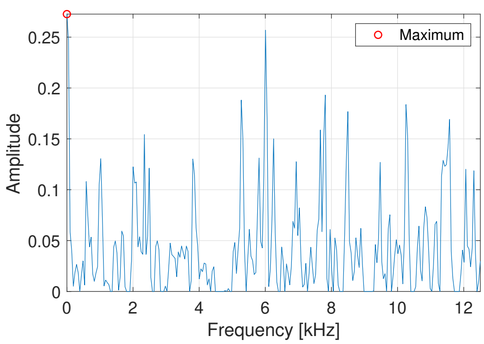


(c) s_3 – non-Gaussian noise

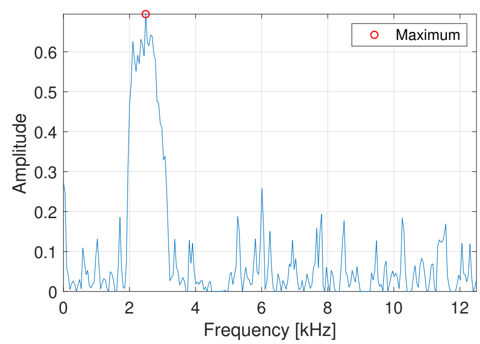


(d) s_4 – non-Gaussian noise with cyclic impulses

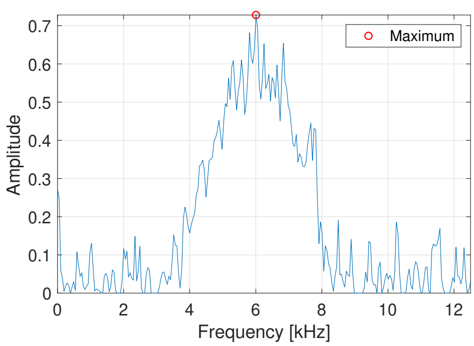
Figure 8. Spectral kurtosis for simulated signals [14]



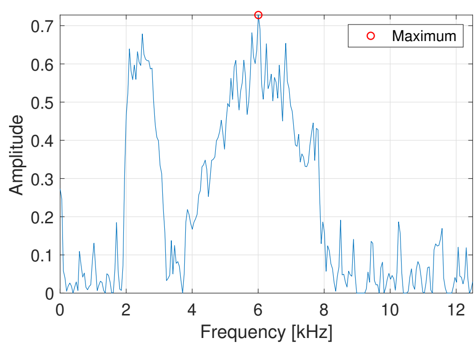
(a) s_1 – Gaussian noise



(b) s_2 – Gaussian noise with cyclic impulses

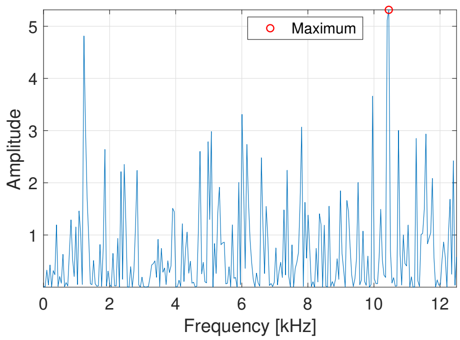


(c) s_3 – non-Gaussian noise

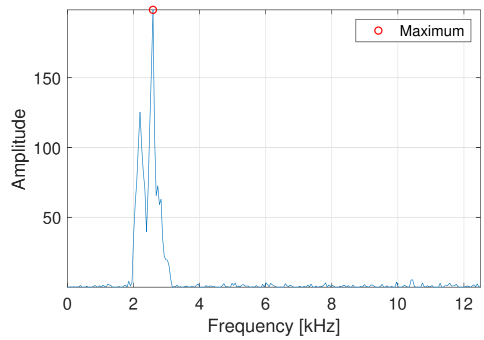


(d) s_4 – non-Gaussian noise with cyclic impulses

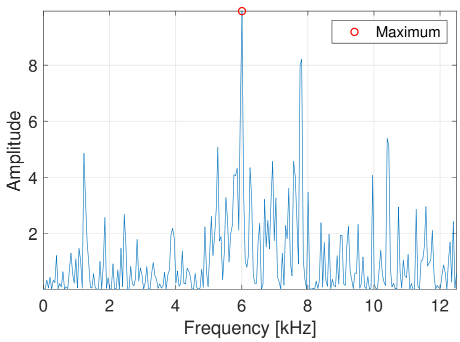
Figure 9. Alpha selector for simulated signals [14]



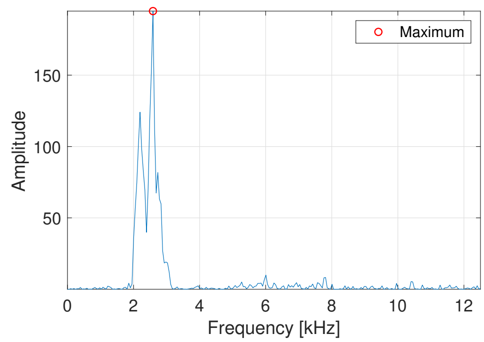
(a) s_1 – Gaussian noise



(b) s_2 – Gaussian noise with cyclic impulses

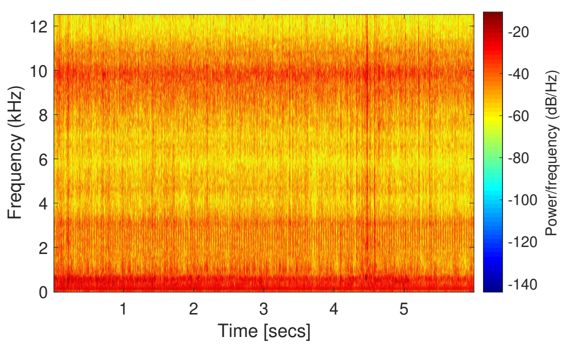


(c) s_3 – non-Gaussian noise

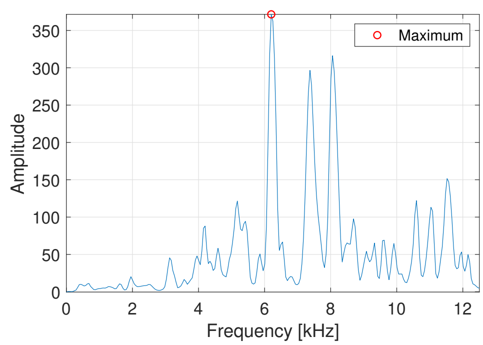


(d) s_4 – non-Gaussian noise with cyclic impulses

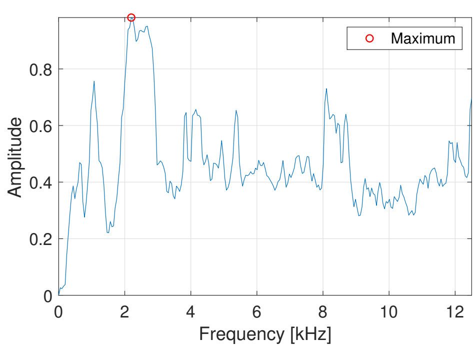
Figure 10. CVB selector for simulated signals [14]



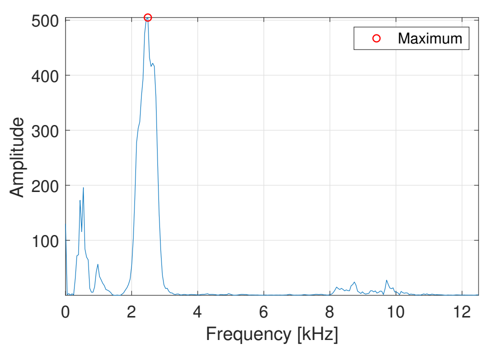
(a) Spectrogram



(b) Kurtosis selector



(c) Alpha selector



(d) CVB selector

Figure 11. Results for real signal from crushing machine [14]

where $\operatorname{cov}(\cdot,\cdot)$ is the covariance function, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y. The empirical equivalence of ρ_{XY} for $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, denoted ρ_{xy}, is defined as [8]

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where $\bar{x}, \bar{y}$ are sample means of data vectors x and y respectively. The Pearson correlation is usually a good dependency measure for finite-variance signals. However, this measure is sensitive to outliers. In our study, the Pearson correlation coefficient is applied to the sub-signals $S_1, S_2, \ldots, S_k$; see Figure 1 (b).

5.2 Spearman correlation map

Let us again consider a bi-dimensional sample $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ corresponding to a random vector (X, Y). Each pair (X, Y) and its empirical counterpart (x_i, y_i) correspond to a pair (Q, W) and (q_i, w_i), where q_i is the rank of the observation x_i in the sample $x_1, x_2, \ldots, x_N$ and w_i is the rank of the observation y_i in the sample $y_1, y_2, \ldots, y_N$. The Spearman rank correlation coefficient for the vector (X, Y) is defined as [23]

$$r_{XY} = \frac{\operatorname{cov}(Q, W)}{\sigma_Q \sigma_W},$$

where $\operatorname{cov}(\cdot,\cdot)$ is the covariance function, σ_Q and σ_W are the standard deviations of the rank variables. The empirical version of the Spearman correlation coefficient is given by

$$r_{xy} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(q_i - \bar{q})(w_i - \bar{w})}{\big[\frac{1}{N-1}\sum_{i=1}^{N}(q_i - \bar{q})^2 \frac{1}{N-1}\sum_{i=1}^{N}(w_i - \bar{w})^2\big]^{1/2}},$$

where $\bar{q}$ and $\bar{w}$ are sample means in the relevant rank samples.

The Spearman correlation takes values in the interval $[-1, 1]$ and investigates a monotonic relationship, in contrast to the Pearson correlation, which analyses a linear relationship. Outliers do not disturb the Spearman correlation, whereas they do in the case of the Pearson correlation.

As we did for the Pearson correlation, in our study, we apply the Spearman correlation to the sub-signals $S_1, S_2, \ldots, S_k$; see Figure 1 (b).

5.3 Kendall correlation map

The formula for the Kendall τ coefficient can be written as follows [22]:

$$\tau = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} J((x_i, y_i), (x_j, y_j)),$$

where

$$J((x_i, y_i), (x_j, y_j)) = \operatorname{sgn}(x_i - y_i)\operatorname{sgn}(x_j - y_j)$$

and $J((x_i, y_i), (x_j, y_j)) = 1$ if a pair (x_i, y_i) is concordant with a pair (x_j, y_j), i.e. if $(x_i - x_j)(y_i - y_j) > 0$; $J((x_i, y_i), (x_j, y_j)) = -1$ if a pair (x_i, y_i) is discordant with a pair (x_j, y_j), i.e. if $(x_i - x_j)(y_i - y_j) < 0$.

The Kendall correlation coefficient is based on the difference between the probability that two variables are in the same order (for the observed data vector) and the probability that their order is different. The Kendall coefficient indicates not only the strength but also the direction of the dependency. Like the Spearman correlation, it is resistant to outliers. In our study, the Kendall correlation is applied to the sub-signals from the time-frequency representation of the signal.

In the study [32], additional enhancements of the proposed dependency maps were proposed. We refer the readers to this position for more details.

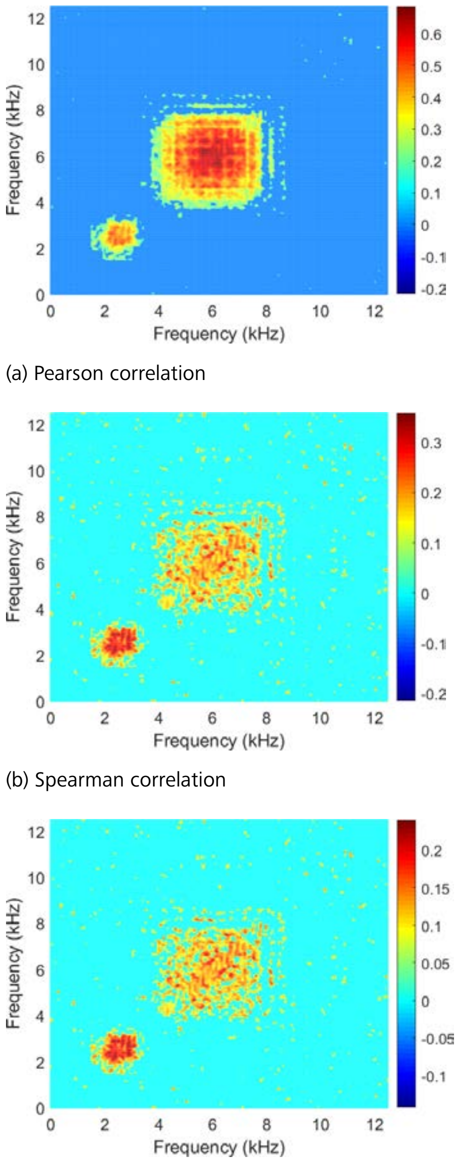
6 Comparative study of dependency measure applications – analysis of simulated and real signals

In this section, we present results of IFB selection by using the above-mentioned dependency measures. The analyses are performed for the simulated signal s_4, presented in Figure 6 (d). Recall that this is the signal that contains the non-Gaussian noise with cyclic impulses. In Figure 12, we present the enhanced dependency maps for the three correlation coefficients and for the signal s_4 by using the procedure presented in [32].

As one can see in Figure 12, in the case of the Pearson correlation map, the result differs from the other correlation maps. The correlation values at the location of the non-cyclic impulses are higher than for the cyclic impulses. In the other cases, one can see the opposite result. This result indicates that the application of more robust dependency measures may help to identify the cyclic impulsive behaviour in the case of impulsive noise.

Finally, we apply the dependency measures to the real signal from the crushing machine presented in Section 4. The real signal is depicted in Figure 4. In Figure 13, we present the enhanced dependency maps based on the Pearson, Spearman and Kendall correlations discussed in the previous section. We can see that the clear picture can be obtained using the robust measures (i.e. Spearman and Kendall correlation maps), where the IFB is indicated properly. The Pearson correlation map reacts to the frequency band related to the non-cyclic impulses; thus it cannot be used as a proper measure for the cyclostationary analysis in the case of non-Gaussian noise. Related discussions are presented in [25].

Acknowledgements. Let me thank Krzysztof Burnecki for the kind invitation to write this article. I would also like to thank Radosław Zimroz and the scientists in the Digital Mining Center (Faculty of Geoengineering, Mining and Geology at Wrocław University of Science and Technology, WUST) for our almost ten years of

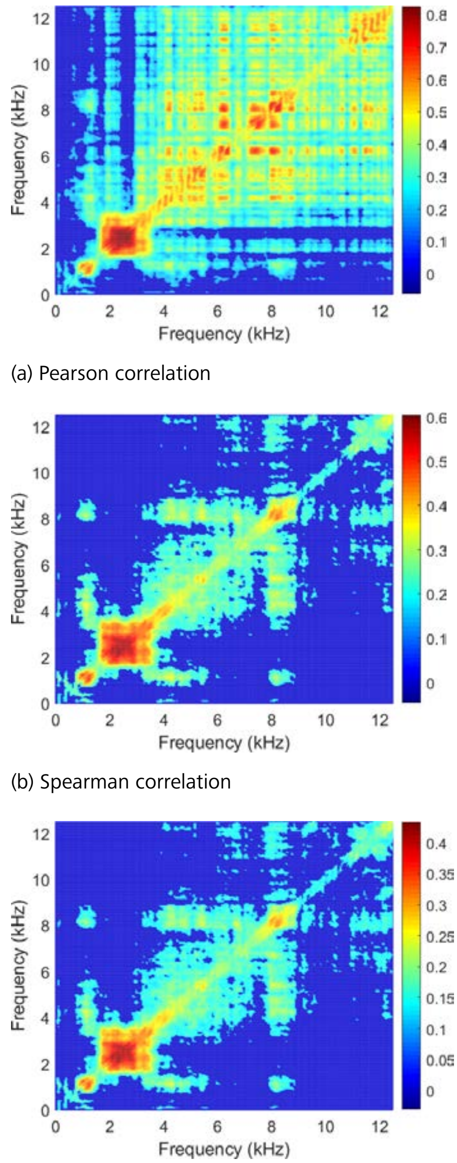


(a) Pearson correlation

(b) Spearman correlation

(c) Kendall correlation

Figure 12. Enhanced correlation maps for the simulated signal s_4 [32]



(a) Pearson correlation

(b) Spearman correlation

(c) Kendall correlation

Figure 13. Enhanced correlation maps for the real signal s_4 [32]

interdisciplinary research on new methods for real signal analysis. Thank you for showing me that mathematics can really be useful in practical applications, e.g. in the mining industry. Finally, I thank my PhD students and indeed all my students for providing new insights into our research, and positive energy.

The research was done within the Hugo Steinhaus Center at WUST whose goal is to organise, encourage and support research and education in numerical and stochastic techniques as applied in science and technology.

References

[1] J. Antoni, Fast computation of the kurtogram for the detection of transient faults. *Mech. Syst. Signal Process.* **21**, 108–124 (2007)

[2] J. Antoni, Cyclostationarity by examples. *Mech. Syst. Signal Process.* **23**, 987–1036 (2009)

[3] B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press (2015)

[4] S. Borak, W. Härdle and R. Weron, Stable distributions. In *Statistical tools for finance and insurance*, Springer, Berlin, 21–44 (2005)

[5] P. Borghesani and J. Antoni, Cs2 analysis in presence of non-Gaussian background noise–effect on traditional estimators and resilience of log-envelope indicators. *Mech. Syst. Signal Process.* **90**, 378–398 (2017)

[6] P. Borghesani and M. R. Shahriar, Cyclostationary analysis with logarithmic variance stabilisation. *Mech. Syst. Signal Process.* **70**, 51–72 (2016)

[7] F. Chaari, J. Leskow, A. Napolitane, R. Zimroz and A. Wylomanska, *Cyclostationarity: theory and methods – IV*. Springer (2020)

[8] O. J. Dunn and V. A. Clark, *Basic statistics: a primer for the biomedical sciences*. John Wiley & Sons, Hoboken, NJ (2009)

[9] Z. Feng, M. Liang and F. Chu, Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples. *Mech. Syst. Signal Process.* **38**, 165–205 (2013)

[10] G. Gelli, L. Izzo and L. Paura, Cyclostationarity-based signal detection and source location in non-Gaussian noise. *IEEE Trans. Commun.* **44**, 368–376 (1996)

[11] V. Girondin, K. M. Pekpe, H. Morel and J.-P. Cassar, Bearings fault detection in helicopters using frequency readjustment and cyclostationary analysis. *Mech. Syst. Signal Process.* **38**, 499–514 (2013)

[12] J. Hebda-Sobkowicz, J. Nowicki, R. Zimroz and A. Wyłomańska, Alternative measures of dependence for cyclic behaviour identification in the signal with impulsive noise – application to the local damage detection. *Electronics* **10** (2021)

[13] J. Hebda-Sobkowicz, R. Zimroz, M. Pitera and A. Wylomanska, Informative frequency band selection in the presence of non-Gaussian noise – a novel approach based on the conditional variance statistic. *Mech. Syst. Signal Process.* **145** (2020)

[14] J. Hebda-Sobkowicz, R. Zimroz and A. Wyłomańska, Selection of the informative frequency band in a bearing fault diagnosis in the presence of non-Gaussian noise – comparison of recently developed methods. *Applied Sciences* **10** (2020)

[15] A. Janicki and A. Weron, *Simulation and chaotic behavior of α-stable stochastic processes*. Monographs and Textbooks in Pure and Applied Mathematics 178, Marcel Dekker, Inc., New York (1994)

[16] A. K. Jardine, D. Lin and D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **20**, 1483–1510 (2006)

[17] P. Jaworski and M. Pitera, The 20-60-20 rule. *Discrete Contin. Dyn. Syst. Ser. B* **21**, 1149–1166 (2016)

[18] D. Jelito and M. Pitera, New fat-tail normality test based on conditional second moments with applications to finance. *Statistical Papers* **62**, 2083–2108 (2020)

[19] J.-H. Jeon and Y.-H. Kim, Localization of moving periodic impulsive source in a noisy environment. *Mech. Syst. Signal Process.* **22**, 753–759 (2008)

[20] D. Joanes and C. Gill, Comparing measures of sample skewness and kurtosis. *J. R. Stat. Soc. Ser. D Stat.* **47**, 183–189 (1998)

[21] V. Katkovnik, Robust m-periodogram. *IEEE Trans. Signal Process.* **46**, 3104–3109 (1998)

[22] M. G. Kendall, A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938)

[23] M. G. Kendall, *Rank correlation methods*. Griffin (1948)

[24] I. A. Koutrouvelis, Regression-type estimation of the parameters of stable laws. *J. Amer. Statist. Assoc.* **75**, 918–928 (1980)

[25] P. Kruczek, R. Zimroz and A. Wyłomańska, How to detect the cyclostationarity in heavy-tailed distributed signals. *Signal Processing* 107514 (2020)

[26] R. Liu, B. Yang, E. Zio and X. Chen, Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process.* **108**, 33–47 (2018)

[27] Y. Liu, T. Qiu and H. Sheng, Time-difference-of-arrival estimation algorithms for cyclostationary signals in impulsive noise. *Signal Processing* **92**, 2238–2247 (2012)

[28] Y. Liu, Q. Wu, Y. Zhang, J. Gao and T. Qiu, Cyclostationarity-based DOA estimation algorithms for coherent signals in impulsive noise environments. *EURASIP Journal on Wireless Communications and Networking* **2019**, 81 (2019)

[29] J. Lundén, S. A. Kassam and V. Koivunen, Robust nonparametric cyclic correlation-based spectrum sensing for cognitive radio. *IEEE Trans. Signal Process.* **58**, 38–52 (2010)

[30] J. H. McCulloch, Simple consistent estimators of stable distribution parameters. *Comm. Statist. Simulation Comput.* **15**, 1109–1136 (1986)

[31] A. Napolitano, Cyclostationarity: New trends and applications. *Signal Processing* **120**, 385–408 (2016)

[32] J. Nowicki, J. Hebda-Sobkowicz, R. Zimroz and A. Wyłomańska, Dependency measures for the diagnosis of local faults in application to the heavy-tailed vibration signal. *Applied Acoustics* **178**, 107974 (2021)

[33] J. Obuchowski, A. Wyłomańska and R. Zimroz, Selection of informative frequency band in local damage detection in rotating machinery. *Mech. Syst. Signal Process.* **48**, 138–152 (2014)

[34] J. Obuchowski, R. Zimroz and A. Wylomanska, Identification of cyclic components in presence of non-Gaussian noise – application to crusher bearings damage detection. *Journal of Vibroengineering* **17** (2015)

[35] R. B. Randall, *Vibration-based condition monitoring: industrial, aerospace and automotive applications*. John Wiley & Sons (2011)

[36] R. B. Randall and J. Antoni, Rolling element bearing diagnostics – a tutorial. *Mech. Syst. Signal Process.* **25**, 485–520 (2011)

[37] R. B. Randall, J. Antoni and S. Chobsaard, The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals. *Mech. Syst. Signal Process.* **15**, 945–962 (2001)

[38] N. A. Salehzadeh and A. M. Ferri, *Vibration-based techniques for damage detection and localization in engineering structures*. World Scientific (2018)

[39] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes*. Stochastic Modeling, Chapman & Hall, New York (1994)

[40] W. A. Smith, P. Borghesani, Q. Ni, K. Wang and Z. Peng, Optimal demodulation-band selection for envelope-based diagnostics: A comparative study of traditional and novel tools. *Mech. Syst. Signal Process.* **134**, 106303 (2019)

[41] D. Wang, Some further thoughts about spectral kurtosis, spectral L2/L1 norm, spectral smoothness index and spectral gini index for characterizing repetitive transients. *Mech. Syst. Signal Process.* **108**, 360–368 (2018)

[42] Y. Wang, J. Xiang, R. Markert and M. Liang, Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications. *Mech. Syst. Signal Process.* **66**, 679–698 (2016)

[43] P. H. Westfall, Kurtosis as peakedness, 1905–2014. R.I.P. *Amer. Statist.* **68**, 191–195 (2014)

[44] J. Wodecki, P. Kruczek, A. Bartkowiak, R. Zimroz and A. Wyłomańska, Novel method of informative frequency band selection for vibration signal using nonnegative matrix factorization of spectrogram matrix. *Mech. Syst. Signal Process.* **130**, 585–596 (2019)

[45] J. Wodecki, A. Michalak, A. Wyłomańska and R. Zimroz, Influence of non-Gaussian noise on the effectiveness of cyclostationary analysis – simulations and real data analysis. *Measurement* **171**, 108814 (2021)

[46] A. Wylomanska, R. Zimroz and J. Janczura, Identification and stochastic modelling of sources in copper ore crusher vibrations. In *Journal of Physics: Conference Series*, 628, IOP Publishing, 012125 (2015)

[47] A. Wyłomańska, R. Zimroz, J. Janczura and J. Obuchowski, Impulsive noise cancellation method for copper ore crusher vibration signals enhancement. *IEEE Trans. Ind. Electron.* **63**, 5612–5621 (2016)

[48] G. Yu, C. Li and J. Zhang, A new statistical modeling and detection method for rolling element bearing faults based on alpha–stable distribution. *Mech. Syst. Signal Process.* **41**, 155–175 (2013)

[49] X. Zhao, Y. Qin, C. He, L. Jia and L. Kou, Rolling element bearing fault diagnosis under impulsive noise environment based on cyclic correntropy spectrum. *Entropy* **21**, 50 (2019)

[50] Y. Zhao, T. Li, X. Zhang and C. Zhang, Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews* **109**, 85–101 (2019)

Agnieszka Wyłomańska received her M.Sc. degree in Financial and Insurance Mathematics from the Institute of Mathematics and Computer Science at the Wrocław University of Technology (WUT), Poland, in 2002, and her Ph.D. degree in Mathematics from WUT in 2006. Currently, she is a Professor at Wrocław University of Science and Technology (WUST) and a member of the Hugo Steinhaus Center. In 2015 she received a D.Sc. degree in Mining and Geology from Faculty of Geoengineering, Mining and Geology, WUST. Her areas of interest relate to time series analysis, stochastic modeling, and statistical analysis of real data (especially technical data related to the mining industry, indoor air quality, and financial time series). She is the author of more than 100 research papers, and cooperates regularly with industrial companies, especially from the mining industry.

agnieszka.wylomanska@pwr.edu.pl

A Diophantine equation concerning *epimoric* ratios

Preda Mihăilescu and Daniel Muzzulini

In this paper, we solve an interesting Diophantine equation that is born from classical questions of music theory.

1 Introduction

This paper investigates a Diophantine equation derived from a principle of construction for musical harmonies and scales advocated in antiquity by Claudius Ptolemy (c. 85–160 AD), known also for his geocentric model of celestial motion. Ptolemy's "Harmonics" was one of the main sources for Greek music theory in the Middle Ages and remained influential in the Renaissance up to the 17th century [3, 4, 16].

Retrospectively, the theory of proportions in the Pythagorean tradition can be considered a theory for rational numbers greater than one under multiplication[1]. In its application to musical harmonies and scales, adding musical intervals corresponds to the multiplication of rational numbers, and therefore piling equal intervals, i.e., multiplying a given interval by an integer, is equivalent with raising its ratio to the respective integer power[2].

According to Ptolemy's music-aesthetic premises, the multiple and the so-called *epimoric* ratios are the building blocks of musical harmonies and scales. Epimoric ratios (also called *superparticular* ratios) are positive rational numbers of the form $\frac{n+1}{n}$, whereas multiple ratios of the form $\frac{n}{1}$ are ordinary natural numbers. Since they can be written as the unit plus a unit fraction $(1 + \frac{1}{n})$, epimoric ratios can be regarded as an elementary form of improper fractions accessible to perception guided by the intellect, see [3, pp. 60–62], [4] and [7, pp. 191–200]. The role of epimoric ratios in music theory and pitch perception was repeatedly emphasized and problematized throughout the course of history – in the second half of the 16th century for instance, they were debated by Gioseffo Zarlino (1517–1590) and Vincenzo Galilei[3] (1520–1591) [14, 15]. On the other hand, epimoric ratios also played a crucial role in the development of novel calculation techniques: Simon

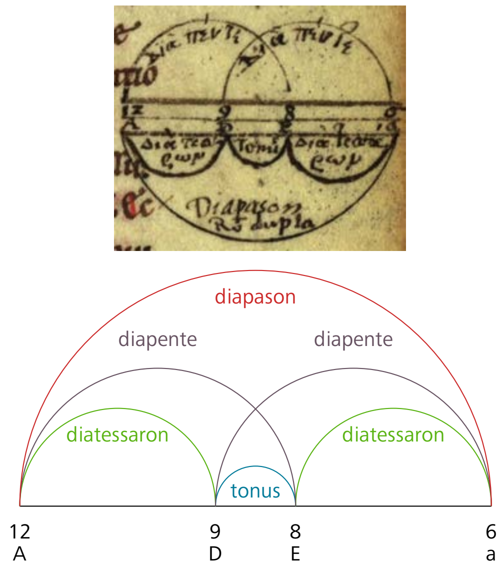


Figure 1. Arc diagram showing a complete graph with epimoric ratios. The nodes are labelled with numbers and note names, and the arcs with the Greek and Latin names of the related intervals. The six pairs of numbers *tonus* (whole tone), *diatessaron* (fourth), *diapente* (fifth), *diapason* (octave) form epimoric ratios. The diagram is remarkable because it represents equal intervals, i.e., logarithms of ratios, as equal semicircles, and not on a linear scale. It is a marginal note by Swiss music theorist Heinrich Loriti Glareanus (1488–1563) added to a manuscript copy from c. 1200 of the "Micrologus" by Guido of Arezzo (c. 991–992 until after 1033).

[1] We are aware that the antique ratios are not yet fully fledged rational numbers, i.e., classes of ordered pairs of integers. Whereas ratios were considered equivalent to their representations in lowest terms, the order of the terms of a ratio was not constitutive. The ratios 3 : 4 and 4 : 3, for instance, represent the same relationship (the *epitriti* or *sesquitertia*) corresponding to the musical interval of the fourth (the diatessaron, see Figure 1) in the sense of a perceptual distance. For our purpose the restriction to rational numbers greater than 1 is sufficient and convenient.

[2] The term musical interval refers to logarithms of frequency ratios. Many quantifiable sensory phenomena and their physical counterparts are in a logarithmic or nearly logarithmic relationship, as loudness or brightness sensation with respect to the intensity of sound or light.

[3] The father of Galileo.

1	4	16	64	256	1024
	5	20	80	320	1280
		25	100	400	1600
			125	500	2000
				625	2500
					3125

Figure 2. Triangular table used by Boethius to calculate finite geometric sequences of integers with the common epimoric factor 5/4. Only the first row requires repeated multiplication by 4. The numbers in any row can be found by adding neighbors from the previous one:

$$1 + 4 = 5, \quad 4 + 16 = 20, \quad 16 + 64 = 80, \quad \text{etc.}$$

The main diagonal direction holds the powers of 5, and the columns contain the geometric sequences in lowest terms. From the fourth column one can read that three major thirds 5/4 are smaller than an octave because $(5/4)3(128/125) = 2$. The method can be used for arbitrary epimoric ratios. There are examples for 3/2, 4/3, 5/4 and 9/8 in medieval Boethius manuscripts.

Stevin (1548/49–1620) used various epimoric bases in his tables of compound interest, and Jost Bürgi (1552–1632) created a fine-grained and very accurate exponential table with more than 23,000 entries for the epimoric base $1 + \frac{1}{10{,}000}$, see [17], [18, p. 75] and [11, pp. 199–200, 209–210].

Archytas of Tarentum (c. 420–c. 350 BC) proved that the equation

$$\left(\frac{n+1}{n}\right)^k = \frac{s+1}{s}$$

has no integer solutions in n, s for integer exponents $k > 1$ (see [2]). Archytas' reasoning was discussed by Boethius (c. 477–524 AD) [9, pp. 451–470], who also quotes a more general result and its proof from the Euclidean "*Sectio Canonis*"[4], stating that *epimoric ratios can be decomposed in no way into products of two or more equal integer ratios* – [5, Inst. Mus. IV.2, p. 303], [6, p. 118].

In other words, musical intervals of an epimoric ratio, such as the octave ($s = 1$) or the fourth ($s = 3$), cannot be divided equally into smaller intervals of epimoric ratios, i.e., equal division results in "irrational ratios"[5]. This fact makes it impossible to construct musical scales or scale segments with several equal intervals of epimoric ratios spanning for example a fourth or an octave. Our theorem, the main result of this paper, proves that a similar restriction holds for the partition of an epimoric ratio into a power of an epimoric ratio and a single epimoric cofactor: It shows that for positive integers q, r, s and exponent $k > 2$, equation (1) below has no solutions[6]. Introducing an epimoric cofactor, $\frac{r+1}{r}$, into the decomposition raises the upper bound k for solvability only by 1.

[4] See [1, p. 195]; the assignment of the "*Sectio Canonis*" to Euclid is insecure.

[5] Irrationality had a precarious ontological status as being defined only ex negativo, and it was linked to incommensurable (geometric) quantities, see also [12].

2 The main theorem

We will prove the following theorem.

Theorem 1. *The Diophantine equation*

$$\left(\frac{q+1}{q}\right)^k \cdot \frac{r+1}{r} = \frac{s+1}{s}, \quad q, r, s \in \mathbb{N}, \, k > 2, \tag{1}$$

has no integer solutions.

The following derivation holds for all $k \geq 2$. The value $k = 2$ allows for an infinity of solutions; it will be considered in detail in a separate paper. By removing denominators, we obtain the following two equivalent formulations of (1):

$$(q+1)^k \cdot (r+1)s = q^k r(s+1); \tag{2}$$

$$((q+1)^k - q^k)(r+1)s = q^k(rs + r - rs - s) = q^k(r - s). \tag{3}$$

The last identity in (3) implies that $r - s > 0$, so we can in particular divide by $r - s$.

We define $\delta = (q+1)^k - q^k$ and note that

$$\begin{aligned} \delta &= kq^{k-1} \cdot \left(1 + \frac{k-1}{2}\frac{1}{q} + O\left(\frac{1}{q^2}\right)\right) \\ &< kq^{k-1} \cdot (q(e^{1/q} - 1)). \end{aligned} \tag{4}$$

In both equations (2) and (3), we encounter various pairs of factors of the type $(x, x+1)$ for some $x \in \mathbb{N}$, for instance $s, s+1$ in (2). These are coprime and in order to exploit this useful fact, we define a series of factors whose existence follows from such relations of coprimality. Namely, we set $A = (r+1, r-s) = (r+1, s+1)$ and $B = (s, r-s) = (r, s)$. We assume $A, B > 0$ and let the cofactors be $AC = s+1$ and $BD = s$ for some $C, D \in \mathbb{N}$. Since $(q^k, (q+1)^k) = 1$, it follows that $q^k \mid (r+1)s$ and $\delta \mid r - s$. By combining the last relations of divisibility with the definitions of A and B, we get

$$\begin{gathered} (r+1)s = AB \cdot q^k, \quad r - s = AB \cdot \delta, \\ \frac{q^k}{\delta} = s \cdot \left(1 + \frac{s+1}{r-s}\right) > s. \end{gathered} \tag{5}$$

Using the upper bound on δ in (4), we obtain the following estimates for s.

[6] The conjecture that (1) has no solutions was formulated by the second author, from the study of music theory and on the basis of his mathematical background. It was eventually settled in a joint effort of the two authors.

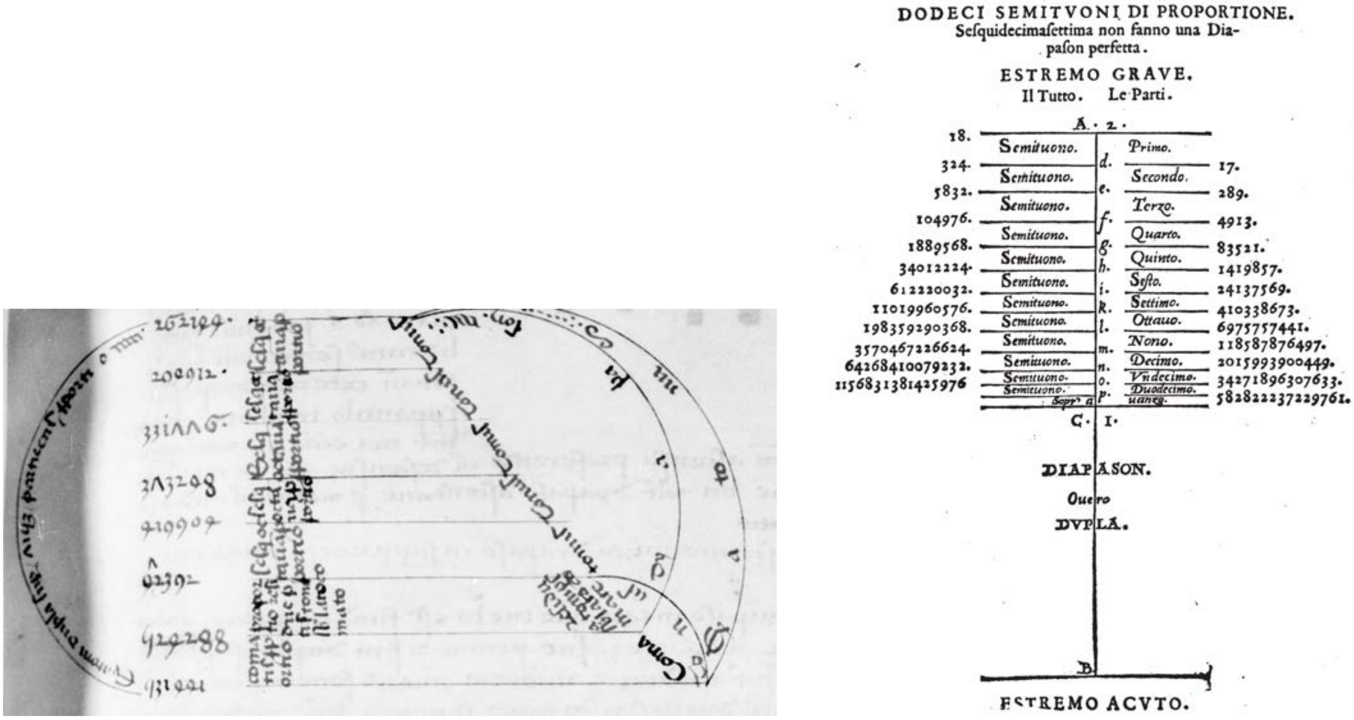


Figure 3. Left: Since twice the number on top 262,144 ($= 8^6$) is less than 531,441 ($= 9^6$), six epimoric tones (9/8) are by a Pythagorean comma ($531{,}441/524{,}288 = 3^{12}/2^{19}$) greater than the octave, as the diagram by Jacobus Leodiensis (14th century) illustrates. *Right*: Twelve epimoric semitones of the ratio 18/17, however, are a little smaller than an octave, as Gioseffo Zarlino's monochord calculations show. The epimoric ratio 18/17 was proposed by Vincenzo Galilei as a substitute for the 12th root of two – it is the best epimoric approximation to the semitone of our modern piano tuning.
The monochord string AB is divided at C in the middle, the horizontal system of lines $d, e, f, \ldots, p$ indicate fret positions on the string, which is to be plucked between the frets and B (*estremo acuto*) to give the corresponding notes of this regular scale with a small non-epimoric gap between p and C.

Lemma 1. *Assuming that* (1) *has non-trivial integral solutions, s obeys the bounds*

$$\frac{q}{k\mu} < s < \frac{q^k}{\delta} < \frac{q}{k}, \quad \text{with } \mu = q(e^{1/q} - 1) \cdot \left(1 + \frac{s+1}{r-s}\right). \tag{6}$$

In particular, $q > k$.

Proof. The upper bound for s follows from (5); the lower bound follows from the same identity, in conjunction with the upper bound in (4). Since $s \geq 1$, we obtain our first lower bound for q, namely $\frac{q}{k} > s \geq 1$, hence $q > k$. ∎

Our next task is to derive from the above and some additional bounds, a tight interval which must contain s; en route we also obtain sharper lower bounds on q.

Lemma 2. *Under the same assumption as above, we let*

$$Q = \left(\frac{q}{k} + 1\right) \cdot \frac{q}{k} \quad \text{and} \quad U \coloneqq \frac{q^k - Q}{\delta}, \quad V \coloneqq \frac{q^k - 1}{\delta}.$$

Then

$$s \in I \coloneqq (U, V). \tag{7}$$

Moreover, $q > (k-1)^{k+1}$ and there is at most one integer $\sigma \in I \cap \mathbb{N}$. In particular, if (1) *has a solution, then $s = \sigma$.*

Proof. We have

$$r = s + r - s = AB \cdot \left(\frac{q^k}{r+1} + \delta\right).$$

Now, $(B, r+1) = 1$ and $A \mid (r+1)$, so the previous becomes

$$r = AB\delta + B\frac{q^k}{(r+1)/A}.$$

Since $(B, r+1) = 1$, it follows that $(r+1) \mid Aq^k$, and in (5), we find

$$B \cdot \left(\frac{Aq^k}{r+1}\right) = s.$$

Since $(A, s) = 1$, we get

$$\left(\frac{s}{B}\right) = D \mid q^k,$$

hence $s = BD \mid rq^k$. Reinserted in (5) with the definition $r \coloneqq r' \cdot B$, this leads to

$$r + 1 = A\frac{q^k}{D} \quad \text{and} \quad r' - D = A \cdot \delta.$$

So

$$r = A\left(\frac{q^k}{D}\right) - 1 \quad \text{and} \quad r' = \frac{r}{B} = A\delta + D,$$

thus

$$s = \frac{Aq^k - D}{A\delta + D} \in \mathbb{N};$$

consequently,

$$s \cdot \delta + \frac{sD}{A} = q^k - \frac{D}{A}, \tag{8}$$

$$CD = D\,\frac{s+1}{A} = q^k - s\delta = q^k(s+1) - s(q+1)^k.$$

Since $CD \le s(s+1) < Q$, we conclude that

$$\frac{q^k - 1}{\delta} \ge s = \frac{q^k - CD}{\delta} > \frac{k^2 q^k - q(q+k)}{k^2((q+1)^k - q^k)}.$$

Statement (7) follows from these inequalities, by inserting the definitions of U and V. The length of the interval I is $\ell = \frac{Q-1}{\delta}$; the improved lower bound on q will show that $\ell < \frac{1}{2}$ for $k \ge 3$, and thus the interval I contains at most one integer, which confirms the statement on σ.

Now $D \mid q^k$, so

$$B = \frac{\frac{q^k}{D} - C}{(q+1)^k - q^k}.$$

Since $D \le s < \frac{q}{k}$, we also have $\frac{q^k}{D} > kq^{k-1}$. This will lead to the bound for q. *Assume first* that $\frac{q^k}{D} \equiv 0 \bmod q$. Then $B + C \equiv 0 \bmod q$. But $B + C \le 2s + 1 < 2\frac{q}{k} + 1 < q$ for $q > k > 2$, so we obtain a contradiction to $B + C \ge 0$, and thus $B = C = 0$, which is absurd.

It remains to treat the case $\frac{q^k}{D} \not\equiv 0 \bmod q$. We decompose $D = ad^k$, so that all primes p that divide a are either coprime to q, or occur in a with a power less than k, while $d \mid (q, D)$. Then $\frac{q^k}{D} = a(\frac{q}{d})^k \equiv 0 \bmod \frac{q}{d}$ and (6) implies a fortiori that $Bkd^k < q$ holds along with $B + C \equiv 0 \bmod \frac{q}{d}$. Since $B, C > 0$, we have

$$\frac{q}{k} > C \ge \frac{q - Bd}{d} \implies dq > kq - Bkd > kq - \frac{q}{d^{k-1}},$$

and thus $d - (k - \frac{1}{d^{k-1}}) > 0$ and a fortiori $d \ge k - 1$. In particular, q must be *large*, namely

$$q > B(k-1)^{k+1} \ge (k-1)^{k+1}, \tag{9}$$

as claimed. Using this bound, a straightforward verification shows that $\ell < \frac{1}{2}$, and this completes the proof of Lemma 2. ∎

We finally use the bound (9) and sharper estimates for δ to complete the proof of Theorem 1. If $I \cap \mathbb{N} = \varnothing$, then there are no solutions, and we are done. Otherwise, we let σ be the unique integer in the interval I. Since $s \in I$ is also an integer, it follows that $s = \sigma$.

Proof of Theorem 1. We determine σ in terms of $\frac{q}{k}$. We have

$$\delta \coloneqq (q+1)^k - q^k = q^{k-2}\left(qk + \binom{k}{2} + \frac{1}{q}\binom{k}{3} + \rho\right),$$

with

$$|\rho| \le \begin{cases} 0 & \text{for } k = 3, \\ \dfrac{k^4}{4!\,q^2} < \dfrac{1}{20q} & \text{for } k > 3, \\ & \text{using } q > (k-1)^{k+1}. \end{cases}$$

Consequently,

$$\delta\sigma = q^{k-2}\sigma\left(qk + \binom{k}{2} + \frac{1}{q}\binom{k}{3} + \rho\right) = q^k - CD,$$

$$\sigma \cdot \left(qk + \binom{k}{2}\right) + \frac{\sigma}{q}\binom{k}{3} = q^2 - \rho_1,$$

where

$$|\rho_1| = \left|\frac{CD}{q^{k-2}} - \sigma\rho\right|.$$

Since $\sigma < \frac{q}{k}$, there is a number e of the form $e = \{\frac{q}{k}\} + n$, with $n \in \mathbb{Z}_{\ge 0}$ and $\{\frac{q}{k}\}$ denoting the fractional part of $\frac{q}{k}$, such that

$$\sigma = \frac{q}{k} - e = \left[\frac{q}{k}\right] - n, \quad ek = q - k\sigma.$$

We claim that $n = 0$. The definition of U implies that

$$nk = q - [q - k\sigma] \le \frac{kq^k(1 + \frac{k-1}{2q} + O(\frac{1}{q^2})) - kq^k + \frac{q^2}{k}}{\delta}$$

$$\le \frac{\binom{k}{2}q^{k-1} + \frac{q^2}{k} + O(q^{k-2})}{kq^{k-1}} < \left\lceil \frac{k-1}{2} \right\rceil \le \frac{k+1}{2}.$$

Since n is an integer and $0 \le n < \frac{k+1}{2k}$, it follows that $n = 0$ and $0 < ek < k$, as claimed.

Thus, $\sigma = \frac{q}{k} - e = [\frac{q}{k}]$. Inserting this value of σ in (8) yields

$$q^2 - eqk + \frac{k-1}{2}q - \frac{ek(k-1)}{2} + \frac{(k-1)(k-2)}{6} - \binom{k}{3}\frac{e}{q} = q^2 - \rho_1,$$

hence

$$q \cdot \left(\frac{k-1}{2} - ek\right) = R' \coloneqq -\frac{ek(k-1)}{2} + \frac{(k-1)(k-2)}{6} + \rho_2, \tag{10}$$

with

$$|\rho_2| = \left|\rho_1 - \binom{k}{3}\frac{e}{q}\right| \le |\rho_1| + \frac{k(k-1)(k-2)}{6q}.$$

We have seen that $ek \in \mathbb{N}$, so if the left-hand side of (10) does not vanish, then its cofactor is an integer or a half-integer; if it does not vanish, its absolute value will exceed $\frac{q}{2}$. We denoted the right-hand side of (10) by R', so

$$|R'| \le |R| + \frac{k-1}{2} \cdot \left|\frac{k-2}{3} - ek\right| < \frac{k(k-1)}{3} + |R|.$$

For the right-hand side, small values of k allow for larger values of ρ_2, so we first assume $k \geq 4$. In this case

$$|\rho_2| \leq R := \frac{Q}{q^{k-2}} + \frac{\sigma}{20q} + \frac{k(k-1)(k-2)}{6(k-1)^5} < \frac{2}{(k-1)k}.$$

Now $|R| < (\frac{1}{k^2} + \frac{1}{qk} + \frac{1}{20k}) + \frac{6}{(k-1)^{k-2}} < \frac{1}{k} \cdot (\frac{2}{k} + \frac{1}{20})$ and inserting this in the bound for R', we see that $|R'| < (k-1)^2 < \frac{q}{2}$. Since the left-hand side is at least $\frac{q}{2}$ in absolute value, if it does not vanish, we conclude that the two sides must vanish simultaneously. Thus, $e = \frac{k-1}{2k}$ and the right-hand side is $\frac{k^2-1}{12} - \rho_2$. Since $|\rho_2| < 1$ and $\frac{k^2-1}{12} > 1$ for $k > 3$, the last expression cannot vanish for $k > 3$, so there are no solutions in this case.

The case $k = 3$ is more delicate; recall that in this case $\rho = 0$ and thus $\rho_2 = \frac{CD-e}{q}$. The best bound for the error term is now $0 < |\rho_2| < \frac{q}{9}$, so in (10)

$$q = 3e(q-1) + \frac{1}{3} + \rho_2 = 3e(q-1) + \frac{1}{3} + \frac{CD-e}{q}.$$

We generate a contradiction by a case-by-case examination. We know that $3e < k = 3$, so $3e \in \{0, 1, 2\}$. The cases $3e = 0$ and $3e = 2$ are easily seen to be impossible. In the first case, the right-hand side is too small, while in the second case it is too large, compared to q, as one verifies from the definitions.

If $3e = 1$, we obtain

$$q = q - 1 + \frac{1}{3} + \frac{CD - \frac{1}{3}}{q} \implies \frac{q + 3CD - 1}{3q} = 1,$$

thus $CD = \frac{2q+1}{3} = \frac{2q+1}{k}$. We have seen above that $D \mid q^k$, while $3CD = 2q + 1$ implies $D \mid (2q+1, q^3)$, and thus $D = 1$. But then $C = CD = \frac{2q+1}{3} > \frac{q}{3} + 1$, contradicting the upper bound $C \leq s + 1 < \frac{q}{k} + 1$ established above. We conclude that there are no solutions for $k = 3$ either, and this completes the proof. ∎

3 Remarks and comments

Here we provide some historical details that place our result in its musical context. For additional reading we recommend the excellent modern introduction to superparticular ratios[7] by Halsey and Hewitt [8].

3.1 Music theory

In order to briefly elucidate the musical context of the theorem, we give some examples. Historically, partitions of ratios are frequently written as ordered multi-term proportions within arc diagrams. Arrangements as proportions in lowest terms corresponding to the left-hand side of the following equalities are given in brackets.

With $k = 1$, the octave (2/1) can be divided into a fifth (3/2) and a fourth (4/3):

$$\frac{3}{2} \cdot \frac{4}{3} = \frac{2}{1} \quad (2 : 3 : 4)$$

and with $k = 2$, into two fourths and a whole tone (9/8), see Figure 1:

$$\left(\frac{4}{3}\right)^2 \cdot \frac{9}{8} = \frac{2}{1} \quad (6 : 8 : 9 : 12). \tag{11}$$

Likewise, the fifth (3/2) can be partitioned with two minor thirds (6/5) and a chromatic semitone (25/24):

$$\left(\frac{6}{5}\right)^2 \cdot \frac{25}{24} = \frac{3}{2} \quad (20 : 24 : 25 : 30).$$

However, no epimoric musical interval can be divided into four epimoric smaller intervals, of which three are equal $(k = 3)$. For example, the Pythagorean division of the fifth into three whole tones and a non-epimoric remainder,

$$\left(\frac{9}{8}\right)^3 \cdot \frac{256}{243} = \frac{3}{2} \quad (192 : 216 : 243 : 256 : 288), \tag{12}$$

or the division of the octave into three major thirds (5/4) and a *diesis* (128/125),

$$\left(\frac{5}{4}\right)^3 \cdot \frac{128}{125} = \frac{2}{1} \quad (64 : 80 : 100 : 125 : 128),$$

are prototypical: Whatever cubed epimoric ratio is chosen, the cofactors to 3/2 and 2/1 are never epimoric. The latter example illustrates that the just intonation major third (5/4) is an approximation to the problem of *doubling the cube*, whereas the *irrational* major thirds of the present-day equal division of the octave is a true solution beyond antique ratio theory[8].

Ptolemy's *tetrachords* (divisions of the fourth) involved three different epimoric ratios, as in

$$\frac{9}{8} \cdot \frac{10}{9} \cdot \frac{16}{15} = \frac{4}{3}.$$

Combining this with (11) results in the octave division

$$\left(\frac{9}{8}\right)^3 \cdot \left(\frac{10}{9}\right)^2 \cdot \left(\frac{16}{15}\right)^2 = \frac{2}{1},$$

where 9/8 and 10/9 define two varieties of whole tones and 16/15 a semitone larger than the Pythagorean 256/243. This partition can be used to define the diatonic scale in just intonation, see Figure 4 (left). The Pythagorean example (12) which fails to be made up solely of epimorics, is an indication for the origin of our problem (1).

[7] This is an other expression for *epimoric* ratios.

[8] Illustrations for $k = 6$ and $k = 12$, where $s = 1$ (the octave), from sources of the 14th and 16th century are given in Figure 3.

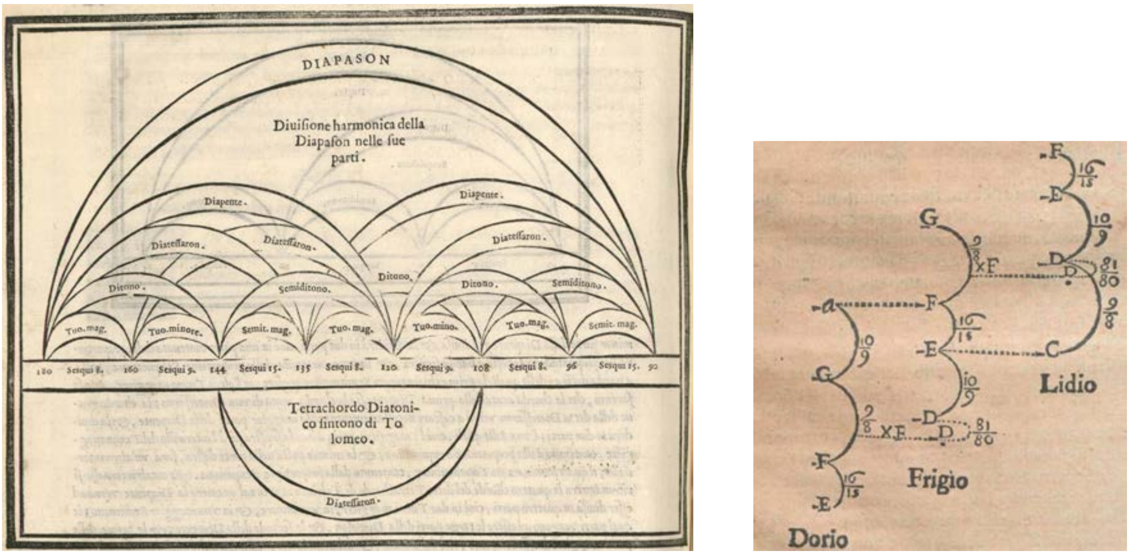


Figure 4. Left: Ptolemy's diatonic scale according to Gioseffo Zarlino. The baseline holds the proportion 90 : 96 : 108 : 120 : 135 : 144 : 160 : 180 (from right to left) with three major tones (9/8 = Sesqui 8), two minor tones (10/9 = Sesqui 9) and two (major) semitones (16/15 = Sesqui 15). The fourth (*diatessaron*) of the Ptolemaic tetrachord highlighted at the bottom consist of a semitone (16/15), a major tone (9/8) and a minor tone (10/9). Only intervals of epimoric ratios are labelled in this almost complete graph with eight nodes. *Right*: Greek tetrachords in the 17th century. Doni's system with three "epimoric tetrachords" E–A (dorian), D–G (phrygian) and C–F (lydian) uses two varieties of whole tone steps (9/8 and 10/9), semitones E–F (16/15) as well as two pitches for D differing by a syntonic comma (81/80) resulting in a fine grained system of pitches [10, pp. 62–69].

3.2 Diophantine equations

Music Theory stood more than once at the origin of fascinating Diophantine equations. For instance, the reputed Catalan equation[9] $x^u - y^v = 1$, stating that 8 and 9 are the only successive non-trivial powers of integers, generalizes the original question about $3^x - 2^y = 1$ considered by Philippe de Vitry (1291–1361) in relation with *harmonic numbers* and Platonic music theory, thus a Diophantine equation with actual connection to music. Levi ben Gershon (1288–1344) had proved that this particular equation does not have other solutions than $9 - 8 = 1$, and this already in the 13th century. Leonhard Euler (1717–1783) switched exponents and bases in the musical equation, and finally Catalan (1814–1894) allowed both bases and exponents to vary: both latter variations had left the common field of music and mathematics, and Diophantine equations were investigated for their pure mathematical interest.

The present equation (1) still has a lively connection to music theory. Were this of no more concern, one could imagine generalizations of (1) such as

$$\left(\frac{q+1}{q}\right)^k \cdot \left(\frac{r+1}{r}\right)^l = \left(\frac{s+1}{s}\right)^m, \quad q,r,s \in \mathbb{N},\, k,l,m > 2,$$

or, defining

$$\ell(q,m) = 1 + \frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{n+m},$$

one could look at

$$\ell(q,n)^k \cdot \ell(r,n)^l = \ell(s,n)^m, \quad q,r,s \in \mathbb{N},\, k,l,m,n > 2,$$

All this recalls the falling powers

$$x^{\underline{n}} = x(x-1)\cdots(x-n+1),$$

dear to Isaac Newton (1643–1727), and one finds a variant of Fermat's Last Theorem, that cannot be found among the dozen of variants mentioned in Ribenboim's *13 Lectures* [13], probably the most adequate source for verifying if a variant of Fermat's Equation has already received attention. This one apparently did not:

$$x^{\underline{p}} + y^{\underline{p}} = z^{\underline{p}}, \quad (x,y,z) = 1 \quad \text{and} \quad x,y,z > n+1.$$

We stop here and invite the reader to imagine his own favorite generalization, leaving it to the future to decide whether some of these variations will capture the attention of a larger number of mathematicians, professional or not.

[9] Catalan proposed this equation in a French journal in 1841 and it appeared in Crelle's Journal as a note to the Editor, in 1844.

Image sources

Figure 1: Guido of Arezzo (c. 1200), Micrologus, Ms. 8 Cod. Ms. 375 (Cim 13), fol. 53r. Source: München, Universitätsbibliothek

Figure 2: A. M. S. Boethius (early 10th c.), De institutione arithmetica., fol. 4v. Source: Medeltidshandskrift 1 (Mh 1), Lund University Library

Figure 3, left: Jacobus (Leodiensis) (15th c.), Speculum musicae, Ms. Latin 7207, Vol. III, Cap LXXXV, fol. 46r. Source: gallica.bnf.fr/ Bibliothèque nationale de France

Figure 3, right: G. Zarlino (1588), Sopplimenti musicali, Venetia: Francesco de Franceschi, Sanese, Lib IV, p. 205, https://s9.imslp.org/files/imglnks/usimg/d/d1/IMSLP129044-PMLP252086-terzo_volume.pdf (accessed February 21, 2022)

Figure 4, left: G. Zarlino (1562), Le istitutioni harmoniche, Venice, Italy, p. 122. https://digital.library.unt.edu/ark:/67531/metadc25955/ (accessed February 21, 2022), University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu; crediting UNT Music Library

Figure 4, right: G. B. Doni (1635), Compendio del Trattato de' Generi e de' Modi della Musica, Roma, p. 41. Source: Mus.th. 7234, Bayerische Staatsbibliothek München

References

[1] A. Barker, *Greek musical writings: II Harmonic and acoustic theory*. Cambridge University Press, Cambridge (1989)

[2] A. Barker, Ptolemy's Pythagoreans, Archytas, and Plato's conception of mathematics. *Phronesis* **39**, 113–135 (1994)

[3] A. Barker, *Scientific method in Ptolemy's "Harmonics"*. Cambridge University Press, New Haven (2000)

[4] A. Barker, Mathematical beauty made audible: Musical aesthetics in Ptolemy's "Harmonics". *Classical Philology* **105**, 403–420 (2010)

[5] Boethius, *De institutione arithmetica libri duo: De institutione musica libri quinque*. B. G. Teubner, Leipzig (1867)

[6] Boethius, *Fundamentals of music*. Yale University Press, New Haven (1989)

[7] P. Gozza, *Number to sound: The musical way to the scientific revolution*. The Western Ontario Series in Philosophy of Science 64, Springer, Dordrecht (2000)

[8] G. D. Halsey and E. Hewitt, More on the superparticular ratios in music. *Amer. Math. Monthly* **79**, 1096–1100 (1972)

[9] C. A. Huffman, *Archytas of Tarentum: Pythagorean, philosopher and mathematician king*. Cambridge University Press, Cambridge (2007)

[10] M. Kirnbauer, *Vieltönige Musik – Spielarten chromatischer und enharmonischer Musik in Rom in der ersten Häfte des 17. Jahrhunderts*. Schola Cantorum Basiliensis Scripta 3, Schwabe, Basel (2013)

[11] D. Muzzulini, The geometry of musical logarithms. *Acta Musicologica* **87**, 193–216, DOI 10.5281/zenodo.5541790 (2015)

[12] P. Pesic, Hearing the irrational: Music and the development of the modern concept of number. *Isis* **101**, 501–530 (2010)

[13] P. Ribenboim, *13 lectures on Fermat's last theorem*. Springer, New York (1979)

[14] B. V. Rivera, Theory ruled by practice: Zarlino's reversal of the classical system of proportions. *Indiana Theory Review* **16**, 145–160 (1995)

[15] F. Sargolini, La critica di Vincenzo Galilei al misticismo numerico di Gioseffo Zarlino. *Nuncius* **15**, DOI 10.1163/182539100X00038 (2000)

[16] J. Solomon, *Ptolemy, "Harmonics": Translation and commentary*. Brill, Leiden (2000)

[17] S. Stevin, *Tafelen van interest, midtgaders de constructie der selver …* Christoffel Plantijn, Antwerpen (1582)

[18] S. Stevin, *The principal works of Simon Stevin. Vol. IIA: Mathematics*. C. V. Swets & Zeitlinger, Amsterdam (1958)

Preda Mihăilescu studied mathematics and computer science in Zürich, receiving his PhD from ETH Zürich. He was active during 15 years in the industry, as a numerical analyst and cryptography specialist. In 2002, he proved Catalan's Conjecture. This number theoretical conjecture, formulated by the French mathematician E. C. Catalan in 1844, had stood unresolved for over a century. The result is known as Mihăilescu's Theorem. He is currently a professor at the Institute of Mathematics of the University of Göttingen.

preda@uni-math.gwdg.de

Daniel Muzzulini studied mathematics, musicology, physics and philosophy at Zurich University, where he received his PhD with "Genealogie der Klangfarbe" (Peter Lang 2006). He had been assistant at ETH Zürich and research associate at Zurich University, and after studying informatics at the University of Applied Sciences (now FHNW) in Basel, he worked as a software developer until 2002. Afterwards, he was a teacher of mathematics and informatics at Alpenquai College in Lucerne until 2019. Since 2015 he has been the manager of the project "Sound Colour Space – A Virtual Museum" at Zurich University of the Arts (ZHdK). His main topics of research are the history of (mathematical approaches to) music theory and perception psychology as well as the history and theory of scientific diagrams.

daniel.muzzulini@zhdk.ch

The Newton Project

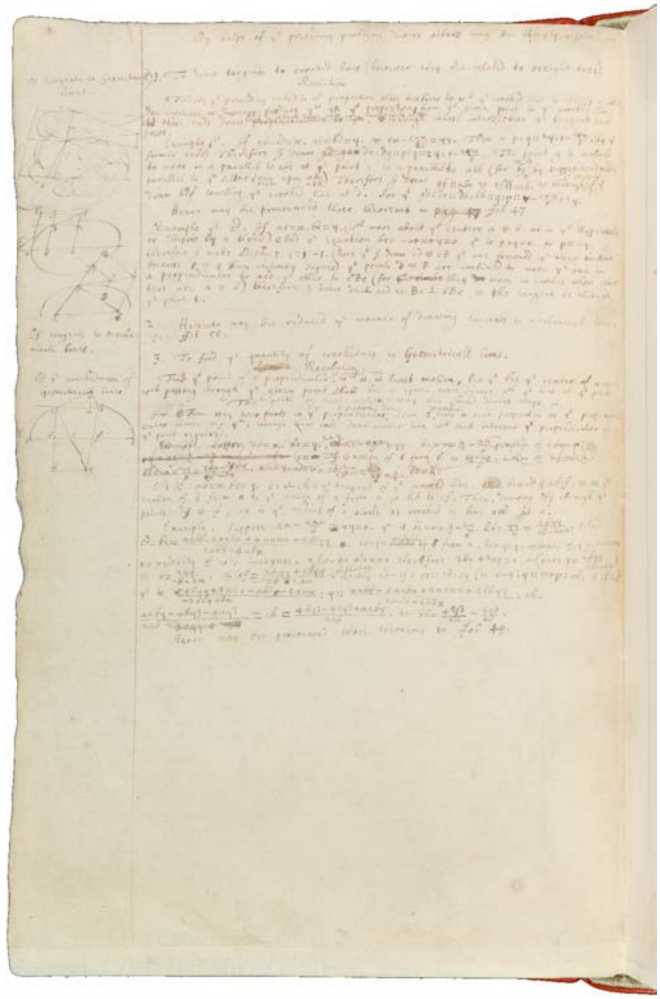
Scott Mandelbrote

In early 2022, the next phase of the Newton Project (newtonproject.ox.ac.uk) will have gone live, with a presentation of materials from the National Archives relating to Isaac Newton's stewardship as Warden and, later, Master of the Mint in the period from 1696 to his death in 1727. At its heart lies the transcription of the contents of Mint 19. This was originally the largest single lot (number 327) in the sale at Sotheby's in July 1936 of the residue of Isaac Newton's papers, which had been consigned by Viscount Lymington, who had inherited them in direct line from Newton's niece, Catherine Barton, and her husband, John Conduitt, also Newton's successor at the Mint. Lot 327 made £ 1400 and was sold to the Hungarian-American book dealer, Gabriel Wells, who played perhaps the critical role in the dispersal of items from the sale. The philanthropist, Charles Wakefield (1st Viscount Wakefield), subsequently bought Newton's Mint papers at cost from Wells on condition that he give them to the nation. Now published with accompanying material, in particular from the papers of the Treasury, Mint 19 makes available Newton's work in fulfilling the recoinage of England's silver currency, the administration of the London Mint and its various English outposts, the reorganisation of the Edinburgh Mint, the prosecution of clippers and coiners, the design of medals, the consideration of the value of coin and the provision of money for the colonies or for the payment of the army in the Low Countries. It shows him at work in what became his primary responsibility, both at a theoretical level (for example, in calculating the role of gold currency in the money supply) and at a practical one (for example, assessing the capacity of the relatively new machinery for pressing coins in use at the Mint).

The Newton Project was founded in 1998 by Rob Iliffe and Scott Mandelbrote, at the prompting of Harvey Shoolman, with the intention of producing an edition of those parts of the manuscript remains of Isaac Newton (1642–1727) that had so far remained unedited. Between 1959 and 1984, sixteen volumes of Newton's correspondence, mathematical papers, and unpublished optical lectures had been printed by Cambridge University Press, under the editorship (respectively) of H. W. Turnbull, J. F. Scott, A. Rupert Hall and Laura Tilling, D. T. Whiteside, and Alan E. Shapiro. Last year, one further magisterial volume, edited by Shapiro, and considering the preliminary work for the published *Opticks* appeared. With the partial exception of the *Correspondence*, which necessarily cast its net much more widely, most of the material edited in these books derived from the holdings of Cambridge University Library, where items deposited by Newton himself had been vastly augmented by the gift in 1872 from his heirs of the Portsmouth Collection, consisting of almost all his surviving mathematical and scientific papers and manuscripts. Whiteside, in particular, had been able to supplement material in the Portsmouth Collection with information drawn from manuscripts that had left Newton's possession in his lifetime and which had passed through the hands of William Jones (1675–1749) into the possession of his pupil, the Earl of Macclesfield, at Shirburn Castle. By great good fortune (notably the intervention of the Heritage Lottery Fund and the Andrew W. Mellon Foundation), Cambridge University Library was able to acquire those papers, including a long run of Newton's correspondence, in 2000. This accession largely brought to an end the peregrination of Newton's papers, which started as early as the 1670s, but which really took off after 1936, with the scattering, initially to booksellers and private collectors, and later (largely through donation) to libraries across the world of imperfectly described fragments from Newton's manuscript writings. A few items from the 1936 sale remain publicly unaccounted for. Most of these were among the Newton papers that resurfaced at Sotheby's in New York in December 2004 from the purchases of Emmanuel Fabius. In general, the surviving folios of Newton's writing that remain in private hands or in the trade derive from that sale. Several have been recycled through the auction rooms multiple times, including, most recently, from the discredited Aristophil collection built up by Gérard Lhéritier.

One aim of the Newton Project, which has largely been accomplished, was to provide an up-to-date record of the dispersal of Newton's papers consequent on the 1936 sale. This was achieved through the cataloguing work of Rob Iliffe and of John Young, who also led the first phases of the transcription work undertaken by the project. They built on the efforts of Peter Murray Jones, who had superintended the microfilm edition of most of Newton's manuscripts issued by Chadwyck-Healey in 1991. Several important changes of ownership, in addition to those noted above, have occurred since 1991: above all, perhaps, the transfer of the Babson collections, formed largely after 1936, to the library of Bern Dibner, and its subsequent peregrination via Cambridge, Massachusetts, to the care of the Huntington Library in San Marino, California. For the last twenty years, the bulk of the work of ordering and transcribing Newton's unpublished materials has been done online by the team of transcribers working for the Newton Project. The decision to present materials in an electronic format was taken from the start and has necessarily altered the type of editorial work involved.

The Newton Project, initially funded in 1999 by the Arts and Humanities Research Board, and since supported by a wide variety of public and private donors (including the Arts and Humanities Research Council, the Joint Information Systems Committee, the Royal Society, and the Winton Foundation), is an evolution of one of the earliest born-digital research endeavours. The fragmentation of the sources on which it worked (initially in the field of Newton's theological writings), together with their sheer bulk, quickly indicated that a print edition of the kind undertaken by Whiteside or Shapiro would be impossible. For all his tremendous skill as an editor, Whiteside had, moreover, taken decisions about the date of composition of materials and their relationship to one another that, faced with multiple and disordered drafts, the directors of the Newton Project felt it unwise as well as impossible to emulate. The Newton Project has now existed for slightly longer than the period that Whiteside required for editing *The Mathematical Papers*. In all other respects, the two works are incomparable. Whiteside recreated Newton's progress and achievement as a mathematician, working over time to particular themes that the editor identified. The Newton Project has not reconstructed Newton's processes of thinking or discovery but has instead presented the range of surviving evidence for his activities across many fields in a form that assists scholars or ordinary readers to make their own judgements. Presenting both a diplomatic and a normalised transcription of documents, the Newton Project allows readers to search Newton's previously unpublished writings by words and phrases, as well as to order his writings by their current location or (less accurately) by date. It allows one quickly to trace Newton's activity across a vast corpus of materials scattered throughout the world and is thus the starting point for future scholarly endeavour rather than an end point in itself. At its heart is the coding (based on TEI guidelines, version P5) provided by the transcribers and taggers (notably John Young, Cesare Pastorino, Yvonne Martin-Portugues Santacreu, Linda Cross, Raquel Delgado-Moreira, Will Scott and Kees-Jan Schilt) who have worked on the project and ultimately superintended by Michael Hawkins, who has organised its technical side. This is and always has been work that has involved large-scale human input to produce materials of adequate accuracy and consistency to enable effective searches. It has been accomplished by editorial effort as well as by technological innovation and requires constant intervention to keep it up-to-date and fully functional. Remarkably, it has been possible to maintain the Newton Project as a freely accessible site to any user throughout its life.



Page 57v of Newton's Waste Book (MS Add. 4004), from Cambridge Digital Library cudl.lib.cam.ac.uk/view/MS-ADD-04004/118. Reproduced by kind permission of the Syndics of Cambridge University Library.

As the possibilities of digital editions have grown, so has the Newton Project. From the start, the intention was to provide readers with translations (many by Michael Silverthorne) as well as transcriptions, and this has proved more straightforward than the original intention also to annotate documents. The growth of digital photography and especially the development of IIIF imaging that allows the viewer to explore at high resolution all aspects of the original document has transformed the possibilities for accompanying transcription with images. In its collaborations with both the Cambridge Digital Library (cudl.lib.cam.ac.uk/collections/newton/1) and the National Library of Israel (nli.org.il/en/discover/humanities/newton-manuscripts), the Newton Project has thus been able to provide searchable transcriptions to run side-by-side with document images that enable the online researcher to see more (in many cases) and read faster than the reader in the search room could do.

This breakthrough has occurred at precisely the moment when the rising value of scientific manuscripts and increasing concern for the fragility of Newtonian materials have meant that access to the originals has become more restricted. Online media have also allowed the growth of the Newton Project, alongside the Cambridge Digital Library, to include interpretative essays on materials and presentations in the format of sound and video recordings of scholarly assessments of Newton's writings. These supplement the library of reference materials about Newton that have been incorporated into the Newton Project's own website to create a wide-ranging and diverse presentation of Newton's writings accessible to many different kinds of user. In the Newton Project's current work on the Mint Papers, this will include the possibility of following Newton's movements, documented in the National Archives, on a contemporary map of London, as well as a series of interpretative essays (many of them provided by the current postdoctoral researcher on the Newton Project, Alice Marples). The result has been the development of something closer to a digital research environment for Newton, rather than the straightforward fulfilment of the original, more limited, idea of a digital edition.

The range of materials covered by the Newton Project has expanded from its original focus on theological writings to include most of Newton's correspondence, much of his scientific writing (including drafts for the *Principia* and the *Opticks*), and now his administrative papers. In preparing transcriptions, work has been assisted by the contribution of Stephen Snobelen and the Newton Project Canada (including Liz Smith and Niki Black). New collaborations are being taken forward actively, for example with the team at the Bodmerlab in Geneva who are digitising and studying the large collection of Newton's drafts erroneously known by the title "Of the Church" and owned by the Fondation Martin Bodmer. The most significant long-standing collaboration, however, has been with the Chymistry of Isaac Newton Project (webapp1.dlib.indiana.edu/newton/), led by William R. Newman and funded by the National Science Foundation and the National Endowment for the Humanities at the University of Indiana since 2005. Newman and his team (including Wallace Hooper, John Walsh, and Will Cowan) have similarly pioneered multi-media presentation of images, transcriptions, recreations of experiments, and scholarly commentary as part of their development of a digital edition of Newton's chymical and alchemical manuscripts, also accessible through the Newton Project. As is the case with some of the materials presented through the Cambridge Digital Library, the Chymistry of Isaac Newton Project has sought to appeal to teachers and school projects as well as to professional academics in its interpretation of Newton's manuscripts.

The possibilities available to the Newton Project have been heavily limited by the availability of funding for what remains a complex and resource-intensive form of scholarship, intended to make materials fully accessible to many different kinds of online user. Commitments to funders have affected when material has been made available and which materials have been prioritised. Funding constraints have also contributed to the choice of source materials: originally microfilm copies checked (where possible) against the originals; more recently greyscale or even high-quality colour images provided by the libraries and archives with which the Newton Project has collaborated. Some of these factors explain a small level of duplication in the materials offered across the Newton Project and the sites of its collaborators. They also contribute to the ongoing problems of inaccuracy, particularly at the level of transcription, which can now be very easy for a user (with access to IIIF images) to spot, but which remain time-consuming to correct and re-encode. The difficulty of proof-reading in a digital edition that is now closing in on ten million words and that has involved the labour of more than fifty collaborators is considerable. The management of the Newton Project has largely been the achievement of Rob Iliffe, who has led it across successive migrations of infrastructure and personnel from Imperial College, London, to the University of Sussex, and now to the University of Oxford.

Future development of the Newton Project, which may include closer collaborations with partner libraries, will also depend on the availability of further elements of funding. They may include steps to annotate materials with reference to Newton's library and will certainly involve the improvement of the material descriptions of manuscripts and their relationship to one another. This is one outcome to be expected from another ongoing collaboration associated with the Newton Project and funded at present by the Arts and Humanities Research Council and the National Endowment for the Humanities. In collaboration with the National Archives, Cambridge Digital Humanities, the Huntington Library, and the Science History Institute, the Newton Project and the Chymistry of Isaac Newton Project are seeking to develop new guidelines for the photography of early modern manuscripts and digital techniques for the analysis of watermarks in early modern paper. This work builds on tools developed by the École des Chartes (Marc Smith) and the École des Ponts ParisTech (Mathieu Aubry) for the computer-assisted visual analysis of watermarks (filigranes.hypotheses.org). At a technical level, the intention is to use the results of transmitted and reflected light photography and multi-spectral imaging to generate clean images of watermarks previously hidden by handwriting on the sheets of manuscripts. It is then hoped to develop a computer-vision tool to match watermarks automatically in the resulting images. Applied to the corpus of Newton's manuscripts, this will allow more accurate analysis of the paper stock used by Newton and its matching to dated examples provided by his correspondence. It will form the basis for enhancing both the images associated with transcriptions through the Newton Project and for improving the material descriptions of the manuscripts themselves.

The work of the Newton Project represents therefore the development of a digital research environment that facilitates both traditional scholarship in the history of science and the traditional disciplines of curatorship that form part of manuscript librarianship. It allows for widespread access to fragile and valuable materials and for the continuing enhancement of their online description and presentation. The Newton Project provides a model for large-scale, collaborative projects in the humanities, in particular by demonstrating what is possible in a site which remains free to users. It has a record of producing skills and knowledge, as well as presenting material for analysis by current and future researchers. At the same time, the experience of the Newton Project underlines the complexity and fragility of large-scale endeavours in the digital humanities. The long-term maintenance and development of the Newton Project remain uncertain, even as its latest major intervention comes on-line.

Select bibliography of works about the Newton Project or compiled using it

[1] N. Guicciardini, Digitizing Isaac Newton. *Isis* **105**, 403–409 and 827 (2014)

[2] R. Iliffe, *Priest of nature. The religious worlds of Isaac Newton*. Oxford University Press (2017)

[3] S. Mandelbrote, *Footprints of the lion: Isaac Newton at work*. Cambridge University Library (2001)

[4] A. Marples, The Science of Money: Isaac Newton's Mastering of the Mint. *Notes and Records of the Royal Society* (forthcoming)

[5] W. R. Newman, *Newton the Alchemist*. Princeton University Press (2019)

[6] C. J. Schilt, *Isaac Newton and the Study of Chronology*. Amsterdam University Press (2021)

Scott Mandelbrote is Fellow, Director of Studies in History, and Perne and Ward Librarian at Peterhouse, one of the Colleges of the University of Cambridge. He is an editorial director of the Newton Project. Recent publications include *The Reception of Isaac Newton in Europe* (Bloomsbury, 2019), 3 vols (with Helmut Pulte).

shm24@cam.ac.uk

Sir David Cox: A wise and noble statistician (1924–2022)

Christos P. Kitsos

Sir David, David Roxbee Cox, was born on 15 July 1924 in Birmingham, United Kingdom, were he attended Handsworth Grammar School. (The aeronautical engineer Harold R. Cox was a distant cousin.) David received his Master of Arts in mathematics at St John's College, Cambridge. Referring to his time as a student in Cambridge, he very often mentioned Harold Jeffreys, and to a lesser extent J. O. Irwin. Sir Harold Jeffreys, FRS (1891–1989) was mathematician, statistician, geophysicist and astronomer. His book *Theory of Probability* (1939) discussed the objective Bayesian view on probability, which Sir David referred to even in his last seminars, see also [41]. Jeffreys was also involved in mathematical physics, the favorite subject of David Cox in his very early steps. He eventually concentrated on statistics in the early 1950s. As for Joseph Oscar Irwin (1898–1982), he was a key person in the middle third of the 20th century, linking theoretical statistics to applications in medicine, an area that Sir David respected during all his research. Moreover, Irwin was one of the very few statisticians who worked with both Pearson and Fisher and was able to maintain cordial relations with these strong personalities in the statistics world of the 20th century [40]. I certainly think that both Jeffreys and Irwin influenced the statistical line of thought of Sir David in his future important work. David Cox obtained his PhD from the University of Leeds in 1949, supervised by Prof. Henry Daniels, FRS, and Prof. Bernard Welch, a founder of the Industrial and Agricultural Research Section of the RSS. His dissertation was entitled "Theory of Fibre Motion". Below is a list of milestone dates in his career:

1944–46	Royal Aircraft Establishment;
1946–50	Wool Industries Research Association of Science and Technology;
1950–55	Assistant Lecturer in Mathematics, University of Cambridge;
1955–56	Visiting University of North Carolina, Princeton, and Berkeley;
1956–61	Reader in Statistics, Birkbeck College, London;
1961–65	Professor of Statistics, Birkbeck College, London;
1965	Member of Technical Staff, Bell Laboratories;
1966–88	Professor of Statistics, Imperial College of Science and Technology, London;
1969–73	Head of Department of Mathematics, Imperial College;
1983–88	SERC Senior Research Fellow;
1988–94	Warden, Nuffield College, Oxford;
1994–2022	Honorary Fellow of Nuffield College.

Sir David Cox at the RSS, 2016. Photo by the Royal Statistical Society.

David Cox was married with Joyce Drummond since 1947, with four children. He was knighted in 1985 and received the Copley Medal, the Royal Society's highest award, in 2010.

David Cox served as President of the Royal Statistical Society (1980–82) and the International Statistical Institute (1995–97). In this capacity I had the honor to meet him at the 51st Session of ISI in Istanbul and discuss in detail the satellite conference on Industrial Statistics we held in Athens [44]; this among other occasions of meeting him in various countries. I still remember that discussion and the comments-questions he asked when I listed the papers presented at the Athens satellite conference. Later I became aware of the contributions to industry he made during his first work steps, at the Royal Aircraft Establishment and the Wool Industries Research Association. At that time, in 1949, he published his first two papers [7] (part of his doctoral dissertation, related to industrial problems), and the discussion of quality control ideas [6]. In 1998 he visited Greece, the University of Business and Economics, Department of Statistics, where he was awarded the honorary doctor degree. A complete list of about 384 publications of Sir David Cox can be found on the internet.

David Cox was a doctoral advisor for several distinguished statisticians, among them David Hinkley (with whom he published in 1974 the book *Theoretical Statistics*), Peter McCullagh (who received the 1983 Karl Pearson Prize of the ISI), Henry Wynn (in design theory; Wynn was the first RSS president elected by a contested vote in 1977). Sir David authored a great number of pioneering works, offering an elegant statistical background and appropriate solutions to real life problems. Most of us worked with a range of his concepts and methods, including the Cox process, Cox models and the Cox's direction. Cox's 1972 survival analysis paper accounted for over 26 % of the citations to papers in *Series B* of the *Journal of the Royal Statistical Society*, something like more than 50 000 citations! He was awarded the International Prize in Statistics, recognizing him specifically for his 1972 paper [22], in which he developed the proportional hazards model that today bears his name, and which changed the way we understand and analyze risk factors.

We shall try to provide here a compact review of his work, specifically, at least of the part that has received a great number of citations and covers different fields in statistics.

1 Experiment design – regression

Following the line of thought of [37] and his pioneering work, Sir David, worked in his early research on the book *Planning Experiments* [13], one of his favorite texts. The book is devoted to all sorts of experimental design models, and although there are discussions on error reduction, it does not contain an optimal design approach, as it has been treated in [52] by S. D. Silvey, a close fellow to Sir David, or later by his student H. P. Wynn in [53]. The experiment design point of view was also discussed, among several very helpful statistical ideas for the cancer problem, in the papers [48, 49], which account for over 2000 citations. Cox came back to the design of experiments and regression in the paper [26], written for the 150th anniversary of the RSS, with a list of 22 essential points stated in an appendix, points that offered vital lines of thought for the interested researcher. I think that trying to address the point 22, concerned with "the prediction (via intervals) of future values", I came across the idea that for applications, and for "future observations" it is better to adopt tolerance intervals rather than confidence intervals, see [45, 46]. Adopting regression and working on the general definition of residuals, Cox and E. J. Snell [33] came across an application of their method to a nonlinear model for leukemia data, where crude and modified residuals are evaluated. On the subject of regression, the paper [19] offers very nice, in my opinion, "miscellaneous and isolated comments". The paper [34] can be considered as a continuation of the existing common work, treating the problem of variable selection. Therein a series of criteria are mentioned, especially Mallow's C_p statistic, recommending the general points a researcher should follow. The medical line of thought, for the practical applications, is also present in this paper: the relation between time to death y and the level of some prescribed "dose" x is considered, and the possible analyses and classification of variables such as age and sex are discussed. The problem of selection of variables in linear regression was essential at that time (see, e.g., Hocking [39]), and appropriate "routines" were discussed in [47]. Working with mixtures of experiments, Cox presented new such models in [21]. In principle, the centroid of a constrained region is the reference mixture, while the effect of the i-th component is measured along a line connecting this centroid to the corresponding vertex $x_i = 1$; it is the direction of this line that is known as Cox's direction. I believe that all this demonstrates an essential characteristic of D. R. Cox's scientific life: he was present, during all his active years, with his own contributions to various problems, at the right time, offering new ideas and clarifying existing ones. Sir David returned to the experiment design theory in the book [31], this time with a modern notation, discussing recent methods (in Chapter 8), nonlinear design, and optimal designs (Chapter 7). Although the spirit of [13] was preserved, the presentation of the work is different, with the addition of the new ideas that emerged since then.

2 Survival analysis – binary data

The sigmoid curve $p(z) = (\exp(z))/(1 + \exp(z))$ is known as "logistic curve", due to Adolphe Quetelet's student Pierre Francois Verhulst (1804–1849). It was J. Berkson who devoted his statistical work to "logit models" [3], according to Bliss pioneering work [4] on "probit models", and then later D. Finney coined the term bioassay [36].

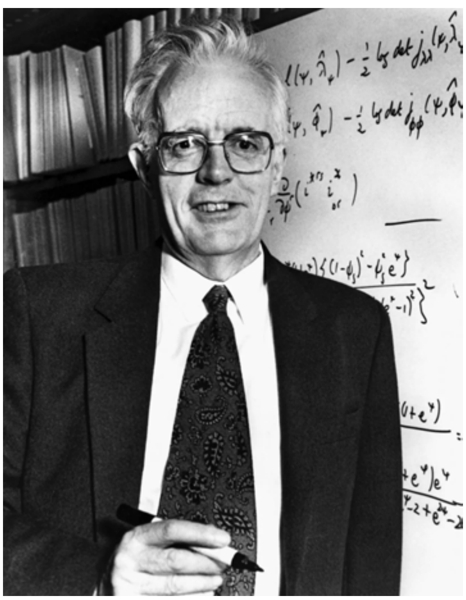
The binary response problem was extensively discussed by Cox [14]. Later, in [20], he cemented the theory of the binary response problem, so useful in biostatistics and crucial in data analysis. In this way a systematic and strong framework was constructed for binary data, analogous to the least squares method and extending the probit analysis. The "covariate paper" [51] is concerned with the existence of the MLE working with binary response problems on the Analysis of Binary Data [43]. Some years later an improved version of the 1969 book was published [35]. As computing technology was changing rapidly, binary analysis became increasingly popular. Going from one variable to two, the problem can be simply described as follows.

Let S_1 and S_2 be two dependent Bernoulli variables. Let x be a covariate associated to the distribution of S_1 and S_2. In [23] Sir David works within the framework of the analysis of multivariate binary data, adopts logistic models (see [23, Table 2]), and views as a special case the joint distribution of (S_1, S_2). The possible outcomes are $(1,1)$, $(0,1)$, $(1,0)$, $(0,0)$. Eventually the bivariate distribution of (S_1, S_2) can be expressed as products of (ordinary) logistic functions and thus the likelihood function and the information matrix can be evaluated (see also [35]).

The year 1972 is crucial for David Cox's research. He presented important results in [22, 23], which changed the way of thinking on what a risk factor is in survival analysis. This paved the way for powerful scientific research and discoveries to take place, which had a lasting impact on human health worldwide. He introduced statistics in medical applications that J. O. Irwin (see Section 1 above) could not have even imagined! His mark on research is so great that his 1972 paper is one of the three most-cited papers in statistics and ranked 16th in *Nature*'s list of the top 100 most-cited papers of all time for all fields. So indeed 1972 was a golden year for Sir David. Since then we are referring to Cox's proportional hazards model. We shall refer to his paper for the essential newly proposed relation (9) therein: the typical hazard function $h(t)$ is, in principle, specified by the assumed probability model to identify etiological agents for the risk problem under investigation. Let us suppose, at first, that two explanatory variables, x_1 and x_2, say, are of interest, and that these do not vary with time. We can assume that $h(t)$ is a linear function of x_1 and x_2, $h(t; x_1, x_2)$, say. Recall that we assume that $h(t) > 0$, and this might not be the case for the postulated linear function. If it is assumed that the function $\ln[h(t)]$ is linear, with an extra linear term to consider time, there are still problems, even if the parameters can be estimated. Not only is difficult to define how the hazard function depends on time, but if it is also assumed to be non-monotonic (does not increase or decrease with time), then it is difficult to find an appropriate explicit such function to include in the model. The Cox proportional hazards model provides the solution. It defines $h(t; x)$, with $x_j = (x_{1j}, x_{2j}, \ldots, x_{pj})$, $j = 1, 2, \ldots, n$, the vector of p covariates associated to each individual j, as

$$h(t; x) = h_0(t) \exp(xb). \tag{1}$$

Here b is a $p \times 1$ vector of unknown parameters, and $h_0(t)$ an unknown function, which provides the hazard function at $x = 0$, known as baseline function. The above relation (1) is revolutionary. In Sir David's words: "My model is used to compute the probability of anything from earthquakes to bankruptcies". Definition (1) and the related theory and computations are widely used in the analysis of survival data. They enable researchers to easily identify the risks of specific factors for mortality. Certainly the model can be applied to other "survival outcomes", as in electronics among groups of materials with disparate characteristics, or in economics when risk factors are under investigation. The whole analysis is based on the Maximum Likelihood, as all his work is "Fisherian". It is remarkable the way he treats the likelihood now, ignoring some of its terms.



Sir David Cox. Photo by National Cancer Institute.

3 Stochastic processes

Although David Cox agreed (see [50]) that there is too much in his paper [12], "Doubly stochastic Poisson process, all sorts of tests to do with empirical series, of points events …", this paper is certainly his first mentioned contribution to the field of stochastic processes. Most of these ideas were present in his doctoral dissertation, while his interest in queues (see [32]) originated from his work in the textile industry. Today some people are regarding queuing theory as a branch of operational research, but nobody denies that it is inextricably linked to the stochastic processes, including the adoption of Kendall's notation, in his excellent work [42].

The realistic line of thought rather than the technicalities is clear in the two papers [10, 12], published in the same proceedings volume, where it is shown how a non-Markov process can be built into a Markov process. David Cox remained faithful to "the spirit of Bartlett's great masterpiece [2], which is a difficult read, but not because of an overelaborate mathematical formalism" [50]. The covariance counting problem in physics was successfully tackled as a stochastic process [28]. We recall the pioneering work of Maurice Bartlett (1910–2002), devoted mainly to the analysis of data with spatial and temporal patterns, also known from Bartlett's method in analysis of time series. In his book on stochastic processes [2], Bartlett summarizes all his work on the subject, and Sir David is referring to it as a "masterpiece". Bartlett sometimes criticized Fisher, but he was a pioneer in the field (see also [1]). Sir David expressed in [50] the opinion that somebody might study stochastic processes without a heavy mathematical background or by adopting an overelaborate mathematical formalism, even for renewal theory [17, 30].

I feel that there are times when mathematical technicalities are not helpful at all to the experimentalists, and so Cox's line of thought is well accepted. Still there are cases, like the theories of stochastic birth-death processes, where technicalities can be useful for modeling processes of carcinogenesis. Moreover, since the pioneering seminar of Karl Pearson in 1896 on "Regression, Heredity and Banmixia", linear algebra became an important tool in statistics. Then a new chance was offered for more mathematical methods to enter statistical theory, and some indeed proved useful. In principle, I believe, it takes time for a mathematically oriented idea to be absorbed in practical problems, if adopted.

4 The separate families problem

A very interesting problem, known as the "separate families of hypotheses", was introduced in [15, 16]. Cox then returned to this problem later, in [27]. A compact formulation of the problem reads as follows: Let X_i, $i = 1, 2, \dots, n$, be independent identically distributed (i.i.d) random variables from a population with density function f. Let θ be a parameter with values in a parameter space Θ, and ξ be parameter with values in a different parameter space Ξ. Consider the distribution functions $g = g(x; \theta)$ and $h = h(x; \xi)$ associated with the parameter spaces Θ and Ξ, respectively, as well as the resulting families of distributions $G = \{g = g(x; \theta),\ \theta \in \Theta\}$ and $H = \{h = h(x; \xi),\ \xi \in \Xi\}$. It is assumed that all the distribution functions are associated with the same baseline measure. The problem is to test, under smoothness conditions on g and h,

$$H_0\colon f \in G \quad \text{vs} \quad H_1\colon f \in H.$$

The method is applicable for the "one-hit" or the "two-hit" models in the binary response theory [20], so essential for statistics problems concerned with cancer. The paper [16] considered the difference of the log-likelihoods for g and h, denoted $D_l = l(g; \mathrm{est}(\theta)) - l(h, \mathrm{est}(\xi))$, with $\mathrm{est}(d)$ being the estimate of the parameter d, and the expected value $E(D_l)$ of D_l with respect to $g(x; \mathrm{est}(\theta))$, say. Thus, the paper worked with the test statistic $T = D_l - E(D_l)$. It was really a very interesting line of thought, based on fundamental statistical principles.

5 Other fields

The concepts of marginal and conditional likelihood were clarified by David Cox in the paper [24], where he also treated the likelihood of the hazards proportional model and proved that it fits the partial likelihood definition he proposed. He also worked on the concept of likelihood in [25], where he proved that the maximum likelihood estimation of a simple model retains high efficiency in the presence of modest amounts of overdispersion.

Dealing with sequential likelihood ratio tests, David Cox proposed in [18], under mild assumptions, an easy-to-handle method based on an approximation providing numerical evaluations. Moreover, in [9], he devised a unified method under which sequential tests can be obtained for composite hypotheses. Therein he considered the problem of discriminating between the hypotheses H_0 and H_1 concerning two different Bernoulli trials. There are several papers based on [9]; I think [38] tried to offer a mathematical justification for this excellent paper on sequential analysis, where the calculations, eventually, establish the validity of the theoretical considerations for the main argument of sequential analysis: there is a gain in the sampling units, providing a discussion for sufficiency and invariance.

An interesting contribution to sampling is the two-stage sampling [8], which provided food for thought for a two-stage optimal experimental design [43], while the sequential procedures offered a solution to the estimation problem for the nonlinear optimal experimental designs. His contribution to time series is not reduced to the paper [11], very rich in ideas; we should mention also [29], where the trend is investigated, and a smooth function of the time t of the form $a[\exp(bt)]$ or $a(t + t_0)^b$, where t_0 is known and a, b are unknown parameters, is introduced. Importantly, in [11] a point process which is a generalization of a Poisson process, also known as a Cox process, was introduced. It is interesting that David Cox uses the same notation, lambda, for this function, as in the proportional hazards model (see [22]).

There is no textbook on regression analysis that does not refer to the Box and Cox transformations, and to the masterpiece source [5] for teaching all sorts of statistics subjects to graduate students. Both J. Tukey and M. S. Bartlett, in a discussion of [5], stated: "the authors have made a major step forward". It is indeed a marvelous contribution, widely adopted, especially in applications.

We tried to survey briefly a small, but – we believe – representative part of David Cox's extended scientific research, in almost all fields of statistics. One should emphasize that all his papers (despite including often in the title the word "notes"), are rich in new, pioneering ideas, and always provide helpful examples.

6 Discussion

Sir David was particularly known for adopting a pragmatic rather than a dogmatic perspective on the Bayesian/frequentist controversy and described this position at his very interesting RSS seminars and accompanying videos. He was also referring to "foundation" with the well-known comment of Fisher, about "building a basement". His line of thought was clearly referring to "theoretical statistics" and not to "mathematical statistics"; needless to say, he was faithful to this line of thought until the end. Model adequacy was crucial to him, though probably he did not persuade

Bronze portrait of Sir David, by the sculptor Martin Jennings, in the Senior Common Room of the Nuffield College, Oxford. With the kind permission of Martin Jennings and the Nuffield College.

everybody working in the field of medical statistics. He received many honors: the Guy Medal in Gold of the RSS, the inaugural International Prize for Statistics, and the Copley Medal of the RSS (as Carl Friedrich Gauss once did!). Sir David will be remembered as an incredibly generous and supportive friend. I had the honor to receive his friendly comments and advice in many discussions, and especially at the ISI Session in Istanbul while discussing industrial statistics and a cancer problem. Only a small sample of his over 350 papers are mentioned here. He was the editor of the journal *Biometrika* for an extraordinary span of 25 years, from 1966 to 1991, and was a co-editor with Professor D. M. Titterington, head of the Department of Statistics of Glasgow University in the 1980s (and my supervisor in Glasgow!), of a volume dedicated to the centennial anniversary of Biometrika. He served terms as President of the Royal Statistical Society (1980–82) and the International Statistical Institute (1995–97). In his words, "people say theoretical work in statistics should be motivated by applications because it's a practical subject" [50]. That is in accordance with his good relationship with John Tukey, during his visit to USA, but mainly provides evidence for the general line of thought David Cox followed, often stressing how hard it was for him to get to grips with ideas and to solve the impressive, for us, problems that he formulated in his pioneering work in statistics. He did not hesitate to work on the improvement of his own books: he returned to and with D. Reid [31] revised the experiment design book [31], and with E. Snell revised the *Analysis of Binary Data* [35].

As Professor F. Downton stated in his discussion for the [22] paper: "Professor Cox has been too modest", and he lived in modesty all his productive life, one could add.

Acknowledgements. I would like to thank Alex Rigas for helpful discussions during the preparation of this paper. I am also grateful to Steven Roy and Fernando Pestana da Costa for their help.

References

[1] M. S. Bartlett, Some evolutionary stochastic processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **11**, 211–229 (1949)

[2] M. S. Bartlett, *An introduction to stochastic processes*. Cambridge University Press, London (1955)

[3] J. Berkson, Why I prefer logits to probits. *Biometrics* **7**, 327–339 (1951)

[4] C. L. Bliss, The method of probits. *Science* **79**, 38–39, 409–410 (1934)

[5] G. E. P. Box and D. R. Cox, An analysis of transformations (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **26**, 211–252 (1964)

[6] A. Brearley and D. R. Cox, *An Outline of statistical methods for use in the textile industry*. Wool Industries Research Association, Leeds (1949)

[7] D. R. Cox, The theory of drafting wool silvers. *Proc. Roy. Soc. London Ser. A* **197**, 28–51 (1949)

[8] D. R. Cox, Estimation by double sampling. *Biometrika* **39**, 217–227 (1952)

[9] D. R. Cox, Sequential tests for composite hypotheses. *Proc. Cambridge Philos. Soc.* **48**, 290–299 (1952)

[10] D. R. Cox, The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc. Cambridge Philos. Soc.* **51**, 433–441 (1955)

[11] D. R. Cox, Some statistical methods connected with series of events (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **17**, 129–164 (1955)

[12] D. R. Cox, A use of complex probabilities in the theory of stochastic processes. *Proc. Cambridge Philos. Soc.* **51**, 313–319 (1955)

[13] D. R. Cox, *Planning of experiments*. A Wiley Publication in Applied Statistics, John Wiley & Sons Inc., New York (1958)

[14] D. R. Cox, The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **20**, 215–242 (1958)

[15] D. R. Cox, Tests of separate families of hypotheses. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, Univ. California Press, Berkeley, 105–123 (1961)

[16] D. R. Cox, Further results on tests of separate families of hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **24**, 406–424 (1962)

[17] D. R. Cox, *Renewal theory*. Methuen & Co. Ltd., London (1962)

[18] D. R. Cox, Large sample sequential tests for composite hypotheses. *Sankhyā Ser. A* **25**, 5–12 (1963)

[19] D. R. Cox, Notes on aspects of regression analysis (with discussion). *J. Roy. Statist. Soc. Ser. A* **131**, 265–279 (1968)

[20] D. R. Cox, *Analysis of binary data*. Chapman & Hall, London (1969)

[21] D. R. Cox, A Note on polynomial response function for mixtures. *Biometrika* **58**, 155–159 (1971)

[22] D. R. Cox, Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34**, 187–220 (1972)

[23] D. R. Cox, The analysis of multivariate binary data. *Appl. Stat.* **21**, 113–120 (1972)

[24] D. R. Cox, Partial likelihood. *Biometrika* **62**, 269–276 (1975)

[25] D. R. Cox, Some remarks on overdispersion. *Biometrika* **70**, 269–274 (1983)

[26] D. R. Cox, Present position and developments: Some personal views. Design of experiments and regression. *J. Roy. Statist. Soc. Ser. A* **147**, 306–315 (1984)

[27] D. R. Cox, A return to an old paper: "Tests of separate families of hypotheses". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**, 207–215 (2013)

[28] D. R. Cox and V. Isham, A bivariate point process connected with electronic counters. *Proc. Roy. Soc. London Ser. A* **356**, 149–160 (1977)

[29] D. R. Cox and P. A. W. Lewis, *The statistical analysis of series of events*. Methuen & Co. Ltd., London (1966)

[30] D. R. Cox and H. D. Miller, *The theory of stochastic processes*. Methuen & Co. Ltd., London (1965)

[31] D. R. Cox and N. Reid, *The theory of the design of experiments.* Monographs on Statistics and Applied Probability 86, Chapman & Hall, Boca Raton (2000)

[32] D. R. Cox and W. L. Smith, *Queues*. Chapman & Hall, London (1991)

[33] D. R. Cox and E. J. Snell, A general definition of residuals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **30**, 248–275 (1968)

[34] D. R. Cox and E. J. Snell, The choice of variables in observational studies. *Appl. Stat.* **23**, 51–59 (1974)

[35] D. R. Cox and E. J. Snell, *Analysis of binary data*. Second ed., Monographs on Statistics and Applied Probability 32, Chapman & Hall, London (1989)

[36] D. J. Finney, *Probit analysis*. Third ed., Cambridge University Press, London (1971)

[37] R. A. Fisher, *Design of experiments*. Oliver and Boyd, Edinburgh (1935)

[38] W. J. Hall, R. A. Wijsman and J. K. Ghosh, The relationship between sufficiency and invariance with applications in sequential analysis. *Ann. Math. Statist.* **36**, 575–614 (1965)

[39] R. R. Hocking, The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–49 (1976)

[40] J. O. Irwin, The place of mathematics in medical and biological statistics. *J. Roy. Statist. Soc. Ser. A* **126**, 1–44 (1963)

[41] H. Jeffreys, On the relation between direct and inverse methods in statistics. *Proc. Roy. Soc. London Ser. A* **160**, 325–348 (1937)

[42] D. G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Statistics* **24**, 338–354 (1953)

[43] C. P. Kitsos, *Design and inference in nonlinear problems*. Unpublished PhD thesis, Glasgow University (1986)

[44] C. P. Kitsos and L. Edler, *Industrial statistics.* Physica-Verlag (1997)

[45] C. P. Kitsos and T. L. Toulias, Confidence and tolerance regions for the signal process. *Recent patterns on signal processing* **2**, 149–155 (2012)

[46] C. P. Kitsos and V. Zarikas, On the best predictive general linear model for data analysis: A tolerance region algorithm for prediction. *J. Appl. Science* **13**, 513–524 (2013)

[47] L. R. La Motte, The SELECT routines: A program for identifying best subject regression. *Appl. Stat.* **21**, 92–93 (1972)

[48] R. Peto, C. M. Pike, P. Armitage, E. N. Breslow, D. R. Cox, V. S. Howard, N. Mantel, K. McPherson, J. Peto and C. P. Smith, Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part I. *Br. J. Cancer* **34**, 585–612 (1976)

[49] R. Peto, C. M. Pike, P. Armitage, E. N. Breslow, D. R. Cox, V. S. Howard, N. Mantel, K. McPherson, J. Peto and C. P. Smith, Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II. *Br. J. Cancer* **35**, 1–39 (1977)

[50] N. Reid, A conversation with Sir David Cox. *Statist. Sci.* **9**, 439–455 (1994)

[51] M. J. Silvapulle, On the existence of maximum likelihood estimators for the binomial response models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **43**, 310–313 (1981)

[52] S. D. Silvey, *Optimal design*. Monographs on Applied Probability and Statistics, Chapman & Hall, London (1980)

[53] H. P. Wynn, Results in the theory and construction of *D*-optimum experimental designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34**, 133–147, 170–186 (1972)

Christos P. Kitsos obtained his PhD from the Department of Statistics at the University of Glasgow in 1986. From 1991 to 2008 he participated in a CCMS/NATO Pilot Study on experimental carcinogenesis. He received an *Exzellenzstipendium des Landes Oberösterreich* in 2015 at the Johannes Kepler Universität Linz. Since 2018 he is emeritus professor of statistics at the Technological Educational Institute of Athens, University of West Attica. Moreover, he is a part-time full professor in the *Doutoramento em Matemática Aplicada e Modelação* at the Universidade Aberta in Lisbon. He is co-editor of a number of volumes in risk analysis.

xkitsos@uniwa.gr

Mathematics at the Institute of Science and Technology Austria (ISTA)

Florian Schlederer

What is mathematics, if not the most practical art? A precise formalism for abstract thinking and the most powerful language to gain insight into nature's phenomena. Oftentimes the most playful scholars gravitate to its immaterial beauty. It serves as a rigorous backbone supporting and connecting all kinds of scientific disciplines. Foundational mathematical research at the Institute of Science and Technology Austria (ISTA) embraces exactly these qualities. It is defined by a strictly curiosity-driven approach, the multidisciplinarity in the Institute's DNA, and the scientific excellence of its international faculty. From discrete/computational topology over number theory to mathematical physics and analysis, eleven groups make up a selected, successful, and fast-growing faculty.

Lab Buildings East (left) and West (right). The Institute of Science and Technology Austria (ISTA) offers state-of-the-art research facilities and scientific support units in all major disciplines. © Peter Rigaud/ISTA

Not long ago, Klosterneuburg near Vienna was only known for its magnificent 900-year-old monastery and fertile vineyards on the hills. Since 2009, it has become the home of a brand-new scientific institute, the Institute of Science and Technology Austria (ISTA). This bold initiative from the Austrian government aims to foster research and graduate education in the natural sciences. While so young and with 68 research groups comparably small, ISTA already ranked third in the normalized Nature Index amongst the world's best research institutions. How did an institute located at the edge of the tranquil Vienna Woods achieve this distinction?

The formula for world-class research

ISTA was founded based on a set of principles distilled from the most successful governance systems and ideas in the world of research institutions. In fact, Professor Haim Harari, former president of the Weizmann Institute of Science, played a pivotal role in applying proven methods to the founding of ISTA. First President of ISTA is Professor Thomas A. Henzinger, a leading computer scientist with roots at the University of California in Berkeley, USA, and the EPFL in Lausanne, Switzerland. He shaped the Institute into what has become a model of academic institution to emulate.

All researchers at ISTA pursue their interests without constraints or predefined research topics. The Institute does not allocate a predetermined number of positions to different disciplines; excellence and promising research are the exclusive hiring criteria for all scientists at ISTA. Groups are supported by state-of-the-art infrastructure and scientific service units. Professors are mostly hired early in their careers on a tenure track system, usually between three to eight years after their PhD. This system provides them scientific independence and a career perspective. Since their basic research may lead to unforeseen but useful discoveries and applications, intellectual property and technology transfer are also important objectives of the Institute. The technology park adjacent to the Institute, IST Park, has developed rapidly since opening in 2019 and is already home to ten companies, including three start-ups that are based partly on Institute research.

To foster a creative atmosphere at ISTA, separating organizational structures such as faculties and scientific departments are avoided. Also, the Institute brings together researchers from all major scientific disciplines. Their communication and collaboration across fields is actively encouraged, though not mandatory – the fact that some principal areas of mathematics do not have a direct interdisciplinary component is fully recognized.

Another core mission of ISTA is science education. The Institute sees itself as a place of inquiry and reason that translates scientific practice and its results for the general public via a diverse science outreach program. Furthermore, it is a PhD-granting institution that offers fully funded PhD positions to highly qualified candidates with a bachelor's or master's degree in the natural sciences. Training the next generation of scientific leaders is done on graduate level, which favors advanced and research-related teaching.

ISTA's mathematics groups

- Nick Barton: Mathematical Models of Evolution
- Tim Browning: Analytic Number Theory and its Interfaces
- Herbert Edelsbrunner: Algorithms, Computational Geometry, and Computational Topology
- László Erdős: Mathematics of Disordered Quantum Systems and Matrices
- Julian Fischer: Theory of Partial Differential Equations, Applied and Numerical Analysis
- Tamás Hausel: Geometry and its Interfaces
- Vadim Kaloshin: Dynamical Systems, Celestial Mechanics, and Spectral Rigidity
- Matthew Kwan: Combinatorics and Probability
- Jan Maas: Stochastic Analysis
- Robert Seiringer: Mathematical Physics
- Uli Wagner: Discrete and Computational Geometry and Topology

A growing community

From the beginning, ISTA has recruited around five professors per year. From more than 15,000 applications, 68 professors are currently under contract. In the fall of 2021, the government committed another 3.3 billion Euro to secure further growth of the Institute for the next 15 years. By 2026, ISTA will reach 90 research groups, and then up to 150 groups by 2036. This growth trajectory enables it to transcend its status as a small institute and to become a substantial European hub in the global network of frontier science. New laboratory buildings, research facilities, and a visitor center are under construction and further ones are in planning to meet the needs of a growing and diversifying community.

The clear decision towards growth is based on the scientists' achievements since the founding of the Institute. This encompasses more than 3,600 scientific publications, numerous memberships in prestigious academies and societies, and a list of renowned awards

Leading a world-class research institution. The management of ISTA consists of (from left to right) Vice President for Science Education Gaia Novarino, Managing Director Georg Schneider, President Thomas A. Henzinger, Executive Vice President Michael Sixt, Dean of the Graduate School Eva Benková, and Vice President for Technology Transfer Bernd Bickel. © Peter Rigaud/ISTA for example in mathematics: the Leonard Eisenbud Prize from the American Mathematical Society and the Erwin Schrödinger Prize from the Austrian Academy of Sciences to László Erdős, the Promotion Prize of the Austrian Mathematical Society to Julian Fischer, and twice the Ferran Sunyer i Balaguer Prize to Tim Browning.

So far 54 European Research Council (ERC) grants have been awarded to 45 professors. In mathematics, ten ERC grants for eleven research groups constitute a considerable portion of the total grant sum of 22.7 million Euro for mathematicians at ISTA. Among the graduates and alumni from mathematical groups, there are personalities like Phan Thành Nam, professor of analysis and mathematical physics at the Ludwig Maximilian University (LMU) in Munich, Germany, and Tanya Kaushal Srivastava, assistant professor of algebraic geometry at the Indian Institute of Technology (IIT) Gandhinagar.

PhD graduate Laura Schmid. Her research in computer-aided verification and game theory supervised by professor Krishnendu Chatterjee lies at the cross-section of mathematics, evolution, and computer science. © Peter Rigaud/ISTA

The newly opened Sunstone Building. It provides space for first-year students, hosts the Institute's library, a nuclear magnetic resonance facility, and state-of-the-art laboratories for fundamental research in chemistry and materials. Further expansion of the campus is in construction and planning. © Peter Rigaud/ISTA

Becoming a mathematician at ISTA

Two of the eleven mathematics groups began their work last year. Since hiring is done on an institute-wide level, attracting many excellent candidates in mathematics is key to maintaining its presence at ISTA. While the hiring concept is open to every direction in mathematics, ISTA strives for a healthy balance.

"Looking at the mathematics groups here, I can roughly identify three areas," says Robert Seiringer, professor and area chair for mathematics and physics. "Five groups are working at the cross-section of mathematical physics and analysis, three groups are concerned with combinatorics and discrete/computational topology, and two groups work on algebra and number theory." He hopes to hire colleagues in other key fields such as geometry and statistics. The envisioned development plan lays out a mosaic of mathematical expertise, where every new group – much like a tessera – feels connected to at least one existing group without duplicating it adding to a full display of cutting-edge mathematics.

Out of the present 534 scientists 280 are PhD students, 29 of which are currently working in mathematics groups. The graduate school follows the US admission pattern: ISTA admits students once a year, with a mid-January application deadline. Students are provided with a generous fellowship for the whole duration of their studies, assuming satisfactory progress. Their first year is defined by three to five rotations, where students experience different groups in various fields until they choose an affiliation for their PhD.

Postdoctoral programs are centrally financed by the ISTA Fellowship; applications are solicited with deadlines of March and September. Once again, candidates compete across disciplines for admissions. Other possibilities for fully funded, two-year post-

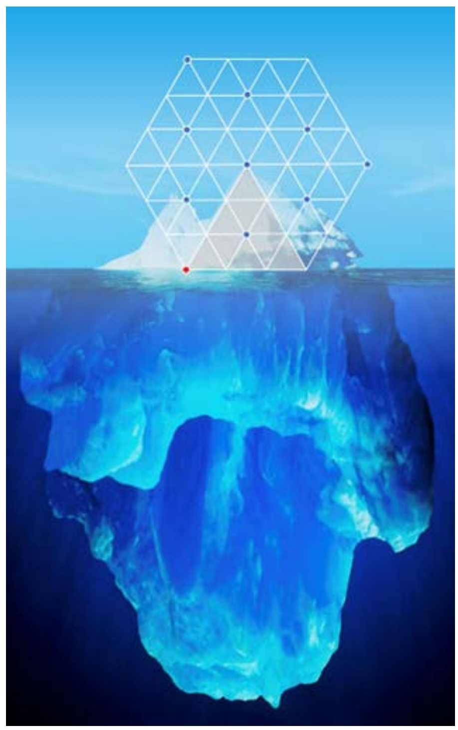
The historic central building of ISTA. A place of curiosity-driven inquiry and excellent science at the borders of the peaceful Vienna Woods. © ISTA

doctoral positions are the twelve IST-BRIDGE fellowships every six months until November 2023; or the two annual NOMIS fellowships, backed by the NOMIS Foundation to advance interdisciplinary research. Brazilian mathematician Gonçalo-Oliveira, for instance, embodies the outstanding talent which ISTA and the NOMIS Foundation strive to inspire. He works at the intersection of Tamás Hausel's and László Erdős' groups.

For long-term visitors, there is a program with partial financing for periods of three to twelve months. Finally, ISTA offers an internship program – the ISTernship – for talented and enthusiastic undergraduate students, who would like to spend eight to twelve weeks at the Institute during the summer. A more detailed description of these opportunities may be found on the ISTA website www.ista.ac.at.

Recent research highlights in mathematics

Three recent highlights give insight into the current mathematical activities at ISTA, each exemplifying one strand of current research: algebraic geometry, mathematical physics, and number theory.



Mathematics from tip to icy toe. Professor Tamás Hausel investigates the weight diagrams in the nilpotent cone to gain knowledge about the whole underlying representation theory of Lie groups. © ISTA/Ruslan Tagiyev

The tip of the mathematical iceberg

Mathematics may strike you as less adventurous than a polar expedition, but the exquisiteness of this conquered abstract iceberg could change your mind. The pioneers of this quest, ISTA-professor Tamás Hausel and Oxford scholar Nigel Hitchin, collaborated at the intersection of differential and algebraic geometry, connecting the distant fields of physics and number theory.

The analogy with a floating iceberg shows the significance of their mathematical expedition. The iceberg is attached to a Lie group: Most of its characteristics lie hidden beneath the surface. Down there, the interesting useful properties reside. Hausel's and Hitchin's elegant construction uses an abstract mathematical object from the Lie group in question, a so-called nilpotent cone of Higgs bundles. The nilpotent cone refers to the iceberg. Fortunately, the tip of its structure is completely understandable in terms of weight diagrams, which serve as visual representations of the characteristic notions of the Lie group. From the tip, they can infer knowledge about the mysterious bottom, and may even reconstruct the whole representation theory of Lie groups from it [2].

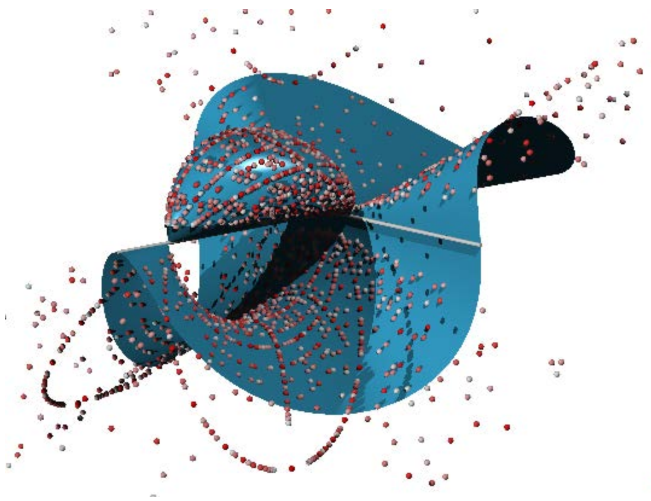
Unifying definitions of jellium

Jellium, or the homogeneous electron gas, is a fundamental system in quantum physics and chemistry. For instance, it is used for describing the deep interiors of white dwarfs and the valence electrons in alkaline metals. It was originally defined as an infinite gas of electrons in a positively charged uniform background, and the thermodynamic limit of the system was established rigorously. Another system, the *uniform* electron gas, is similar, but there is no background and the electron density is constant everywhere. In yet another model, electrons interact with periodic images of themselves.

It was conjectured in the 1980s that the ground state energies of these three systems coincide in the thermodynamic limit, but rigorously establishing this remained elusive – until now. Using a novel "floating crystal" trial state, Professor Robert Seiringer and his co-authors were able to show that these three systems are equivalent. Their argument involves "melting" a layer of crystal close to the boundary, then replacing it by an incompressible fluid. This allowed them to compensate for the charge fluctuations that occur at the system boundaries [3].

When is necessary sufficient?

Polynomials with integer coefficients and their integer solutions have long been studied and continue to fascinate researchers up to this day. Moreover, the very *existence* of solutions is an important part of this: In 1900, Hilbert famously challenged researchers to design an algorithm that could say whether a polynomial equation with integer coefficients has an integer solution. We now know that in general this is not the case. However, we can say that if integer solutions exist, two conditions must hold, one having to do with solutions that are real numbers, the other with the divisibility of solutions by integers. A central conjecture in number theory suggested that if the number of variables (call this number n) of a polynomial is greater than the degree d, then passing these two tests almost always means there exists an integer solution. In 2020, Professor Tim Browning and his coauthors proved the conjecture to be true in all cases except when $n = 4$ and $d = 3$, a family known as cubic surfaces. Their result provides the key to creating an algorithm – of the kind Hilbert asked for – for nearly all polynomials whose number of variables is greater than their degree [1].



The cubic surface $x^3 + xy + yz^2 = 0$. It passes the two tests and has numerous integer solutions (red dots), but the conjecture remains open for cubic equations in general. © Ulrich Derenthal/Leibniz University Hannover

References

[1] T. Browning, P. Le Boudec and W. Sawin, The Hasse principle for random Fano hypersurfaces, preprint, arXiv:2006.02356 (2020)

[2] T. Hausel and N. Hitchin, Very stable Higgs bundles, equivariant multiplicity and mirror symmetry. *Invent. Math.*, DOI 10.1007/s00222-021-01093-7 (2022)

[3] M. Lewin, E. H. Lieb and R. Seiringer, Floating Wigner crystal with no boundary charge fluctuations. *Phys. Rev. Lett. B* **100**, 035127 (2019)

Florian Schlederer is an Austrian book author, science curator, and one of three science writers at the Institute of Science and Technology Austria (ISTA). After having studied both physics and philosophy at the Universities of Vienna, Oxford, and Tokyo, he currently searches for intriguing story angles to disseminate mathematics, physics, computer science, and climate science into society.

communications@ista.ac.at

Live and remote: Nairobi workshops in algebraic geometry

Chiara Bonadiman, Diletta Martinelli, Jared Ongaro and Balázs Szendrői

Introduction

Since August 2015, a yearly workshop in algebraic geometry has been held at the University of Nairobi, a regular event that helped to establish a strong connection between a group of international researchers and local professors and students in Nairobi. The origins of the workshops go back to a 2014 summer school on algebraic geometry held in Mombasa, Kenya [1]. This school was run by a small but active local group: a senior algebraist, Claudio Achola, and two young geometers, Damian Maingi and Jared Ongaro, under the auspices of the Eastern African Universities Mathematics Programme[1]. At the school, discussions between locals and internationals led to the idea that the growth of the local research community would be best helped by a regular event.

The workshops held[2] between 2015 and 2019 varied in length, content, and scope, from a few days of lectures by one or two internationals to large-scale meetings. One of their core functions was to provide ideas for Masters projects for the local student cohort; indeed, during this time, 11 students completed Masters theses in geometry in Nairobi, some moving on to international PhD positions. Another role that the workshops have taken on is to establish an Africa-wide network of algebraists and geometers.

In 2021, during the second year of the COVID-19 pandemic, we organized a remote workshop. The aim of our article is to compare this event to one of the earlier ones, and to draw some conclusions.

Participants of the first workshop in 2015

The 2018 workshop

In August 2018, the fourth workshop in algebraic geometry was held in Nairobi over two weeks. Participants included several internationals, over 20 African students (both MSc and PhD), as well as senior African mathematicians.

Finding good topics for such a workshop always presents a challenge: on the one hand, one would like to introduce students to current research areas; on the other hand, there is the danger of choosing topics that are too advanced and do not connect with their mathematical background. This is particularly true for a workshop in algebraic geometry, a highly technical subject with many requisite preliminaries.

On this occasion, the first week allowed us to try to bridge this gap. Introductory lectures were given on the general theory of algebraic varieties, on blow-ups and resolution of singularities, as well as on elliptic integrals and Picard–Fuchs equations. A more challenging lecture provided an explanatory overview of the birational classification of algebraic varieties and the minimal model program. The morning lectures were complemented by exercise classes.

We found that the level of prior mathematical knowledge of the students varied a lot: some of them had already been exposed to advanced material, while others had less background in pure mathematics. To reach such a diverse audience, the lectures were made as self-contained as possible, occasionally referring to deeper mathematical content that could be appreciated by the more experienced. A detailed list of references was also prepared to help students start approaching the literature. For the exercise sessions, problems had been prepared for a variety of skill levels. An informal atmosphere developed which stimulated discussions, helped also by social interactions during lunches and coffee breaks. We noted

[1] www.isp.uu.se/what-we-do/mathematics/networks/eaump/

[2] https://sites.google.com/site/nairobiagworkshop/home

Participants of the 2018 workshop

a significant change in the students' attitude over the course of the week; by the end, they were much more relaxed and eager to ask questions.

The second week consisted of research-level talks, to present students with active research areas and to engage the local group of researchers. The topics varied from enumerative problems to moduli theory and mirror symmetry. There was time for students' presentations, allowing junior participants to present their projects in front of an international audience and get constructive feedback. A professional development session was also held, including advice on CV writing, networking, grants and funding opportunities, resources, and workshops. International lecturers agreed to be available for further advice to students; some relationships continued for months, and sometimes years, incorporating mathematical discussions and career advice.

To end this section on a personal note, we would like to mention that one of the speakers during the second week of the workshop was Professor Marco Garuti, Academic Director of AIMS Cameroon, on leave from the University of Padova. For many of us, this was our only occasion to meet Marco; we got to know a considerate man, a skilled academic leader and someone strongly committed to the cause of African mathematics. His untimely death in July 2021 touched us all deeply.

The 2021 workshop

A planned 2020 workshop had to be postponed due to the onset of the COVID-19 pandemic. One year later, the global health crisis still made it impossible to organize an in-person international meeting. At the same time, the community in Nairobi felt the need for an event to support master students, and to maintain established links. We therefore decided to organize the sixth workshop as a hybrid event in August 2021. Groups of participants at Nairobi and Makerere Universities gathered in lecture rooms equipped with projectors and blackboards for exercise sessions. Further individuals from other Kenyan universities, Nigeria, Cameroon, Congo and South Africa participated from private computers or phones. For those who wished to actively participate, the quality of the local internet connection appeared to be largely satisfactory. The total number of active participants was around 30.

We used Zoom for lectures and problem sessions, while a Slack space was set up to share materials and communications; dedicated channels helped with the organization of workshop activities, student presentations, and the professional development session. Students found Slack a very useful tool to access all the information and materials, as well as to contact the lecturers and the other students in an informal and quick way.

The workshop consisted of mini-courses on toric varieties, Hurwitz theory and computational methods with applications to chemical reaction networks and robotics. Problem sessions were run in Zoom break-out rooms, separately for the groups of students gathered at Nairobi and Makerere, and for individual students elsewhere, with tutors surfing from one Zoom room to another to follow the students' work and to answer questions. The engagement of the students was noticeably better and the interactions more intense in the groups that were physically in the same room and could discuss at a blackboard. In the fully remote group, the participants sometimes used the whiteboard option on Zoom and sometimes typed solutions to exercises, but it was hard to incorporate mathematical arguments without an appropriate editing software. The online setting was better suited to the computer-based problem sessions of the last mini-course: students could share their screens and show directly their lines of codes to the rest of the group. The group in Nairobi also used the opportunity to test the Sangaku online tutorial system, developed by Neil Strickland, University of Sheffield, to be employed alongside a video meeting system such as Zoom. It allows users to type in answers rendered in LaTeX markup language or upload a scan of a solution using an in-built phone or webcam interface. This approach was not uniformly used across the workshop and requires further development.

A day during the week was set aside for a joint session with the African Mathematics Seminar[3], with talks by senior speakers Diane Maclagan and Bernd Sturmfels, on subjects such as tropical geometry and computer-assisted enumerative geometry. As in 2018, there was a professional development session, as well as a session for student presentations.

To better understand the student experience, feedback was collected using questionnaires as well as direct interviews with participants ranging from BSc to PhD students. Students were generally satisfied with the workshop and its organization; they found the courses interesting and useful, with the mathematical content at the right level. Many participants stressed the relevance

[3] https://sites.google.com/view/africa-math-seminar/home

of facing new topics or being presented with known topics from a different angle. As one student put it,

> *If you know something or you are already interested in something then* [attending the workshop] *helps you to grasp more and learn more but if you don't know* [yet the material discussed], *like for the undergraduate students,* [then participating in the lectures] *helps you to really open up and see these are problems, potential problems that arise in this area.*

Students who attended the workshop fully remotely highlighted the fact that they would have preferred to follow problem sessions, and to interact with other people to solve problems, in person, making the following points:

> *When* [the workshop] *is physical, I can move around the room, it gives you the whole feeling of the mathematics in the environment. When people are talking about maths, when people are trying to solve some problems it's better to be in person. I think that we can follow the conference online but when we need to discuss between students I think it's better* [to be together] *because some students sometimes are not concentrating 'cause when they find that a problem is difficult they just try to leave the problem, they don't discuss, they just keep quiet. But if we are together we can try to see, ok what we can do here, we can discuss a little bit more.*

Being in the same physical space gives one the opportunity to discuss a problem and involve mates in the resolution, but also to understand other groups' issues, help them, and learn from them. This kind of direct communication is very difficult in an online format. The intensity of remote interaction strongly depends on the students' approach, but also on the organization and the management of the sessions. Future online events will benefit from dedicated strategies[4] that further encourage debate and discussion among students.

Conclusions

The 2018 workshop provided many opportunities for informal exchanges during problem sessions, lunch and coffee breaks, that created a highly collaborative atmosphere. In contrast, the 2021 remote workshop largely lacked these personal interactions. It is also likely that the intensity of the students' attention to lectures was reduced compared to a live event. Nevertheless, in our view, the remote event was a successful experiment; it elicited positive feedback from the participants. We believe that one of the key reasons for its success was that it built upon years of live connections and collaborations between the international and the local mathematical community.

An online setting presents many challenges, but it also has obvious advantages. Compared to an in-person event, a remote one can be organized with a fraction of the funds, especially in Africa. The reduction of carbon emissions is another major advantage. As opposed to a completely online event, where each participant connects to the call individually, the experience is more engaging and interactive if groups of students can gather in the same location

[4] See [2] for some ideas of how to design activities to stimulate social interactions in an online space.

Speaker dinner in Nairobi, with the late Marco Garuti (second on right)

A cross-section of the participants at the closing ceremony of the 2021 workshop

and attend together. One idea that emerged is that for future events, a group leader could be chosen from each group to enable good communication between the remote session facilitator and the local group.

In conclusion, it seems likely to us that remote events will not fully replace in-person events. But even for live meetings, senior talks and professional development sessions can have remote speakers, allowing for contributions from people from all over the world who are not able to travel to Africa. On the other hand, remote events can form a useful complement to an ongoing in-person engagement programme between different African departments and the international research community.

Acknowledgements. We would like to thank Aurelio Carlucci, Margherita Pagano and Bianca Gouthier for help during the exercise sessions of the 2021 workshop. The workshop was financially supported by EPSRC GCRF grant EP/T001968/1 "Capacity building in Africa via technology-driven research in algebraic and arithmetic geometry", held by BS at the University of Oxford. CB's participation was supported by grant PRIN 2017 "Moduli Theory and Birational Classification", held at the University of Trento.

References

[1] T. Denton, Mathematics in Kenya. *Notices of the AMS* **62**, 754–759 (2015)

[2] A. Matveeva and N. Nikolaev, Junior Global Poisson workshops. *EMS Magazine* **119**, 40–41 (2021)

Chiara Bonadiman obtained her Master's degree at the University of Trento. She is now a PhD student in mathematics education at the University of Florence.

chiara.bonadiman@unifi.it

Jared Ongaro is currently lecturer at the Department of Mathematics, University of Nairobi. His research interests are in enumerative algebraic geometry. He is an e-learning champion in his faculty; he led a team that worked with the University of Sheffield, UK, to create the SOMAS learning platform, built on Moodle but allowing randomization, so that each student receives a randomly generated problem to solve.

ongaro@uonbi.ac.ke

Diletta Martinelli is currently assistant professor at the University of Amsterdam. Her field of research is algebraic geometry with a special focus on higher dimensional birational geometry and moduli theory. She graduated from Imperial College London, and held postdoctoral positions at Edinburgh, MSRI Berkeley and Glasgow. In recent years she has been involved in a variety of projects in developing countries, including masterclasses in several African countries and in Pakistan, supervising Masters students, as well as outreach initiatives and projects aimed at improving the gender balance in academic settings.

d.martinelli@uva.nl

Balázs Szendrői is an algebraic geometer. Having studied in Cambridge, and worked in Warwick and Utrecht, he is currently professor of pure mathematics at the University of Oxford. He is involved in a number of research development projects across Africa; he is also a member of the Committee for Developing Countries of the EMS and the Council of AIMS-South Africa, and a trustee of the charity Supporting African Maths Initiatives.

szendroi@maths.ox.ac.uk

Calculus: A new approach for schools that starts with simple algebra

R. Michael Range

We outline a novel approach to tangents and derivatives that does not use any limits. Instead, it uses elementary algebraic concepts related to the quadratic equation, and therefore fits right into the school curriculum. Adding an elementary estimate to the central algebraic factorization naturally leads to the concept of continuity and thereby reveals that the algebraic derivative can also be captured by an approximation process. This turns out to be critical for handling non-algebraic functions, such as exponential functions. In order to capture the new elusive limits, students recognize that one needs to expand the familiar rational numbers to the much more intriguing real numbers. The solution of the tangent problem for exponential functions leads to the general notion of a differentiable function, in a formulation that is the natural generalization of the algebraic version, and which has been known for over 70 years. This approach gradually proceeds from most elementary concepts to the heart of analysis, making it clear to students along the way why more sophisticated tools are needed, and providing motivation for the more advanced concepts that are indispensable for a proper understanding of calculus.

1 The beginning

The standard introduction to calculus typically involves limits pretty much from the beginning. For example, the calculation of the slope of the tangent to $y = x^2$ at (a, a^2) leads us to consider $\lim_{h\to 0} \frac{(a+h)^2 - a^2}{h}$, whose intuitive answer $\frac{0}{0}$ is meaningless. However, as long as $h \neq 0$, it is easy to simplify the expression to $2a + h$, which makes it obvious that the answer should be $2a$. Yet there still remains the problem with the limit, something that prompted intensive discussions already in the early 18th century, led by the criticisms of the philosopher George Berkeley (1685–1753) (see [5, pp. 628–630]). Eventually, mathematicians in the 19th century placed limits on a solid foundation and since then students have had to learn about them first. But have you ever wondered if there shouldn't be an easier way that avoids limits until they really are needed?

Thanks to René Descartes (1596–1650), there actually is such an elementary way, and students in school can be introduced to it as soon as they have learned the basics of the quadratic function and equation. Tangents had been considered long before Descartes' days. Well over 2000 years ago, Greek geometers studied tangents to simple curves, such as circles, or more generally, conic sections. For them, a tangent to a curve at the point P is *a line that* ***touches*** *the curve at P, but does not* ***cut*** *it* (see [5, p. 120]). While this language is somewhat vague, its intuitive meaning is quite clear, and it turns out that it is very easy to give it a precise algebraic interpretation.

The crux of the matter is visible in Figure 1 that shows the graphs of two quadratic functions of the form $f(x) = (x - b)^2 + c$, where b, c are some fixed numbers.

The curve on the right corresponds to the case where $c < 0$: the x-axis intersects the graph of f in two *distinct* points, and surely it would seem appropriate to say that the axis "cuts" the curve in each of those two distinct points. The curve on the left corresponds to the case $c = 0$. In this case, the two points of intersection have blended together, one on top of the other, and it would be appropriate to say that the x-axis just "touches" the curve at that

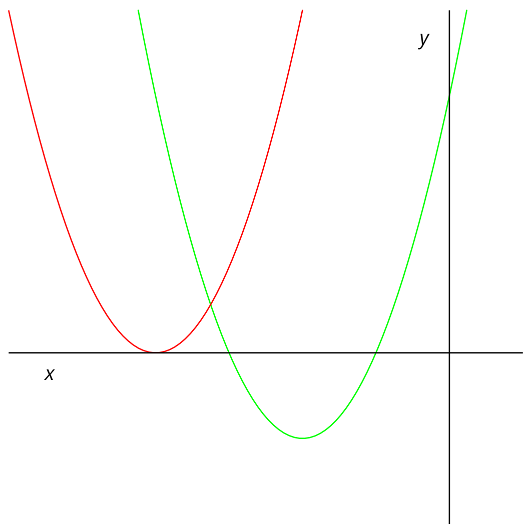


Figure 1

EMS MAGAZINE 124 (2022) — DOI 10.4171/MAG/70

point, but does not "cut" it. So, according to Greek geometers, the x-axis surely is a tangent to the curve at that point. Algebraically, in this case, the point $(b, 0)$ of intersection is determined by the solution(s) of the equation $(x-b)^2 = 0$, which is a quadratic equation for which the two roots coincide; that is, we have a "double root (or zero)", or a root of "multiplicity 2". Geometrically, we have a "double point", that is, two points that just happen to lie on top of each other, so that only one point is visible. Of course, if the curve is just moved down a little bit, the two points of intersection separate and both become visible.

So, by just looking at this most simple example, we conclude that the ancient notion of a touching line is made precise geometrically by recognizing that the special point where the line touches the curve is really a double point, and furthermore, that this is made precise algebraically by saying that the relevant equation that determines the points of intersection has a double zero.

Given that the quadratic equation is a central and important topic in school algebra, which in particular includes the case of a double zero, this seems to be the perfect place to introduce the students to the basics of calculus. All that is needed is to appropriately interpret the picture above to conclude:

> A tangent to a curve at a point P is a line that intersects the curve at P with multiplicity 2 (or higher).

I can hear algebraic geometers say that they have known this for a long time. True, after all Descartes' idea did leave a trail. But have you ever used this idea when you are teaching a calculus class? If yes, great for you! But sadly, I have not yet found any calculus text that utilizes this simple idea to introduce students to tangents.

2 Historical remarks

As we just mentioned, the idea of finding tangents via double zeroes appears first in Descartes' work [3]. Actually, Descartes was more fascinated by *normals* to a curve rather than by tangents[1], but knowing either one determines the other. Perhaps Descartes had simply been thinking of generalizing what had long been known for a circle, namely that the tangent at a point P on a circle is perpendicular to the normal at P, which in this case is simply the line from the center of the circle to the point P. Therefore, in this case, the normal is the obvious tool to find the tangent. Descartes solved the problem for an ellipse and emphasized the generality of his method. Descartes' expositor Frans van Schooten applied the double point method to find the tangent to a parabola directly [9]. On the other hand, attempts to apply this method to more general curves led to formidable complications. The situation was captured by Howard Eves [4, pp. 284–285] as follows: "Here we have a general process which tells us exactly what to do to solve our problem, but it must be confessed that in the more complicated cases the required algebra may be quite forbidding." Thus Descartes' approach was eventually forgotten. On the other hand, as you will see shortly, the implementation of the double point method presented here, first published in [6], is completely elementary, and applies to all curves defined by algebraic expressions. So it is natural to wonder why mathematicians in the 17th century overlooked what today appears to be so obvious. One possible explanation is based on the fact that even though the concept of slope was certainly known at that time ("quotients of infinitesimals" were supposed to capture it), the *point-slope form* to describe lines – something every high school student today is familiar with – was not known or used explicitly at that time. In fact, it seems to have appeared first only more than 100 years later, in a 1784 paper by Gaspard Monge (1746–1818) (see [1, pp. 205–206]). Instead, Descartes, and everybody else around that time, described lines through the point P on the curve by using a second distant point Q on that line. One then changed the position of the line by moving that point Q along some other line, often a line of symmetry of the curve under consideration. But in case of more complicated curves, there was no obvious natural place to move that second point along. Perhaps philosophers and scientists in the 17th century were still so strongly under the spell of Euclid (approximately 325 BC – 265 BC), who had introduced the axiom that a line is defined by two (distinct) points, that they just couldn't conceive of describing a line by a single point and its slope.

3 Polynomials and the Chain Rule

This is not the place to present all the details of our proposed new approach to Calculus (see [8] for some more details). But I want to highlight two simple important results that can readily be presented to students pretty much at the beginning.

First, suppose P is a polynomial of degree n. We want to identify the tangent to its graph at the point $(a, P(a))$. An arbitrary (non-vertical) line through $(a, P(a))$ is described by an equation $y = P(a) + m(x-a)$, where m is the slope. Its points of intersection with our curve are given by the solutions of the equation

$$P(x) = P(a) + m(x-a).$$

Fortunately, we do not need to find all solutions of this equation. We only are interested in the *one* obvious zero, namely $x = a$, and we want to determine when this is a zero of multiplicity greater than 1. We rearrange the above equation in the form

$$P(x) - P(a) - m(x-a) = 0.$$

[1] Descartes wrote: "… when I have given a general method of drawing a straight line making right angles with the curve. And I dare say that this is not only the most useful and most general problem in geometry that I know, but even that I have ever desired to know." [3, p. 95]

A standard result gives the factorization

$$P(x) - P(a) = q(x)(x - a),$$

where q is a certain polynomial of degree $n - 1$. While the existence of such a factorization is not trivial, it is an immediate consequence of the following fundamental result which should definitely be part of any discussion of polynomials in school.

Theorem. *If P is a polynomial of degree $n \geq 1$ that has a zero at the point a, i.e., $P(a) = 0$, then P has a linear factor of the form $x - a$; that is, there exists a polynomial q of degree $n - 1$ such that $P(x) = q(x)(x - a)$.*

In our case, we simply apply this theorem to the polynomial $P(x) - P(a)$, which does indeed have a zero at a.

By using this result, we can now factor our equation $P(x) - P(a) - m(x - a) = 0$ and rearrange it in the form

$$[q(x) - m](x - a) = 0.$$

This clearly shows that a is a solution of the equation – after all we only consider lines that intersect the curve at $(a, P(a))$! Most significantly, this factorization shows that a is a zero of multiplicity greater than one if and only if the factor $[q(x) - m]$ also has a zero at a, and this occurs precisely if the slope $m = q(a)$.

Let us formalize our result in the following theorem that completely solves the tangent problem for any polynomial.

Theorem. *Let P be a polynomial of degree $n \geq 2$ and let $(a, P(a))$ be a point on its graph. Then there exists a unique line through $(a, P(a))$ that intersects the graph at that point with muliplicity greater than 1. The slope m of that line is given by $q(a)$, where q is the polynomial of degree $n - 1$ in the factorization $P(x) - P(a) = q(x)(x - a)$.*

Of course, we call this unique line given by the theorem the *tangent to the graph of P at the point $(a, P(a))$*. Its slope $m = q(a)$ is called the *derivative of P at the point a*, and it is denoted by $D(P)(a)$ or also by $P'(a)$.

Example. If $g(x) = x^2$, then the relevant factorization is $x^2 - a^2 = (x + a)(x - a)$, so $q(x) = x + a$, and hence $D(g)(a) = q(a) = 2a$. So we have obtained the expected result without any limits whatsoever!

It is straightforward to generalize this to an arbitrary power function $P(x) = cx^n$, where c is a constant and n is a non-negative integer. Just as easily one can verify that if f and g are two polynomials, then $(f + g)'(x) = f'(x) + g'(x)$, and consequently one obtains the familiar formula for the derivative of a polynomial, all this without any limits!

Next we want to discuss the Chain Rule. This order also differs from the standard curriculum, which typically discusses product and quotient rules first. Perhaps this is due to the fact that multiplication and division are such standard operations in school algebra that it seems natural to extend these operations to polynomials or more general functions. Unfortunately, the relevant rules for differentiation look quite complicated and mysterious for the beginning student, adding more frustration. In contrast, as mathematicians know, the natural operation to consider for functions in the most general setting is *composition*, that is, apply one function after the other. It therefore should not come as a surprise that the rule for differentiation of compositions is as easy as it gets, with a completely transparent proof.

In detail, suppose f and g are polynomials, and consider the composition $(f \circ g)(x) = f(g(x))$, which is again a polynomial. We want to find its derivative at the point $x = a$. Let $b = g(a)$, and introduce the relevant factorizations

$$f(y) - f(b) = q_f(y)(y - b) \text{ and}$$
$$g(x) - g(a) = q_g(x)(x - a),$$

where q_f and q_g are polynomials. The obvious step is to replace $y = g(x)$ and $b = g(a)$ in the first equation, resulting in

$$f(g(x)) - f(g(a)) = q_f(g(x)) \cdot [g(x) - g(a)],$$

and then to introduce the second equation to obtain

$$\begin{aligned}(f \circ g)(x) - (f \circ g)(a) &= f(g(x)) - f(g(a)) \\ &= q_f(g(x)) \cdot [q_g(x)(x - a)] \\ &= [q_f(g(x)) \cdot q_g(x)](x - a).\end{aligned}$$

Clearly, this is the relevant factorization for $f \circ g$, and it follows that

$$D(f \circ g)(a) = q_f(g(a)) \cdot q_g(a) = D(f)(g(a)) \cdot D(g)(a).$$

We thus have established the Chain Rule.

Yes, you might say that all this only handles the case of polynomials, while in the case of general differentiable functions things surely must be more complicated. Well, not really. As we will show later, the above proof, combined with natural properties of continuous functions, carries over as it stands.

It is natural to introduce inverse functions at this place, and then the relevant rule for derivatives. There is a new twist, as taking inverses where a function is one-to-one usually brings in a new kind of function, so that the relevant factorization needs to be established as well. But we shall skip these details and refer the interested reader to [6, 7]. Also, considering inverses will typically take us beyond the rational numbers, although this can be avoided by restricting the domain of the inverse of a function f defined on the interval I to the image $f(I)$. For example, if we just consider $\sqrt{x}$ on the domain $\{r^2 \mid r \in \mathbb{Q} \text{ and } r \geq 0\}$, everything can be done by just using rational numbers.

4 Rational functions

The next topic I propose for the school curriculum is to investigate the class of rational functions, defined by taking quotients of polynomials. There are two main reasons for this choice. First of all, it takes the student up one step on the ladder to mastering calculus. While no really new ideas are needed in order to handle tangents and derivatives, things do get more complicated algebraically. Also, this seems to be a good place to introduce the quotient rule for derivatives (and of course the product rule as well), since rational functions are the first non-trivial examples where quotients appear. In particular, the quotient rule shows that the derivative of a rational function is again a rational function, with the same domain. The other reason is that the study of the field $\mathbb{Q}(x)$ of rational functions is a good opportunity to strengthen the student's familiarity with the fundamental rules that she/he has learned while mastering the rational numbers $\mathbb{Q}$. And these rules (or axioms) will become even more critical when one must introduce the *real* numbers $\mathbb{R}$, which notoriously are very difficult to describe exactly, so that "following the rules" becomes even more important. Of course, mathematicians are used to dealing with abstract sets, whose elements are required to follow the rules that define the particular structure that is studied. But for students, this is a big leap, and I believe that it is important to help them to prepare to deal with the more axiomatic approach that will be needed for the field of real numbers.

The main lesson is that in order to learn about tangents for rational functions, no new ideas are needed beyond what we used for polynomials. In fact, the methods can be readily extended to all functions built up from polynomials by algebraic operations.

Furthermore, it is remarkable that everything we have covered so far does not involve any limits, and it can be handled by just using the rational numbers $\mathbb{Q}$. Students thus get introduced to tangents and the basic rules of differentiation within the familiar setting of rational numbers. Isn't it amazing how simple school algebra allows to solve a major problem – at least in the algebraic setting – that was the focus of fundamental investigations in the 17th century? Perhaps this will help students to better understand the importance of and to appreciate the basic algebra that they are learning in school.

5 Continuity and approximation of derivatives

At this point, students may wonder if this is it. Perhaps they have heard of exponential functions or trigonometric functions, and if they found the discussion of tangents for rational functions of some interest, they may ask whether all this works for these functions as well. Before we get into that in the next section, we want to discuss another remarkable consequence of the fundamental algebraic factorization that opens the door to the next stage, where we really will be entering a new world.

Consider a polynomial P and its standard factorization $P(x) - P(a) = q(x)(x - a)$ at a fixed point a, where q is also a polynomial. It is easy to see by standard estimations that any polynomial is bounded over any finite interval. In particular, there exists a constant $K > 0$, so that

$$|q(x)| \leq K \quad \text{for all } x \text{ with } |x - a| \leq 1.$$

The factorization then implies the estimate

$$|P(x) - P(a)| \leq K|x - a| \quad \text{for all } x \text{ with } |x - a| \leq 1.$$

Clearly, this estimate shows that as x gets closer to a, then $P(x)$ gets closer to $P(a)$ as well. In symbols, $P(x) \to P(a)$ as $x \to a$. In other words, we have discovered that polynomials enjoy a fundamental property that is known as *continuity* at each point a. In case you are hooked on ε and δ, just choose $\delta = \varepsilon/K$, and you are done. But there is no need to make things so complicated for the students at this point; the estimate speaks for itself, and it confirms what students can readily see by graphing polynomials with a graphing calculator.

Remark. We want to emphasize that no limits are required yet. In fact, it is the above estimate that suggests the idea of *limit* in the simple case where the value of the limit is known and equals the value of the function. So we may introduce the notation

$$\lim_{x \to a} P(x) = P(a)$$

to summarize the statement that $P(x) \to P(a)$ as $x \to a$. Note how this approach is really the reverse of the standard one, which begins with limits, introduces continuity as a special case of limits, proves various limit theorems, and ultimately concludes that polynomials are continuous everywhere. I wonder which approach works better for our students?

Rational functions can essentially be handled by the same method, and similarly more general algebraic functions. It just takes a simple argument to show that every rational function is *locally* bounded near each point in its domain (i.e., where the denominator is *not* zero). So we have a rigorous proof that rational functions are continuous at every point where they are defined.

Let us now apply these new ideas to the polynomial (or rational) factor q in the standard factorization $P(x) - P(a) = q(x)(x - a)$. Using the new notation, we know that

$$\lim_{x \to a} q(x) = q(a) = D(P)(a).$$

This shows that the derivative of P at a is approximated by the values of $q(x)$ as $x \neq a$ gets closer and closer to a. So what are these values $q(x)$? As long as $x \neq a$, we can divide by $x - a \neq 0$, resulting in

$$q(x) = \frac{P(x) - P(a)}{x - a} \quad \text{for } x \neq a.$$

So we have reached the very beginning of the standard introduction to derivatives: the derivative is the limit of average rates of change, or, in geometric language, the slope of the tangent is the limit of the slopes of secants through the points $(a, P(a))$ and $(x, P(x))$ on the curve as $x \to a$. However, from the perspective of the student who is learning calculus as suggested in our approach, this is a new, *non-algebraic* technique to capture derivatives, which so far have been defined exactly by a simple algebraic method. It thus takes the student the first step beyond pure algebra further up the ladder towards mastering calculus. And as we start investigating non-algebraic functions, we will shortly discover how this latest insight will reveal amazing new phenomena that will force us to expand the foundations of the rational numbers $\mathbb{Q}$.

6 Exponential functions and mysterious limits

As we have reached the boundary of the algebraic techniques, the next step involves the study of other – non-algebraic – functions. Perhaps the most widely known such functions are the *exponential functions*. This past year just about everybody heard about the exponential growth of the spread of Covid-19, but other applications, such as compound interest, population growth, radioactive decay, and so on are also widely known. From the perspective of the experienced mathematician, exponential functions are the *eigenfunctions* of the differentiation operator, and therefore understanding them fully should reveal all there is to know about tangents and derivatives.

Exponential functions are easily defined for inputs that are natural numbers; most importantly, this basic case already reveals the fundamental rule

$$b^{m+n} = b^m \cdot b^n \quad \text{for } m, n \in \mathbb{N}.$$

While studying integers, rational numbers, and rational functions, students have learned that "following the rules" is a fundamental principle when one moves into new uncharted territory. So they should not be surprised that the rule we just recalled, that is, the functional equation for the exponential function, will guide us to properly define exponential functions for other numbers, such as negative integers, and then rational numbers. In particular, these rules, when extended to $1/n$ with $n \in \mathbb{N}$, require that the base b must be positive, and most importantly that

$$(b^{\frac{1}{n}})^n = b,$$

that is, $b^{1/n}$ must be the n-th root of b. While this forces us to go beyond the rational numbers, this is not much of a problem at the preliminary stage. After all, already Greek geometers recognized that there is no rational number $\frac{m}{n}$ that satisfies $(\frac{m}{n})^2 = 2$, and philosophers, scientists, and mathematicians lived for over 2000 years with this problem, and still were able to make amazing progress. So students, too, will just have to accept this reality as we move on with a preliminary investigation of exponential functions $E_b(x) = b^x$ with domain $\mathbb{Q}$, where $b > 0$.

So let us look at the tangent problem for the concrete case $b = 2$ at the point $(0, 1)$. Guided by the algebraic approach, we look for an analogous factorization

$$2^x - 2^0 = q(x)(x - 0),$$

and we immediately realize that there is no formula for q that would unambiguously tell us the value of $q(0)$. Just as in the algebraic case, that value would give us the slope of that line through $(0, 1)$ that intersects the graph of E_2 with multiplicity greater than one. Since we do know that $q(x) = \frac{2^x - 1}{x}$ for $x \neq 0$, and since we just learned that in the algebraic case the values $q(x)$ approximate $q(0)$ as $x \to 0$, it thus seems reasonable to study

$$q(x) = \frac{2^x - 1}{x} \quad \text{as } x \to 0.$$

Unfortunately, we have no tools available yet to determine whether such a limit actually exists. And if it does, what is it? If we look at numerical approximations for $q(0)$ by considering inputs $x_k = 10^{-k}$ for $k = 1, 2, 3, \ldots$, we see that the numbers $q(x_k)$ approach a strange number whose decimal expansion begins with 0.6931471805… as k gets larger, so that $10^{-k} \to 0$. To summarize: the student recognizes that we are faced with fundamentally new phenomena that force us to take a deeper look at the numbers that we are using, and to thoroughly study the mysterious limit process that is central to understanding the exponential function.

7 Analysis

It is now time to bring in the real numbers $\mathbb{R}$ by adding one more axiom beyond those that the students have been familiar with all along from the rational numbers $\mathbb{Q}$. We introduce the *completeness* axiom by requiring that the field of real numbers satisfies the *Least Upper Bound Property*, and hence also the analogous *Greatest Lower Bound Property*. This version makes precise something that is intuitively quite obvious, and it provides easy proofs of the existence of the relevant limits that are needed to rigorously complete the discussion of the exponential function. Once students are well familiar with the real numbers $\mathbb{R}$, the domain of exponential functions is extended to all real numbers, and one verifies that the functional equation continues to hold in this more general setting. The tangent problem is solved by verifying in detail the existence in $\mathbb{R}$ of the limit

$$\lim_{x \to 0} \frac{2^x - 1}{x} = \inf\left\{\frac{2^x - 1}{x} \,\middle|\, x > 0\right\} = c_2,$$

where the expression on the right is the Greatest Lower Bound of the relevant set, whose existence is guaranteed by the completeness axiom. The subscript 2 in the notation c_2 relates to the base 2 of the exponential function E_2 that is considered. We then define

$$q(x) = \begin{cases} \frac{2^x - 1}{x} & \text{for } x \neq 0, \\ c_2 & \text{for } x = 0, \end{cases}$$

so the function q is continuous at 0, since $\lim_{x \to 0} q(x) = q(0)$.

We therefore conclude that the function $E_2(x) = 2^x$ has a factorization

$$E_2(x) - E_2(0) = q(x) \cdot (x - 0), \quad \text{with } q \text{ continuous at } 0,$$

and we define its derivative $D(E_2)(0) = q(0)$. Of course, the same results hold for exponential functions E_b with arbitrary base $b > 0$, with

$$D(E_b)(0) = c_b = \lim_{x \to 0} \frac{b^x - 1}{x}.$$

The functional equation readily implies that analogous results hold at every point $a \in \mathbb{R}$, leading to the differentiation formula

$$D(E_b)(a) = E_b(a) \cdot D(E_b)(0).$$

Furthermore, the number $e = 2^{1/c_2}$, where c_2, as introduced above, equals the derivative $D(E_2)(0)$, is the unique number for which $D(E_e)(0) = 1$, so that the *natural* exponential function $E(x) = E_e(x) = e^x$ satisfies

$$D(e^x) = e^x \quad \text{for all } x \in \mathbb{R}.$$

Once we have identified the relevant property for the exponential function that identifies the slope of the tangent, it is natural to use it as the defining property for general *differentiable* functions, as follows.

Definition. The function f defined in an interval surrounding the point a is said to be differentiable at a if there exists a factorization

$$f(x) - f(a) = q(x) \cdot (x - a),$$

where the function q is continuous at a. The value $q(a)$ is called the derivative of f at a and it is denoted by $D(f)(a)$ or $f'(a)$.

Note how this definition naturally generalizes what has been central for polynomials and rational functions. Since in those cases the factor q is of the same algebraic type and known to be continuous, rational functions are trivially differentiable according to this definition at every point where defined. And of course, we just saw that exponential functions are differentiable everywhere.

It is obvious that the definition above is equivalent to the standard definition of differentiability.

This formulation of differentiability has been around for quite a while, but unfortunately it is not widely known. To my knowledge, it was first introduced by Constantin Carathéodory (1873–1950) at least as early as 1950, in his classic text *Funktionentheorie* [2]. It has been used in Germany in courses and books since the early 1960s, and it slowly has appeared in US textbooks since the late 1990s (see [6] and [7, pp. xxvi–xxvii] for more details).

It should be mentioned that Carathéodory's formulation of differentiability naturally suggests the basic relationship $(df)_a = f'(a)\,dx$ between "differentials" that is widely used in applications, namely, that a small change dx in the input leads to a small change $df(dx)$ in the output, which is well approximated by the product $f'(a) \cdot dx$. In fact, the error between $\Delta f = q(x)\Delta x$ and $(df)_a(\Delta x)$ is given by $[q(x) - q(a)]\Delta x$, which, because of the continuity of q at $x = a$, is negligible compared to Δx as long as x is very close to a. Of course, working with differentials df and dx is loaded with much historical baggage, given their classical interpretation as "infinitesimals", while for differential geometers it is standard to view the differential $(df)_a$ at a as a linear function defined on the tangent space at that point. It is hard enough for mathematicians to keep all this straight, and for students this is bound to be very confusing. Clearly, we need to find a way to help our students to sort this out. Perhaps it is time to introduce *different* notations for infinitesimals, small changes in input/output, and the underlying linear function, instead of just using df.

The Chain Rule, revisited. We already introduced and proved this rule for polynomials above. The proof for general differentiable functions follows exactly the same structure as the proof in that case, and is completed by just using natural properties of continuous functions, such as the fact that compositions and products of continuous functions are continuous. Again, the reader should compare this with the various proofs of this important result found in standard textbooks, and decide for her/himself which proof we should teach our students.

Aside from allowing the simplest proof of the Chain Rule known to me, Carathéodory's formulation has other advantages. In particular, it naturally extends to functions and maps of several variables once the appropriate linear algebra is introduced, and it allows simple proofs of the Chain Rule and Inverse Function Rule in the higher dimensional setting.

8 Concluding remarks

This concludes our outline of how we propose to introduce students to differential calculus, beginning with rational numbers and the quadratic equation. Once we have reached the heart of Analysis and in particular the general notion of differentiable functions, the story continues, more or less, along more traditional lines. We hope that this brief outline has convinced the reader that there is an alternative to the traditional approach, a new approach that should make things easier for students. It gradually builds up from simple algebra to lead them to the point where they will recognize the true necessity of deep new ideas involving more advanced mathematical tools that are the heart of analysis. The author is currently working on a detailed textbook that implements this outline and which may be viewed as a substantial expansion of the first half of his earlier book [7].

References

[1] C. B. Boyer, *History of analytic geometry*. Scripta Mathematica, New York (1956)

[2] C. Carathéodory, *Funktionentheorie*. Birkhäuser, Basel (1950)

[3] R. Descartes, *La Géometrie. The Geometry of René Descartes*. Reprint of the 1637 Leiden edition, with English translation by D. E. Smith and M. L. Latham, Open Court Publ. Co., Chicago (1925), reprinted by Dover, New York (1954)

[4] H. Eves, *An introduction to the history of mathematics*. 3rd ed., Holt, Rinehart and Winston, New York (1969)

[5] V. J. Katz, *A history of mathematics*. 3rd ed., Addison Wesley, Boston (2009)

[6] R. M. Range, Where are limits needed in calculus? *Amer. Math. Monthly* **118**, 404–417 (2011)

[7] R. M. Range, *What is calculus? From Simple Algebra to Deep Analysis*. World Scientific Publishing, Singapore (2015)

[8] R. M. Range, Using high school algebra for a natural approach to derivatives and continuity. *Math. Gaz.* **102**, 435–446 (2018)

[9] F. Van Schooten, *Geometria à Renato Descartes. Anno 1637*. Gallice edita, Amsterdam (1659–1661)

R. Michael Range earned his *Diplom in Mathematik* at the University of Göttingen in 1968, where lectures of Hans Grauert got him hooked on multidimensional complex analysis. A Fulbright Fellowship brought him to the United States, where he earned a Ph.D. at UCLA in 1971. He has held academic positions at Yale University, the University of Washington, and at the State University of New York in Albany, as well as research positions at institutes in Bonn, Stockholm, Barcelona, and Berkeley. Range has published numerous research articles, and is the author of *Holomorphic Functions and Integral Representations in Several Complex Variables* (Springer 1986 and 1998), a widely used introduction to the field. He has lectured in many countries, including mini-courses at the Chinese Academy of Sciences in Beijing, at POSTECH in South Korea, and in Trondheim, Norway. More recently he has written historical and expository articles on multidimensional complex analysis, one of which was recognized with the Lester R. Ford award of the Mathematical Association of America in 2004. He also has spent much time thinking about the calculus curriculum, with the goal of making things easier for the ever increasing number of students. Range loves mountains and enjoys hiking, biking, and especially downhill skiing. In 2015 he retired to Park City, high in the beautiful mountains near Salt Lake City in Utah.

range@math.albany.edu

Math curriculum matters: Statistical evidence and the Portuguese experience

Nuno Crato

In this paper, I provide an overview of the recent evolution of Portuguese students' results in elementary and middle school mathematics. I highlight the reforms slowly introduced from 2003 to 2015, and their results. These reforms were pragmatic and made in response to the poor results obtained by Portuguese students in the early TIMSS and PISA studies and got a significant and deliberate boost in 2011, when the government invited experts from the Portuguese Mathematical Society to collaborate on new programs and standards. Results from both PISA and TIMSS after the application of the new standards showed a significant improvement, with 4th grade Portuguese students passing their counterparts from traditionally better performing countries, such as Finland. Subsequent abolition of the new standards and other reforms of the period led to a significant backslide of the educational panorama. However, only now the news of Portuguese successes is spreading, and we must look at what has determined the advances and regressions. In order to understand what leads to good education results, we need to look at what a country did, and not at what it is doing now.

1 Introduction

In a well-known dialogue from the pen of the mathematician-turned-writer Lewis Carroll, Alice asks the Cat: "Would you tell me, please, which way I ought to go from here?" "That depends a good deal on where you want to get to," said the Cat. "I don't much care where –" said Alice. "Then it doesn't matter which way you go," concluded the Cat.

Any reasonable and mature person understands it perfectly: With clear goals, it is easier to progress. Any business- or management-oriented person understands it perfectly: When we set up goals, plans, and monitor a schedule, we progress faster. Any experienced teacher worried about student's progress understands it perfectly: With clear curricular goals and planned lessons, students advance faster and in a consistent way.

However, this Alice dialogue is a paradigm of what happens frequently in educational debates. More often than not, curricular goals are derided in favour of generic "skill development". "Curricular flexibility" is a pretext for not conforming to generic curricular goals. The so-called deep understanding is often a pretext for eschewing the assessment of curricular attainment.

In this paper, I will describe a positive experience in the Portuguese education system along the years 2003–2015, namely from 2011 to 2015, and its backlash after 2016.

2 A push for quality over quantity

Throughout the twentieth century, Portugal struggled against a backwards education environment. In 1970, almost 18 % of the population was still illiterate, 66 % of the 15-year-old hadn't completed any level of formal education, and only 0.9 % of the total population had a higher education degree.[1]

The progresses made during the last decades of the twentieth century are extraordinary. Following the general improvement of economic conditions after the 1960s, the euphoria of a baby boom, a newly installed democracy, and the arrival of European structural funds, we witnessed a school expansion and a complete change of the country. In 30 years, illiteracy dropped from 18 % to 9 %, the percentage of 15-year-olds without any level of formal education dropped from 66 % to 9.2 %, and the fraction of the population with higher education raised from 0.9 % to 8.4 %.[2]

All these successes were essentially quantitative, i.e., they democratized education, expanded the school system, and increased schooling years. But did they bring youngsters to a reasonable level of literacy? A debate about the quality of education began dividing the country.

Shocking news arrived late, with the first international large-scale assessments (ILSA). In 1995, TIMSS showed Math 4th grade Portuguese students at the bottom of the scale, with only two

[1] Pordata, www.pordata.pt, consulted 23 July 2021.

[2] Here and in the following description, I draw heavily from my previous paper on the Portuguese experience [1].

countries, Iran and Iceland, behind[3]. In 2001, as PISA[4] 2000 results were released, Portugal saw its students' score below participant countries average and much below the OECD average in all three areas (literacy, mathematics, and science).

During the ensuing years, the debate continued and took many shapes. Reforms in the school system were at times contradictory and served different purposes, but until 2015 they essentially went in one direction: to pay more attention to the results [5]. Several changes propelled the attention to education outcomes.

In 2001, the government was legally forced to release school exam outcomes that previously were hidden from the public. This was a game changer, as it increased public awareness of the diversity in schools' quality and put pressure on schools and teachers to improve education results. In 2003, a new minister established exams for mathematics and reading at the end of compulsory schooling (then the 9th grade). In 2005, the ministry established priority of mathematics and reading subjects and developed special plans to support teaching of these disciplines. In 2010, a new minister introduced in Portugal the first *learning standards*, following an Anglo-Saxon lead. These standards did not replace the existing curriculum; they simply provide it with a clearer structure.

In Portugal and other countries with a highly centralized system, the curriculum is usually subject-based and essentially consists of a set of official documents called *programs* ("programas"), which detail the topics to be covered in each school discipline or subject. Standards typically organize the course contents sequentially, highlighting the learning goals, and the achievement level desired for each content. They refrain from pedagogical recommendations and favour the setting of detailed learning outcomes.

The first learning standards appeared in 2010 and 2011. Although they represented a progress with respect to the vague programs in place at that time, they still included pedagogical recommendations mixed up with learning outcomes; they still didn't clearly highlight knowledge goals, and they still were vague in some areas.

By December 2010, PISA 2009 results were published and showed an important improvement in all PISA areas. Some analysts stressed the importance newly instituted 9th grade exams may have had, while others emphasized the role of the new policies of increased attention to results in basic subject areas. In my opinion, both points are correct.

A financial and political crisis exploded in early 2011. Portugal was coming to grips with the most serious financial crisis of its recent history. In June 2011, elections were held, a new majority was formed, and a new prime minister was chosen. I was then appointed as an independent minister. Budget was as tight as possible.

Against this background, the education policy had to be very clear and focused. From the start we decided that we should try to "do better with less", i.e., we should focus on the quality of education.

3 To push for quality under budgetary constraints

We can group the main reforms put in place from 2011 to 2015 into five areas. I still believe the most important one was the setup of a clear and demanding curriculum, allowing a rigorous assessment of students. I will develop this idea in the next section, briefly summarizing now the main reform areas.

First, *curriculum*. As I will describe better in the next section, we designed an increasingly demanding and structured curriculum.

Second, *assessment*. When standards related to the curriculum are well defined and well structured, it is possible to align assessment with these standards. This way, assessment can be more rigorous and act as a reference to teachers, students, and parents. The introduction of rigorous, frequent, and varied assessment tools was a crucial part of the 2011–2015 reforms.

Third, a *plan for success promotion*. In parallel to striving for higher academic standards, we devised a series of measures to improve students who trailed behind and, at the same time, to allow more advanced students to thrive by pursuing some of their specific interests. These measures were set out as early as 2012 in a special law,[5] and complemented by regulatory legislation that made compulsory the support to students with academic difficulties.

Fourth, *school autonomy with incentives tied to students' improvement*. After 2012, we developed a complex system to increase resources allocated to schools as they proved to be able to improve students results with these additional resources. Autonomy allowed schools to freely use their resources to put in place the promotion-of-success measures previously referred to.

Fifth, *parallel offers and vocational tracks*. As compulsory schooling was extended from 9th to 12th grade, vocational high-school tracks became a choice for the three years of senior high school.

[3] TIMSS (Trends in International Mathematics and Science Study) is a large-scale assessment designed to inform educational policy and practice by providing an international perspective on teaching and learning in mathematics and science. TIMSS is a project of the International Association for the Evaluation of Educational Achievement (IEA) and is directed by the TIMSS International Study Center at Boston College in collaboration with a worldwide network of organizations and representatives from the participating countries.

[4] PISA is the Organization for Economic Co-operation and Development (OECD) Programme for International Student Assessment. Started in 2000, every three years, it now tests 15-year-old students from almost the entire world in reading, mathematics, and science. The tests are designed to assess how well students master key subjects to be prepared for real-life situations in the adult world.

[5] "Decreto-Lei 176/2012 de 2 de agosto."

4 Everything starts with the curriculum

There are many definitions of curriculum, starting from a more restricted one, which usually understands curriculum as the specification of what is intended to be taught and learned in academic terms, to more general ones, in which the very generic purposes are considered, and methods and materials are included (see, e.g., [3]). For our purposes, we do not need a very precise definition, and will use the common restricted version just outlined.

The characteristics we intended and believed to have essentially succeeded in having in the curriculum are the following.

First, the set of courses offered from 1st to 12th grade should *prioritize the commonly accepted essential subjects* and add complementary subjects such as information technology and the sorts. Among the essential subjects, we may highlight reading, literature, grammar, and writing; arithmetic, geometry, algebra, and basic probability and statistics; country and world basic history; geography; sciences, physics, chemistry, biology, earth sciences; arts and basic art history.

Second, the *foundational subjects such as reading and elementary mathematics should receive special attention*. All subjects are important, but some have precedence along the school years and priority, as they are essential to a civilized life and to progress in the studies.

These two characteristics seem trivial and indisputable, but unfortunately this is not true. Many educational currents abhor the idea of separated subjects and would like to constantly mix all subjects. These educational currents particularly reject the idea of having foundational subjects.

Third, *curricula should be organized into different subjects*, with internal coherence regarding their fields. This seems obvious, but postmodern currents have attacked the idea of organized knowledge. From their point of view, organized subject knowledge is reductive and should be abandoned.

However, organized field knowledge is the way humans found to progressively understand reality. Obviously, when we look at nature, we do not have a biological phenomenon developing outside of a physical world and outside a given planet climate, to name just a few related areas. But these spheres of phenomena have differences and the way we cope with reality is exactly by partitioning it and by studying piece by piece parts of the moving world.

Transmission of school knowledge adds another constraint: knowledge communication should be facilitated by breaking complex concepts into simpler ones and by organizing them progressively. One of the most important findings in modern educational psychology is that comprehension operates through a narrow channel of working memory which only supports a limited cognitive load (see, e.g., [4]).

Fourth, each subject needs an *internal coherence*. This is most obvious for mathematics, but the same applies to all subjects. As an example, consider learning a foreign language. Obviously, there are many possible progression paths. But if a teacher starts by introducing her/his students to basic day-to-day vocabulary and a few verbs related to everyday house life, she or he cannot immediately assign readings related to foreign travel. As another example, consider the study of Brazil's independence. To understand the basic forces behind the independence, a student must first be introduced to Brazil's colonization, King D. João VI's escape from Lisbon to Rio de Janeiro in the aftermath of Napoleon's invasion of the Iberian Peninsula, the long settlement of the crown court in Rio, the development of this city, and the return of the king to Portugal 14 years later. This interdependence can be studied in many ways, but if we want students to have a basic understanding of history as an evolution of trends and not as collection of facts, the study must have an internal coherence.

In mathematics, all this is even more relevant, as mathematics is a hypothetical-deductive discipline. In mathematics there are many ways of establishing logic sequences, but the interdependence is crucial. We usually start with definitions that set up the ground for construction of a theory, then explore basic properties, and establish mathematical facts by logical deduction.

One can be more rigorous or less rigorous, according to the students' level and the purpose of the specific course (basic algebra for vocational training is different from college preparation algebra). One can stress the formal aspects of theorem proving or stress the computational aspects of a particular topic. But in all cases, the development of topics must be internally coherent with the definitions formulated at the start and the various facts established previously.

To give an elementary example, let us consider the definition of the logarithm. One can start with exponents,

$$\log_a x = y \iff a^y = x,$$

or with the integral

$$\log x = \int_1^x \frac{1}{t}\, dt,$$

or proceed formally, e.g., define the logarithm as a strictly monotonically increasing function $L\colon R^+ \to R$ with the following property for all $x, y \in R^+$:

$$L(xy) = L(x) + L(y).$$

Now, suppose a teacher asks a student to prove that $\log(xy) = \log x + \log y$. The expected answer differs according to the definition the class is using. In case they are using the last described definition, the correct answer would simply be "it is the definition".

Mathematics is a structure. By developing only "mathematics experiences" without a structured curriculum, our students would perceive mathematics as a series of disperse facts, or even of disperse tricks. They would hardly learn mathematics. At best, they will acquire a superficial knowledge of mathematical concepts.

Fifth, *curriculum should be organized in a progressive way*, translating the internal coherence and logic of the discipline. There are of course many equally valid alternative sequences, but the curriculum should present at least one of these possible sequences. A teacher may choose a slightly different path at her or his own risk, but this requires a deep knowledge of the subject, a reasonable experience, and some art. Curricula are tasks for expert groups.

In Portugal, we redesigned the curricula by forming expert groups that included university professors of the subject (professional mathematicians, or historians, or biologists, …), experienced teachers, and experts in the science of teaching, namely educational psychologists. This departed with the recent usual practice of having curricula designed by so-called "educational experts", i.e., educationalists with no solid knowledge of the subject areas and often with a deep ideological non-scientific slant.

Sixth, *curricula should be ambitious*, knowledge rich, envision a deep understanding of the concepts and procedures, and help the grasp of the structure of the different subjects. This departs from the recent practice of trying to develop skills without knowledge, as if the ability to apply or develop a mathematical practice could be achieved without any particular substantive knowledge.

Seventh, *curricula should be translatable into assessment and auxiliary materials, namely textbooks*. This means that curricula cannot be vague, otherwise they would be useless for guiding teachers, textbook authors, exam authors, and even families involved in helping their children.

This last point means we are seeing curriculum as the centre of a pedagogical coherence: textbooks and other materials, assessment, including standardized testing, all these should be coherently aligned with the curriculum, turning this set of tools into a coherent instrument for education improvement.

5 After the continuous improvement from 2003 to 2011, a sharp progress from 2011 to 2015 and a decrease after 2016

Figure 1 shows the results of PISA assessments in all three areas along the seven PISA waves that started in 2000 and had the last survey in 2018.

I have already described the main forces that led to this great progress along the years. In essence, after the 2000 PISA shock, all governments paid increased attention to students' results and searched for ways of supporting the basic disciplines development and assessment. Two important factors for improvement played a role: one is the introduction of 9th grade math and reading exams, which led to the sharp increase in 2009, the other is the improvement in the structure of the curriculum, with better standards accompanied with additional standardized assessment in 4th and 6th grades. Later, one cannot help noticing a visible decrease in 2018 results.

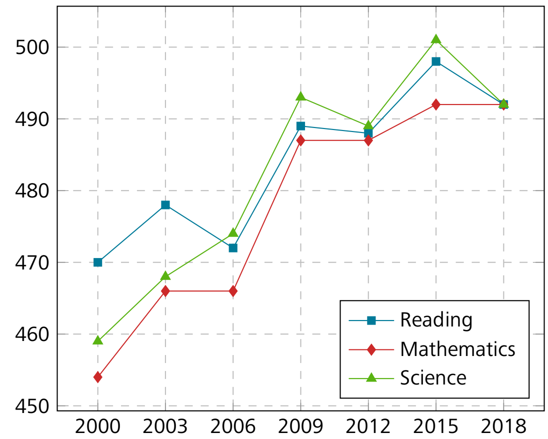


Figure 1. The evolution of Portuguese results since the start of PISA until the last PISA survey. Graph based on OECD data at https://nces.ed.gov/surveys/international/ide/

In the preface to the PISA 2018 report, one can read a surprising praise to Portuguese education achievements. It's worth quoting it: "given the fact that expenditure per primary and secondary student rose by more than 15 % across OECD countries over the past decade, it is disappointing that most OECD countries saw virtually no improvement in the performance of their students since PISA was first conducted in 2000. In fact, only seven of the 79 education systems analysed saw significant improvements in the reading, mathematics, and science performance of their students throughout their participation in PISA, and only one of these, Portugal, is a member of the OECD." This is what the former secretary-general of the OECD had to say.

As a Portuguese citizen and a former minister of education, I can only be elated at this acknowledgement. We conclude that my country is the only OECD country that has constantly improved in PISA. But as an observer, I must be surprised.

Portugal has been improving its results from to 2000 to 2015. However, 2018 is exactly a date when we witnessed a decline: a statistically significant decrease in science, a visible downturn in reading, and a stagnation in mathematics. Why is this precisely the time when we deserve such a public praise from the OECD at its highest level? One can speculate and read on it politically coloured messages, but these statements are surely misleading and similar to frequent misleading references to another country's education policies, namely Finland.

To place the PISA assessments into perspective, it is useful to have a global panorama of PISA math results. Figure 2 presents the evolution of PISA scores for an illustrative set of countries.

Singapore has been at the top, along with a couple of Asian countries and regions, such as South Korea and Macao. For a few years, Finland was the only European country that was close to those top performing countries. Now, it is Estonia that is at the

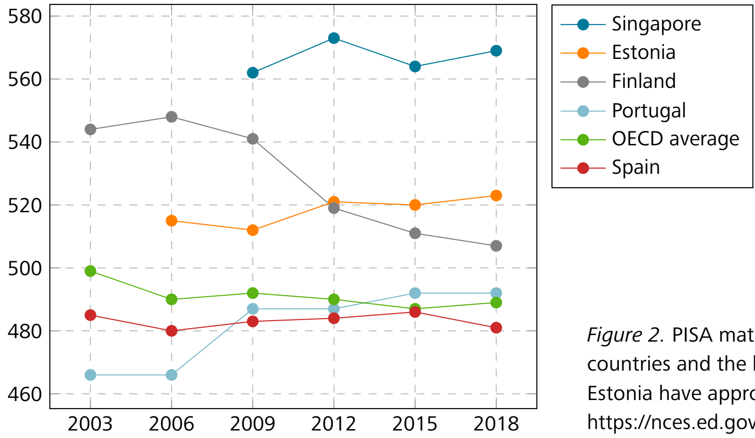


Figure 2. PISA math results show a huge difference between European countries and the high-performing East-Asian countries. Only Finland and Estonia have approached the highest ranked countries. Data retrieved from https://nces.ed.gov/surveys/international/ide/.

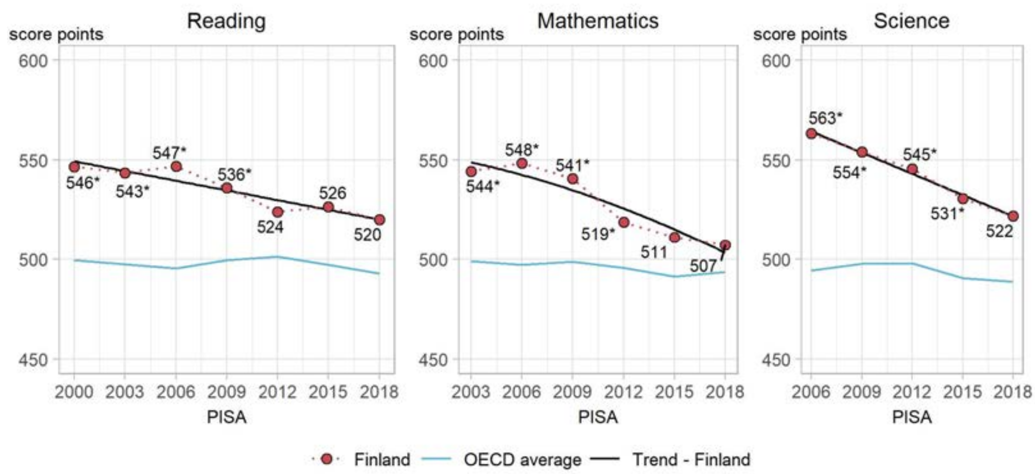


Figure 3. The evolution of Finland PISA results from the start of each assessment to 2018. We notice a sharp decrease since 2006. This means we should study and perhaps emulate what Finland did before this decrease and not try to copy what the country is doing now. Data and graphs from OECD/M. Ikeda and M. Schwabe, 2019, *Finland Country Note*, https://www.oecd.org/pisa/publications/PISA2018_CN_FIN.pdf

top of European results. Countries such as Portugal and Spain are close to the OECD average. Most rich and middle-income countries have been at a similar level[6]. Given the cultural differences between Europe and East Asia, it is natural that Europeans look up to countries such as Finland.

If we want to understand what brought the Finnish system to the comparatively high-performance levels it enjoys in the European panorama, we should look to what Finland did in the past, and not to what Finland is doing now. Given the general laudatory praises Finnish education receives in the western press, it will be probably come as a surprise to the reader the fact that Finland has been declining continuously since 2006 in all three main areas.

This reality can be seen in Figure 3, from the OECD itself, which reveals declines of 41 points in mathematics, 27 points in reading, and 41 in science. To put things into perspective, experts usually estimate that a 30-point change is roughly equal to the difference between two successive school years. We would conclude that in the last 10 or 12 years, Finnish education got worse. Roughly: middle school 15-year-old students now know about the same as 14-year-old students knew in 2006. It is highly debatable whether recent Finland education innovations are an example to follow!

However, many educationists praise current lax programs, multidisciplinarity over disciplinarity, phenomenon-based learning and discovery learning, i.e., praise the changes in Finnish education that are concomitant with its decline (see, e.g., [8,9]). We should instead

[6] Detailed data for PISA and other international surveys are readily accessible at the International Explorer of the National Center for Education Statistics: https://nces.ed.gov/surveys/international/ide/. An overview can be found in [2].

praise the demanding teacher training and teacher selection, the evaluation system, and the valorization of disciplinary knowledge that prevailed during the last decade of the 20th century and the first years of the 21st century – exactly what many educationists abhor, but exactly what raised Finland to its extraordinary level [7]. And do not take this in the wrong sense: Finland's is still a great education system by western standards.

All these points are important when we discuss any national education policy. It is not only the Finnish system that is at stake; it is any other country in the world that would like to learn from international experiences.

I obviously know better the Portuguese situation and history, so I return to this country's experience and now review TIMSS 4th grade math results.

Unlike PISA, which surveys general application skills of 15-year-old students in three main areas and every three years, TIMSS is grade based and more tied to the curriculum. Every four years, it surveys the 4th and 8th grades, both in mathematics and science. Portugal joined TIMSS in 1985 and got terrible results. Only in 2012 it re-joined TIMSS. We are particularly interested in math, as the general purpose of this article is to understand the evolution of this discipline. Additionally, up to 4th grade, sciences are not a particular focus of elementary school.

As we can see in Figure 4, Portuguese students progressed sharply up to 2015. This global picture highlights how the country was able to improve its education system during the last decades. The year 2015 is particularly notable since Portuguese students outperformed their Finnish counterparts. This is partly due to the continuous decrease in Finnish results we can also observe in the graph, but it is nonetheless notable that a medium performing country as Portugal could surpass a still high performing European country such as Finland.

Unfortunately, we also observe a significant decrease in 2018, when the results were even lower than in 2011.

Having described the main factors that led to a remarkable improvement in Portuguese education, specifically in mathematics, it is also important to interpret the decline in results observed both in PISA and TIMSS from 2016 onwards.

The education contexts in 2011 and 2016 could hardly be more different. In 2011 we were building upon a general progress with a better focus on the results, with increased curricular rigour and more evaluation. We were deeply overwhelmed by a financial crisis and the mood of the country was to fight to improve all results, even if with less resources.

The 2011–2015 ministerial team always stressed that we were building upon previous results, but we needed to improve the curriculum and to have better assessment [6].

The curriculum was improved along the orientation described in the previous section. To be specific, we focused on the fundamental subjects, namely reading and mathematics, we better structured each discipline's program, and we designed standards

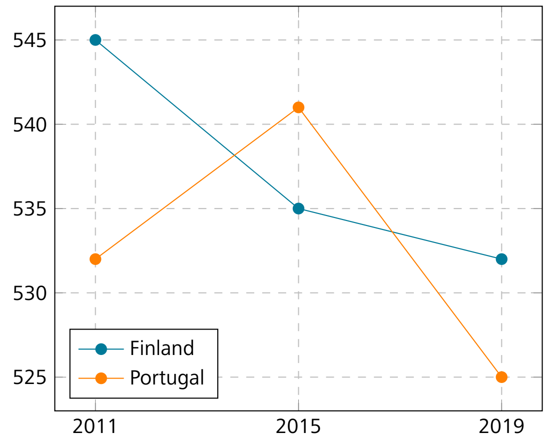


Figure 4. Results for Portuguese students increased remarkably up to 2015 due to an education policy centred on the curriculum and on student outcomes. When this policy changed in favour of a general skills and competencies approach, results decreased. Graph displays all years both countries participated in TIMSS. Data from https://nces.ed.gov/surveys/international/ide/

that established a clear and measurable progression within each subject.

Assessment was improved by creating an evaluation institute[7] with greater autonomy and technical expertise on modern statistical assessment methods, namely item response theory. Assessment was made more frequent and more stable, so that student outcomes could be compared from year to year.

In 2016, the country was coming out of a difficult period and the new parliamentary majority changed course. In education, two of the four standardized assessment moments were abolished. The entrance exam for new teachers was also abolished. The new government directly and publicly attacked previous orientations that were deemed as "elitist", "unrealistically ambitious", and "too content focused", avoiding the "competencies" approach. Instead of focusing on subject knowledge and subject coherence, the ministry then adopted a focus on multidisciplinarity and practical skills and competencies, and eschewed assessment, which was classified as "narrow" and "detrimental to the socioeconomic unfavoured classes".

Later, even after PISA results were made public, the government abolished all programs and standards, replacing them with vague and ill-structured "Essential Learnings" ("Aprendizagens Essenciais").

The country's education mood changed completely, both in administrative orientation and public perception. It is clear that this change is associated with the immediate fall in education outcomes.

[7] Instituto de Avaliação Educativa, IAVE, created by Decreto Lei nº 102/2013, de 25 de julho.

6 Inequality and fairness

One of tenets of the opposition to an ambitious and structured curriculum and to standardized assessments is the faulty idea that ambition and evaluation harm students coming from less privileged classes [6]. It is worthwhile to examine what happened to low-performing students with the 2016 change in education policy.

		2009	2011	2015	2018
Science	High performers	4.2	4.5	7.4	5.6
	Low performers	16.5	19.0	17.4	20.2
Mathematics	High performers	9.6	10.6	11.4	11.6
	Low performers	23.7	24.9	23.8	23.3
Reading	High performers	4.8	5.8	7.5	7.3
	Low performers	17.6	18.8	17.2	19.6

High performers > level 4; low performers < level 2

Table 1. The evolution of high and low performers on PISA in Portugal as percent of the total. Roughly, both proportions moved into a favourable direction up to 2015 and in a negative direction after that date. Table based on PISA data available at https://nces.ed.gov/surveys/international/ide/

It is clear that the high-performers fraction of the students moved into a favourable direction along the years up to 2015, i.e., this fraction increased, but changed course after this year. Although at a much lower rate, it kept the positive movement for mathematics, following the general trend in PISA. As mathematics is a highly cumulative discipline, it is possible that mathematics scores kept more momentum than science and reading.

It is also clear that the low-performers faction of the students moved into a favourable direction along the years up to 2015, i.e., this fraction decreased, but changed course after this year. Although at a lower rate, it kept the positive movement for mathematics, following the general trend in PISA. Again, the mathematics momentum slowed down the relapse.

PISA assesses the general applicable knowledge of 15-year-old, and it is less tied to the curriculum than TIMSS. It is worthwhile to see what happened with 4th grade students as we moved from 2011 to 2015 and 2019. The mathematics momentum I just referenced above for students who have been in school for about nine years would surely be less influential. Results in Table 2 are also very interesting.

In TIMSS, the results are unequivocal. The fraction of high performers increased from 2011 to 2015 and decreased from 2015 to 2019. The fraction of low performers decreased from 2011 to 2015 and increased from 2015 to 2019.

	2011	2015	2019
High performers	8	12	9
Low performers	20	18	26

High performers = level 4; low performers ≤ level 1

Table 2. The evolution of high and low performers on TIMSS in Portugal as percent of the total. Roughly, both fractions moved in a favourable direction from 2011 to 2015 and in a negative direction from 2015 to 2019. Table based on IEA data available at https://nces.ed.gov/surveys/international/ide/

What happened is what economists call a "natural experiment". There are two groups of students in this table. In the first group, students have entered 1st grade in 2011 and have been assessed in 2015 at the end of the first cycle of the elementary school, which in Portugal ends with the 4th grade. In the second group, students have entered 1st grade in 2015 and have been assessed in 2019. In order to understand why one is dealing with two fundamental blocks, one has to be aware that school in Portugal formally begins in 1st grade, when pupils are about 5 or 6 years of age. With the 4th grade ends what is considered the first cycle of elementary schooling, when pupils are about 10 years old. At this moment, students are assessed and most of them progress to a different school, to study at a level which corresponds roughly to what some countries classify as middle school.

This table clearly shows that students from the first group had a much superior learning experience than those from the second group. As these groups do not overlap, except for a very rare number of flunked students, this table complements Figure 2. It is now clear that first-group students not only progressed in the mean, but also their fraction of high performers increased while their fraction of low performers decreased. For the second group, exactly the contrary happened.

What happened is really a natural experiment. The first group of students entered school in a motivating atmosphere of rigour, curricular ambition, and assessment. They had a more demanding curriculum and better structured standards. They knew that they would have a standardized assessment in 4th grade, at the end of this cycle. The second group of students began school in a relaxed atmosphere of less rigour, no curricular ambition, and no assessment. They had a less demanding curriculum and no structured standards. They knew that they would not have a standardized assessment in 4th grade. The results are clear.

While the new policy obsessively dismissed knowledge ambition and assessment as policies that would harm the less privileged students, the reality is that, on the contrary, leniency harms those with a weaker background.

As sometimes economists say, policies should be judged by their results, and not by their stated goals. This is such an occasion.

Acknowledgements. I thank the editor and an anonymous referee for very helpful suggestions. This work was partially supported by Project CEMAPRE/REM – UIDB/05069/2020 financed by FCT/MCTES through national funds.

Disclaimer. For about a decade, I have been serving on the Portuguese Mathematical Society boards (2000–2011) and have been its president (2004–2010). Later, I was appointed minister of education and science of the Republic of Portugal and served as an independent in the government for a whole term (2011–2015). The results I describe in this paper encompass these periods.

References

[1] N. Crato, Curriculum and educational reforms in Portugal: An analysis on why and how students' knowledge and skills improved. In *Audacious Education Purposes*, edited by F. Reimers, Springer, Cham, DOI 10.1007/978-3-030-41882-3_8 (2020)

[2] N. Crato, *Improving a country's education: PISA 2018 results in 10 countries*. Springer, Cham, DOI 10.1007/978-3-030-59031-4 (2021)

[3] A. V. Kelly, *The curriculum: Theory and practice*. 6th ed., Sage (2009)

[4] P. A. Kirschner and C. Hendrick, *How learning happens: Seminal works in educational psychology and what they mean in practice*. Routledge (2020)

[5] J. Marôco, Portugal: The PISA effects on education. In *Improving a Country's Education: PISA 2018 Results in 10 Countries*, edited by N. Crato, Springer, Cham, DOI 10.1007/978-3-030-59031-4_8 (2021)

[6] A. Morris, Student standardized testing: Current practices in OECD countries and a literature review. OECD Working Paper Series No. 65 (2011)

[7] T. Oates, *Finnish fairy stories*. Cambridge Assessment, Cambridge, UK, www.cambridgeassessment.org.uk/Images/207376-finnish-fairy-stories-tim-oates.pdf (2015)

[8] A. Saarinen, Equality in cognitive learning outcomes: the roles of educational practices. Doctoral dissertation, University of Helsinki, https://helda.helsinki.fi/handle/10138/320436 (2020)

[9] G. H. Sahlgren, *Real Finnish lessons: The true story of an education superpower*. Centre for Policy Studies, Surrey, UK, www.ifn.se/media/0irhqejv/2015-sahlgren-real-finnish-lessons.pdf (2015)

Nuno Crato is professor of mathematics and statistics at ISEG, University of Lisbon. His research focus on stochastic models, time series applications, and statistical data-based evaluation of policy measures, namely education. He has published extensively, and has also worked on education policy analysis. He was president of the Portuguese Mathematical Society (2004–2010) and director of the International Institute of Forecasters (2016–2020). From 2011 to 2015, he was the minister of education and science of Portugal. A prolific science writer, some of his books have been translated and published in various countries, including the US, UK, Italy, and Brazil, namely his *Figuring It Out* (Springer 2010). He has received prizes from the European Mathematical Society (2003) and the European Union (2007). He was awarded Commander (2008) and Great-Cross (2015) of Prince Henry Order and Great-Cross (2022) of Public Instruction Order of the Portuguese Republic.

ncrato@iseg.ulisboa.pt

ICMI column

Núria Planas

ICMI AMOR – A collective long-term project with short-term opportunities

In 2017, the International Commission on Mathematical Instruction (ICMI) launched the project ICMI AMOR (ICMI Awardees Multimedia Online Resources) (www.mathunion.org/icmi/awards/amor). The project was initiated by Jean-Luc Dorier and, since then, it aims at building online resources reflecting highly significant and influential research in mathematics education at an international level. The R in AMOR particularly stands for independent units consisting of a series of video modules, with each unit devoted to one mathematics education researcher who was honored with the Felix Klein Award or the Hans Freudenthal Award (www.mathunion.org/icmi/awards/icmi-awards). The Emma Castelnuovo awardees are in the process of being also integrated in the project.

A question at the core of AMOR development is how we narrate the history and present of our research field, or at least of substantial parts of it. Too often, the history of a field is little known to its present participants. Handbooks and compendia volumes are examples of classical resources with an important role in offering representations of knowledge gained in our field, as well as bases for its present and future activity. Yet we cannot expect that one type of resources on its own meets the challenges of communicating the past and present of mathematics education research. Moreover, written communication of advances in the field does not always emphasize the lived experiences and processes of collaboration that made these advances possible. The AMOR units with video modules around contributions made by leading researchers in mathematics education relate their histories of scholarship to the histories behind the pieces of knowledge gained.

The units of Michèle Artigue, Guy Brousseau, Yves Chevallard, Celia Hoyles and Anna Sfard, in different stages of development from finalized to initiated, are currently available on the AMOR website. The unit of Celia Hoyles (Hans Freudenthal Award 2003) is the most recent, now consisting of the introductory Module O presented by Núria Planas, followed by three modules presented by Celia Hoyles and grouped into "Mathematics education in the digital age: Promise and reality": "Setting the scene" (Module 1); "Putting into practice: A curriculum innovation approach" (Module 2); and "Putting into practice: Programming and computational thinking" (Module 3). Alongside work on started units, the AMOR team within the ICMI Executive Committee works on the preparation of newer units. In deciding who to work with, we consider the award year, but also the time availability of the awardees or of close collaborators.

It would be unfair not to give a brief overview of the other units. Anna Sfard's unit has just seen unveiled three more modules on mathematical objects and routines, which are constitutive elements of her influential commognitive approach to mathematics education. Michèle Artigue's unit with nine modules and Module O presented by Jean-Luc Dorier is as much finished as it can be a unit of a researcher who continues to be strongly active after so many decades of incredible work. Jean-Luc Dorier also introduces Guy Brousseau's unit, which together with the six modules presented by Claire Margolinas and Annie Bessot show the impact size of the theory of didactical situations in mathematics. Last but not least, we find Yves Chevallard's unit, with all modules presented by Marianna Bosch, except the introduction (by Jean-Luc Dorier) for communication of how and how much the anthropological theory of the didactic has moved the research field forward. With the involvement of awardees and colleagues, ICMI AMOR keeps going and growing.

Now it is important to use the potentialities of the resources provided by AMOR as much as possible, and as much as needed. In any institution with a PhD and/or Master courses in mathematics education, we could for example draw on these materials to complement the learning experience of the young researchers or newcomers to the field who are in the process of understanding the complexities of doing mathematics education research. The learning opportunities that the use of AMOR can create are not however limited to young researchers; watching the modules can be equally inspiring for mature researchers in the field and for those in the field of mathematics, too. The scientific specificity of the mathematics education research field is wonderfully illustrated through the creative voices of expertise and dedication. Moreover, in a lighter way, some parts of the units can be used in some contexts of mathematics teachers' training.

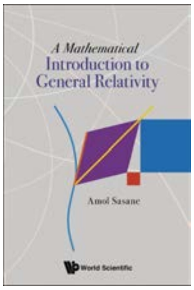
At present, the available AMOR units are a result of invitations and acceptances, and does not necessarily provide broad or strategically intended representations of the research field. We are interested in expanding the project in the direction of broader representations of mathematics education research with awardees from different world regions, and possibly within more diverse theoretical traditions. As it is now, nonetheless, the number of accounts is large enough to be considered a reliable representation of the quality and progress of mathematics education research over the last decades. We thus very much welcome you to visit the ICMI AMOR website and experience the different units. You can send your comments (bad or good) to Jean-Luc Dorier at icmi.secretary.general@mathunion.org.

Núria Planas is professor of mathematics education at Autonomous University of Barcelona, Spain, and honorary research fellow at the Department of Education at the University of Oxford. Her research focuses on multilingual mathematics learning, language use in classroom mathematics teaching, and sociocultural theories of mathematics education. She is a Member-at-Large of the ICMI Executive Committee.

nuria.planas@uab.cat

A Mathematical Introduction to General Relativity
by Amol Sasane

Reviewed by José Natário



Although general relativity is a highly mathematical theory, and arguably one of the main drivers behind the development of Riemannian geometry in the last 100 years, there are relatively few introductory books on this subject that specifically target mathematicians. The book under review is a welcome addition to this scant literature, aiming to introduce Einstein's theory, as well as the needed differential geometry, in a fully rigorous manner. It is interesting to note that the author, a professor of mathematics at the London School of Economics, is not an expert in general relativity, and so is in an ideal position to connect with mathematicians who are encountering the theory for the first time.

The book starts by developing the main ideas of differential geometry, and then goes on to discuss general relativity. It is carefully written, containing numerous appealing figures, and averaging more than ten exercises per chapter (with full solutions provided in an appendix, which is ideal for autonomous study). Moreover, many of the examples and exercises in the differential geometry part are calculations in general relativity (where the author supplies the relevant metrics to be derived in later chapters), which no doubt will appeal to the reader eager to learn general relativity. The level is more elementary than that of other books written in the same mathematical vein, such as "General Relativity for Mathematicians" by Sachs and Wu (which already assumes the differential geometry background), or "Semi-Riemannian Geometry" by O'Neill, and is well suited for mathematics or mathematically inclined physics undergraduate or beginning graduate students.

The detailed plan of the book is as follows: smooth manifolds and smooth maps are introduced in Chapter 1, without assuming point set topology (indeed the prerequisites of the book are simply the usual linear algebra, multivariate calculus and differential equations courses common to most degrees in mathematics, physics or engineering). Chapter 2 discusses tangent vectors, and Chapter 3 studies vector fields. General (mixed) tensor fields are defined in Chapter 4, and semi-Riemannian (in particular Lorentzian) manifolds are introduced in Chapter 5. The Levi-Civita connection, parallel transport and geodesics are discussed in Chapters 6, 7 and 8, respectively, and the notion of curvature is addressed in Chapter 9. Chapters 10 and 11 constitute a digression into differential forms and integration, including the Hodge star (later used to formulate the Maxwell equations); this is a subject not covered in many introductory general relativity books (e.g. O'Neill's "Semi-Riemannian Geometry"). The relativity part of the book starts in Chapter 12 with a discussion of physics in Minkowski spacetime, including a detailed analysis of relativistic velocity addition and electromagnetism. Chapter 13 gives a geometric reformulation of Newtonian gravity and defines the relativistic energy momentum tensor, motivating the introduction of the Einstein field equation in Chapter 14. This chapter also contains a derivation of the Schwarzschild metric and the calculation of the perihelion precession. Chapter 15 introduces black holes, including the Kruskal extension of the Schwarzschild solution, and Chapter 16 briefly discusses cosmology.

On the whole, the book does a good job of introducing differential geometry and general relativity in a mathematically rigorous fashion. It can be used as the textbook for a course on either differential geometry or general relativity (or both) for undergraduate or beginning graduate mathematics or physics students, and is also well suited for autonomous study. My one criticism of the book would be that, after making it through the differential geometry part, the reader should perhaps be rewarded with more general relativity. For example, the discussion of differential forms and electromagnetism in chapters 10, 11 and 12 is nicely followed up by a discussion of the Reissner–Nordström charged black hole solution in Chapter 15, but only as an exercise, with no further exploration of the rich geometry of this spacetime. Other topics of current mathematical and physical interest, such as the linearized Einstein equations, gravitational waves, or the ΛCDM cosmological model for our universe, are likewise only addressed in the exercises, and some other topics, such as the singularity theorems or the

Cauchy problem for the Einstein equations, are not addressed at all. While it is of course unrealistic to ask for a detailed treatment of all these subjects, especially in a book for undergraduates, more steps in that direction could perhaps have been taken. Nevertheless, these small quibbles should not take away from the fact that this book is a valuable addition to the general relativity literature for mathematicians, and one which I highly recommend.

Amol Sasane, *A Mathematical Introduction to General Relativity.* World Scientific, 2021, 500 pages, Hardback ISBN 978-981-124-377-6, eBook ISBN 978-981-12-4379-0.

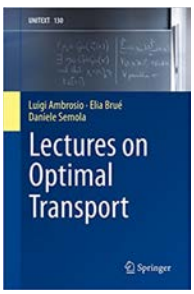
José Natário is a full professor at the Mathematics Department of Instituto Superior Técnico (Universidade de Lisboa). He has authored more than 50 research papers and three books, mostly on geometry and mathematical relativity.

jnatar@math.tecnico.ulisboa.pt

DOI 10.4171/MAG/75

Lectures on Optimal Transport
by Luigi Ambrosio, Elia Brué and Daniele Semola

Reviewed by Filippo Santambrogio



This is the first of the two books that I am reviewing for this issue of the EMS Magazine. It is a textbook on optimal transport (in the same spirit of a book I published in 2015 [9], or of the two books by Cédric Villani [11, 12]), meant yo be used by graduate students. The first author is one of the leading experts on the topic, who has been giving lectures on it for decades at SNS Pisa (by the way, it is in the course that he gave exactly 20 years ago that I started learning about optimal transport). The second and third authors are two of the brilliant students who attended these courses in Pisa.

The book is organized into 19 chapters, each meant to correspond to a single lecture. The duration of a single lecture is not suggested explicitly, but I find the rhythm a little bit slow for graduate students, as I usually cover the material of the first 6 or 7 lectures in approximately 6 hours. Regardless, the idea of organizing the presentation according to teaching time is a very useful pedagogical tool.

The 19 lectures can be roughly divided into four series. Lectures 1 to 7 are essentially devoted to the main theory of the Monge and Kantorovich problems, where two measures are fixed and one looks for the optimal plans or maps to transport the first measure onto the second at minimal cost. At the beginning the cost function is as general as possible, which allows to develop the whole Kantorovich theory, including existence of optimal plans and duality. Only in the last of these lectures the focus is on some precise Euclidean examples, and in particular on the quadratic cost, together with its connections with the Monge–Ampère equation (whose name is spelled correctly all along the book, except for the title of the corresponding lecture where, unfortunately, we can see an acute accent). Another very natural cost, the distance cost originally studied by Monge, is deliberately discussed for only a single page, since it is clearly the goal of the authors to move on to some notions, in connection with PDEs and differential geometry, that are more related to the quadratic cost. Some choices in the proofs or in the presentation could be debatable, for instance regarding duality: the authors do present, shortly, a proof based on rather general convex analysis (the Fenchel–Rockafellar theorem), but devote more space to a full and self-contained proof based on the c-cyclical monotonicity of the support, arguing that it is more constructive, which is absolutely true. On the other hand, this approach might suggest the wrong idea that each optimizer in the Kantorovich problem is associated with a specific maximizer of the dual (the one built from the support of this very optimizer) and this can be seen in the (absolutely classical) proof of uniqueness of optimal transport maps. This proof is based on the clever statement that if every optimal plan is induced by a map, then it is unique, but does not exploit the fact that the map corresponding to a plan can be chosen to be the same for all plans.

After the general presentation of the optimal transport problem, a second series of lectures (8–10) on the Wasserstein distances and Wasserstein spaces follows. Here the authors do a remarkable work by systematically analysing which metric properties of a metric space (X, d) are inherited by the corresponding Wasserstein space $(\mathcal{P}(X), W_2)$ (we see that the focus is explicitly on the case $p = 2$, in order to pave the way for the next part of the book): compactness, completeness, geodesics, … Some parts require the introduction of suitable tools from analysis in metric spaces, in particular the notion of metric derivative, which are independent of optimal transport, but not always well known among graduate students in analysis.

Similarly, the next series of lectures (11–14) is not specifically related to optimal transport: it is devoted to a detailed analysis of gradient flows in Hilbert spaces, paying attention to those notions which can be extended to metric spaces, and in particular the EVI (Evolution Variational Inequality) and the EDI (Energy Dissipation Inequality) formulations. The role played by convexity or λ-convexity is emphasized from the very beginning. A full chapter is devoted to the study of the heat flow as a gradient flow with different choices of the functional and of the Hilbert norm (the heat flow is, for instance, the gradient flow of the Dirichlet energy $u \mapsto \frac{1}{2}\int|\nabla u|^2$ in the L^2 space, but also of the simplest functional $u \mapsto \frac{1}{2}\int u^2$ in the homogeneous H^{-1} space). This is very useful, enabling the reader to realize that one and the same equation can be seen as a gradient flow in many different ways, and that there is an interplay between the functional and the distance (one needs to change both if one is looking for the same equation).

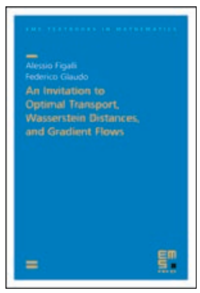
After this detailed discussion about gradient flows, the authors come back for the last lectures (15–19) to the Wasserstein space W_2. First, a long chapter is devoted to the study of various functionals on the space of probability measures and on their variational properties (in particular, lower semicontinuity and geodesic convexity). This recalls what I did in a chapter of [9] and I am glad to see that the authors share my feeling that, in order to address some applications of optimal transport, at some point it is absolutely necessary to clarify what we know and what we should know about the most used functionals. The next lectures are mainly devoted to curves of measures, with a detailed discussion of the continuity equation $\partial_t \mu + \nabla \cdot (\mu v) = 0$, and of the specific case of geodesic curves (for which the velocity field v is related to the gradient of the solution of a Hamilton–Jacobi equation solved by means of the Hopf–Lax semigroup). This is followed by the dynamical formulation of optimal transport proposed by Benamou and Brenier [4] and the characterization of "nice" curves in W_2 as solutions of the continuity equation with L^2 vector fields. I find it unusual to follow this order (I usually consider generic curves and then optimal ones, even if admittedly also in my presentation geodesics are used to build a velocity field), but the exposition is perfectly coherent and clear. I also note a very nice proof of the semicontinuity of the Benamou–Brenier energy based on the interpretation of the L^p norm as dual to the corresponding L^q norm. The last two lectures in this series concentrate more on the heat flow, proving that the solutions of the heat equation are metric EVI gradient flows in the Wasserstein space of the entropy functional. This is the meaning which is given to being a gradient flow in this approach, coherently with the famous book of the first author with N. Gigli and G. Savaré [2]. Since the procedure consists here in checking that an existing solution of a well-known PDE satisfies this notion, this can be seen as an interpretation tool, and not as a way to prove existence of solutions to some PDE with a gradient flow structure in W_2 (in particular, the celebrated Jordan–Kinderlehrer–Otto scheme [7] is only marginally mentioned). Finally, it is shown that the behaviour of the heat flow on Riemannian manifolds is strongly related to conditions on the curvature and to the geodesic convexity of the entropy functional, a fact that has been used as a starting point for a synthetic definition of the notion of curvature bound in metric measure spaces [8, 10] and then for a very general theory of calculus in metric measure spaces [3].

Even if some points in the exposition differ from what I would have done – and I had fun in pointing this out – there is no doubt in my mind that this very well written 250-page book can be an extremely useful tool to teach optimal transport classes or, for a more experienced researcher from a related but different field, to access the subject. It contains both heuristic discussions which could be completed by external reading and rigorous proofs, and covers a large part of the existing theory. Covering the entire theory was of course impossible, and the choice was made to focus on topics that better prepare the reader to deal with some mathematical applications, in particular in connection with differential geometry and partial differential equations from an abstract point of view, and the book is clearly meant for a public of analysts or geometers. And it contains many clever ideas and tricks that the reader will appreciate and re-use in his own work!

Luigi Ambrosio, Elia Brué and Daniele Semola, *Lectures on Optimal Transport.* Springer, 2021, 259 pages, Paperback ISBN 978-3-030-72161-9, eBook ISBN 978-3-030-72162-6.

An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows by Alessio Figalli and Federico Glaudo

Reviewed by Filippo Santambrogio



We now move on to the second book I am reviewing for this issue. It is also a text on optimal transport, and it is also produced by researchers from the same school (i.e., the Italian school on calculus of variations centred at SNS Pisa). This book aims at being a self-contained introduction to optimal transport and some of its applications, with a quite explicit goal to acquaint the reader with the theory and invite her/him to start working on and with it, possibly looking for more detailed developments or proofs elsewhere.

Optimal transport is a very active field and every new monograph that can attract colleagues from related disciplines or can help researchers who need to understand it after a first encounter, is more than welcome. One of the key advantages of this short monograph is the authorship, since one of the authors is well known in the whole mathematical world because of the Fields medal he was awarded exactly for his work on optimal transport: this is very likely to attract more readers than any other manuscript on the topic.

The first important point to notice is the book's size, approximately 130 pages. It is much shorter than other references on the topic, and in this respect it is difficult to compare it to the book by Ambrosio, Brué and Semola or to other classical references [9, 11, 12]. The closest manuscript that one can use as a comparison should probably be *A user's guide to optimal transport* by Ambrosio and Gigli [1], even if inviting readers to discover a topic is not exactly the same as the claimed goal to offer a guide to understand, or at least use, it. In comparing the two texts one has to note that Ambrosio and Gigli's work is in the end only partially about optimal transport, being instead heavily oriented towards metric measure spaces. By contrast, Figalli and Glaudo's work remains focused on optimal transport, which allows for a more complete and useful presentation, despite the small length.

In comparing to classical books, the reference is always the first book by Villani [11], since my own text [9], published 12 years later, aimed at adding developments of the theory that did not exist in 2003, and Villani's second book [12], includes several hundreds of pages on new extensions and connections, in particular in the direction of differential geometry. From this point of view, the present manuscript does not aim at covering new material, as the core of its exposition concentrates on topics already present in [11], and new developments are only worth a few lines in the further reading part. This is a very legitimate choice if one wants to keep the presentation short as well as reasonably self-contained. On the other hand, I would say that, despite aiming at a slightly more pure-math oriented audience, this book shares a "concrete" flavour with [9].

As a short introductory text, the book is composed of only five chapters, and the last one, called "further reading" honestly discusses the other existing references on the topic, and some extensions or connections. Chapter 1 also plays a different role than the others, including some examples of transport maps, some applications (for instance, how to prove the isoperimetric inequality using the Knothe map), and some preliminary background material. The core of the book thus consists of Chapters 2, 3, and 4.

Chapter 2 is devoted to the already classical theory of optimal transport, following more or less the same structure as that of the book by Ambrosio, Brué and Semola (including very similar approaches to duality and to the uniqueness of optimal maps), even if an important role is given to the cost $c(x, y) = -x \cdot y$, which is equivalent to the quadratic cost $\frac{1}{2}|x-y|^2$ and allows for a direct use of convex analysis without the need to introduce c-convexity (or c-concavity). If general costs arrive first in what concerns existence of optimal plans, they appear later in what concerns Kantorovich duality. In this same chapter we can also praise the detailed discussion of the various connections of optimal transport with the incompressible Euler equation (which also allows to underline the multiple roles Yann Brenier played in the theory of optimal transport, see [4–6]).

Chapter 3 and Chapter 4 both include at the same time the second and the third key concepts evoked by the title: Wasserstein distances and gradient flows. Chapter 3 is more metric in nature: it introduces the Wasserstein distances W_p (for every p) and after a short (a few pages, not a few chapters) digression on Hilbertian gradient flows, moves on to the Jordan–Kinderlehrer–Otto scheme (JKO [7]), thus attacking gradient flows via a sequence of iterated minimization problems involving the W_2 distance. This chapter treats only the case of the heat equation, which is the simplest one, but has the drawback of being also a gradient flow in L^2, differently from the Fokker–Planck equation with a potential V which is dealt with in Chapter 4. The main theorem here states that the limit as the time step τ of the JKO scheme tends to 0 is a distributional solution of the heat equation, and in this sense I consider this presentation as more "concrete": it finds a solution – in a very standard sense – to a PDE, and not to a metric condition such as the EVI or the EDI definitions of gradient flows, which could sound unnatural as a definition when PDE tools are available.

Chapter 4 then goes on with the differential and Riemannian structure of the Wasserstein space, discussing geodesic curves and the Benamou–Brenier formula, introducing Otto's calculus in order to endow this space with a formal notion of tangent space and make it a sort of infinite-dimensional Riemannian manifold, and then presenting the notion of geodesic convexity. In a way that I strongly approve, geodesic convexity is shown to be crucial for the study of gradient flows in what concerns finding properties of their solutions, but it is not at all evoked when it comes to proving existence for some PDEs. As an example, the authors then concentrate on the Fokker–Planck equation, adding a convex potential to the heat flow, and prove a series of inequalities which allow to obtain well-known rates of convergence to the steady state for strongly convex confining potentials. They finish the chapter proving convergence to the same steady state in the strong L^1 sense (and not only in the Wasserstein sense, which means weak convergence), providing a very nice proof of a suitable functional inequality (the Csiszár–Kullback–Pinsker inequality; my only criticism here is that the authors claim that a certain step of the proof, establishing that a certain function is negative on the boundary in order to apply later a sort of maximum principle, is "easy", while it required me some work to reach this conclusion).

Besides Chapters 1–5, the book also contains two appendices, both including exercises. Appendix B aims at providing a proof of the disintegration theorem in measure theory via a series of guided exercises, while Appendix A is a collection of 11 fully solved exercises (though I think many readers would have liked to see more exercises than these 11).

Book reviews are not meant for authors, but for potential readers, but in case the authors will read this review they will probably realize, in view of the similarity of some comments and sentences, that I also refereed their manuscript before publication, which means that I had more than one occasion to look at their work (further occasions also include a student whom I supervised who decided to build up her knowledge of optimal transport on this very book). I must say that I get more and more convinced every time I take a closer look: yes, I like this book. It does the job of inviting readers to the field, and it does it well.

Alessio Figalli and Federico Glaudo, *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press, 2021, 144 pages, Hardback ISBN 978-3-98547-010-5, eBook ISBN 978-3-98547-510-0.

References

[1] L. Ambrosio and N. Gigli, A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, Lecture Notes in Math. 2062, Springer, Heidelberg, 1–155 (2013)

[2] L. Ambrosio, N. Gigli and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics, ETH Zürich, Birkhäuser, Basel (2005)

[3] L. Ambrosio, N. Gigli and G. Savaré, Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. *Invent. Math.* **195**, 289–391 (2014)

[4] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84**, 375–393 (2000)

[5] Y. Brenier, Décomposition polaire et réarrangement monotone des champs de vecteurs. *C. R. Acad. Sci. Paris Sér. I Math.* **305**, 805–808 (1987)

[6] Y. Brenier, The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc.* **2**, 225–255 (1989)

[7] R. Jordan, D. Kinderlehrer and F. Otto, The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29**, 1–17 (1998)

[8] J. Lott and C. Villani, Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)* **169**, 903–991 (2009)

[9] F. Santambrogio, *Optimal transport for applied mathematicians*. Progress in Nonlinear Differential Equations and their Applications 87, Birkhäuser/Springer, Cham (2015)

[10] K.-T. Sturm, On the geometry of metric measure spaces. I and II. *Acta Math.* **196**, 65–131 and 133–177 (2006)

[11] C. Villani, *Topics in optimal transportation*. Grad. Stud. Math. 58, AMS, Providence (2003)

[12] C. Villani, *Optimal transport*. Grundlehren Math. Wiss. 338, Springer, Berlin (2009)

Filippo Santambrogio is professor at Université Claude Bernard Lyon 1, where he moved after spending twelve years in the Paris area (in Dauphine and Orsay), and after his studies at SNS Pisa. His research focuses on optimal transport and calculus of variations, both in what concerns the general theory and in their applications to the modeling of traffic systems, crowd motion and optimal location of resources.

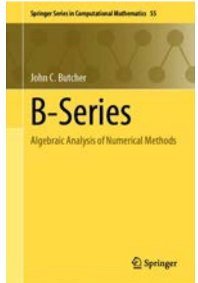
santambrogio@math.univ-lyon1.fr

DOI 10.4171/MAG/88

B-Series: Algebraic Analysis of Numerical Methods

by John C. Butcher

Reviewed by Raffaele D'Ambrosio



The algebraic analysis of numerical methods is driven along a bridge matching numerical analysis, graph theory, group theory, differentiation of vector fields, and so on. The result of the wise union of these aspects is an effective and elegant theory that is useful for investigating meaningful features of numerical methods for differential equations. One of the pioneering sources of this theory is the brilliant and ingenious work of John C. Butcher. He has authored a remarkable number of seminal contributions to the founding, establishment and development of the modern theory of Runge–Kutta methods and its subsequent extensions.

A building block for the algebraic analysis of numerical methods and, in particular, of their order of convergence, is the connection between rooted trees and the differentiation of vector fields. Such a link lies at the basis of Butcher's theory and its precursors[1], such as Arthur Cayley (1821–1895) and Robin H. Merson (1921–1992), a scientist at the Royal Aircraft Establishment, who became popular for his involvement in the computations of an accurate orbit for Sputnik 1, launched in 1957. In the same year, Butcher attended the talk by Merson at the conference "Data Processing and Automatic Computing Machines" held in Salisbury, South Australia, where Merson described the one-to-one correspondence between derivatives and rooted trees. The full theory will only be provided later by Butcher, but it is worth mentioning that Cayley introduced trees with the same purpose as in Butcher theory (namely, to understand and effectively represent the interaction of vector fields repeatedly applied to one another), and then for one century this aspect was totally forgotten in the literature, and reconsidered with effectiveness only when the theory of numerical methods was established with rigor.

Other two building blocks for the algebraic analysis of numerical methods, allowing to detect and elegantly prove properties of numerical methods for differential problems, are well described in the book under review, namely

- Butcher series (in short, B-series), allowing to represent both exact and numerical solutions to a differential problem in terms of series expansions whose coefficients are functions of rooted trees of a prescribed order;

[1] Historical aspects have been described in detail, for instance, in the paper by R. McLachlan et al., *Asia Pacific Math. Newsletter* 7(1), 1–11 (2017).

- the Butcher group (in short, B-group), obtained by equipping the set of homomorphisms of the Hopf algebra of rooted trees into the set of real numbers with an operation of composition of mappings. As mentioned above, the analysis of the group properties of this algebraic structure allows to treat accuracy properties of numerical methods in a very elegant and effective way.

An immediate benefit of the B-series/B-group theory (originally developed for Runge–Kutta methods and subsequently extended to multivalue numerical methods) is that one can achieve a minimal number of conditions for the construction of high-order methods. However, confining the effectiveness and importance of this theory to the sole construction of efficient numerical methods would be somewhat reductive. Indeed, as pointed out by A. Connes and D. Kreimer, the Butcher group had arisen independently in their own works on renormalization in quantum field theory.

Characterized by a very clear and self-contained style, the monograph under review is an excellent contribution, consisting of seven chapters, enriched by exercises and their solutions, study notes, and open-ended projects. Chapter 1 provides a collection of topics useful to contextualize the mathematical problem (i.e., systems of differential equations), its well-posedness, and the ideas behind multistage discretizations, namely Runge–Kutta and multivalue numerical methods. The basic accuracy requirements of consistency, stability and convergence are also recalled, as well as some of the topics more widely explained in later chapters. Chapter 2 contains a comprehensive presentation of trees, forests, and operations on them. The exposition is focused on the basic terminology for trees and forests, their graphical structures, the operations useful to build up trees starting from a single node, how partitions act on trees, the ideas of evolution, stump and antipode, with a close relationship, in several cases, with Hopf-algebra terminology.

The core material of the treatise, oriented to the way trees and forests are useful in the numerics for differential equations, starts in Chapter 3, where the notion of B-series (and composition of B-series) is introduced in detail. B-series allow to write the Taylor expansions of the two mappings describing the exact and numerical solutions, respectively, to the problem via elementary differentials, which can be effectively represented in terms of rooted trees. Chapter 4 relates B-series to the algebraic analysis of numerical methods, with particular reference to the properties of the B-group for the analysis of Runge–Kutta methods, explained in depth in Chapter 5. Specifically, the author describes a very elegant and effective way to provide order conditions for Runge–Kutta methods, analyzing a wide selection of explicit and implicit schemes. The treatise moves, through Chapter 6, to the family of general linear methods, characterized by a combined multivalue and multistage strategy, and explaining the way the B-series approach effectively applies to these methods. Chapter 7 gives interesting insights regarding the application of the B-series approach to the development and analysis of numerical schemes useful for solving Hamiltonian problems, in the spirit of the so-called geometric numerical integration, here oriented to analyzing symplectic Runge–Kutta methods, as well as developing multivalue numerical schemes with near-preservation of invariants.

The author uses a very clear writing style, and each chapter contains an outline and summary, many examples, and a final section of way forward, giving a perspective how the treatise goes on. The approach is self-contained and different levels of in-depth analysis intersect, making the book ideal both for advanced courses of numerical analysis and as a reference monograph in research. Moreover, the large set of applications involving B-series makes the book under review a key reference for all scientists who are effectively using computational modelling for evolutive problems in their work.

John C. Butcher, *B-Series: Algebraic Analysis of Numerical Methods.* Springer, 2021, 320 pages, Hardback ISBN 978-3-030-70955-6, eBook ISBN 978-3-030-70956-3.

Raffaele D’Ambrosio is full professor of numerical analysis at the Department of Information Engineering and Computer Science and Mathematics at the University of L’Aquila and former Fulbright Research Scholar. His research mostly focuses on deterministic and stochastic geometric numerical integration as well as the study of long-time behavior of numerical methods for functional problems.

raffaele.dambrosio@univaq.it

DOI 10.4171/MAG/87

Letter of Ukrainian mathematicians

Dear Editor,

On behalf of Ukrainian mathematicians, we would like to thank the EMS for the statement of solidarity published on March 2. At the same time, we are disappointed by the fact that the corporate membership of the two Russian mathematical societies, Moscow and St. Petersburg, wasn't suspended.

The full-scale Russian invasion of Ukraine has been going on for almost a month. Every new day brings immense suffering and losses. The Russian troops destroy our country, they bomb hospitals, kindergartens, schools, and universities. The cruelty against civilians leaves us breathless.

We are aware of the individual protests of the Russian mathematicians against the war. We appreciate the courage of those Russian colleagues who expressed their anti-war position publicly. Nevertheless, the majority of the Russian population continues to support the genocide against Ukrainian people. They are being spoiled by the mass media propaganda and cut from credible sources of information. In these conditions, it is important to reach the Russian people by all possible means, especially through the cultural and academic communities who may influence the public opinion in the country.

There are no messages about the war and expressions of solidarity with the victims at the web pages of the Mathematical Societies of Moscow and St. Petersburg. The learned societies in Russia continue to function as if nothing has happened. At the same time, rectors of all Russian universities signed the infamous letter supporting the invasion of the Russian troops in Ukraine. Publicly, the academic community in Russia is demonstrating support for the criminal actions of the Russian government.

On March 6 a number of Ukrainian mathematicians wrote a letter to the leadership of the EMS asking them to suspend the corporate membership of the two Russian mathematical societies. So far only the membership of the Euler International Mathematical Institute was interrupted.

Therefore, we call once again for the suspension of the membership of the Moscow and St. Petersburg Mathematical Societies in EMS and request the EMS Magazine to publish this open letter.

Kind regards,
Yuriy Drozd, President of the Ukrainian Mathematical Society
Rostyslav Hryniv, President of the Lviv Mathematical Society
Olena Karlova, President of the Chernivtsi Mathematical Society
Sergiy Maksymenko, President of the Kyiv Mathematical Society

Comment on the above letter

In the last 30 years considerable efforts have been made to unite the European mathematical community, which was very much divided until the iron curtain fell 30 years ago. Unfortunately, with the aggressive politics of the Russian government in the last years, the occupation of the Crimea peninsula and the Donbas region, culminating in the Russian war against the Ukraine, the unity achievements are at stake.

Many mathematicians die in this war or become refugees. The EMS declares full solidarity with the Ukrainian people in their fight against the invasion of their country.

But the EMS is also in solidarity with those in Russia who protest against the war and are arrested or have to leave their country.

We are aware that it is very difficult for those who are bombshelled to understand and accept that the continuing cooperation with colleagues and mathematical societies from Russia is still important and necessary, and even if some mathematicians support the Russian government, we should leave the door open for those who despise these aggressive acts and still share the vision that mathematics research and teaching are ideology-crossing and boundary-crossing activities that should not be taken hostage or be used as a political instrument.

Volker Mehrmann
President of the EMS

SwissMAP Research Station (SRS)

2024 call for proposals

The SwissMAP Research Station (SRS) is a joint venture between the University of Geneva and ETH Zurich, organizing international topical conferences and targeted workshops in the fields of mathematics and theoretical physics. It was inaugurated in 2021 and is based in the Swiss Alps, in Les Diablerets. The station has fully equipped conference and meeting rooms.

As many of the events are recorded, a comprehensive video collection of talks covering different topics is available through the SRS website, where you can also subscribe to the mailing list if you wish to keep up to date with the SRS scientific programs and calls for proposals.

Interested in organizing a conference in 2024 at the SwissMAP Research Station?
The 2024 call for proposals is now open. **Application deadline: September 30, 2022**

SRS
SwissMAP Research
Station in Les Diablerets

More information available on the SRS Website
https://swissmaprs.ch





European Mathematical Society

EMS executive committee

President

Volker Mehrmann (2019–2022)
Technische Universität Berlin, Germany
mehrmann@math.tu-berlin.de

Vice presidents

Betül Tanbay (2019–2022)
Bogazici University, Istanbul, Turkey
tanbay@boun.edu.tr

Jorge Buescu (2021–2024)
University of Lisbon, Portugal
jbuescu@gmail.com

Treasurer

Mats Gyllenberg (2015–2022)
University of Helsinki, Finland
mats.gyllenberg@helsinki.fi

Secretary

Jiří Rákosník (2021–2024)
Czech Academy of Sciences, Praha, Czech Republic
rakosnik@math.cas.cz

Members

Frédéric Hélein (2021–2024)
Université de Paris, France
helein@math.univ-paris-diderot.fr

Barbara Kaltenbacher (2021–2024)
Universität Klagenfurt, Austria
barbara.kaltenbacher@aau.at

Luis Narváez Macarro (2021–2024)
Universidad de Sevilla, Spain
narvaez@us.es

Beatrice Pelloni (2017–2024)
Heriot-Watt University, Edinburgh, UK
b.pelloni@hw.ac.uk

Susanna Terracini (2021–2024)
Universita di Torino, Italy
susanna.terracini@unito.it

EMS publicity officer

Richard H. Elwes
University of Leeds, UK
r.h.elwes@leeds.ac.uk

EMS secretariat

Elvira Hyvönen
Department of Mathematics and Statistics
P.O. Box 68
00014 University of Helsinki, Finland
ems-office@helsinki.fi

Join the EMS

You can join the EMS or renew your membership online at euromathsoc.org/individual-members.

Individual membership benefits

- Printed version of the EMS Magazine, published four times a year for no extra charge
- Free access to the online version of the *Journal of the European Mathematical Society* published by EMS Press
- Reduced registration fees for the European Congresses
- Reduced registration fee for some EMS co-sponsored meetings
- 20 % discount on books published by EMS Press (via orders@ems.press)*
- Discount on subscriptions to journals published by EMS Press (via subscriptions@ems.press)*
- Reciprocity memberships available at the American, Australian, Canadian and Japanese Mathematical Societies

* These discounts extend to members of national societies that are members of the EMS or with whom the EMS has a reciprocity agreement.

Membership options

- 25 € for persons belonging to a corporate EMS member society (full members and associate members)
- 37 € for persons belonging to a society, which has a reciprocity agreement with the EMS (American, Australian, Canadian and Japanese Mathematical societies)
- 50 € for persons not belonging to any EMS corporate member
- A particular reduced fee of 5 € can be applied for by mathematicians who reside in a developing country (the list is specified by the EMS CDC).
- Anyone who is a student at the time of becoming an individual EMS member, whether PhD or in a more junior category, shall enjoy a three-year introductory period with membership fees waived.
- Lifetime membership for the members over 60 years old.
- Option to join the EMS as reviewer of zbMATH Open.