DFT21: Discrete Fourier Transform in the 21st century

Prasanta Pal 1, Shataneek Banerjee 2, Amardip Ghosh 2, David R. Vago 2, and Judson Brewer 2

1Brown University
2Affiliation not available

October 30, 2023

Abstract

Knowingly or unknowingly, digital-data is an integral part of our day-to-day lives. Realistically, there is probably not a single day when we do not encounter some form of digital-data. Typically, data originates from diverse sources in various formats out of which time-series is a special kind of data that captures the information about the time-evolution of a system under observation. However, capturing the temporal-information in the context of data-analysis is a highly non-trivial challenge. Discrete Fourier-Transform is one of the most widely used methods that capture the very essence of time-series data. While this nearly 200-year-old mathematical transform, survived the test of time, however, the nature of real-world data sources violates some of the intrinsic properties presumed to be present to be able to be processed by DFT. Adhoc noise and outliers fundamentally alter the true signature of the frequency domain behavior of the signal of interest and as a result, the frequency-domain representation gets corrupted as well. We demonstrate that the application of traditional digital filters as is, may not often reveal an accurate description of the pristine time-series characteristics of the system under study. In this work, we analyze the issues of DFT with real-world data as well as propose a method to address it by taking advantage of insights from modern data-science techniques and particularly our previous work SOCKS. Our results reveal that a dramatic, never-before-seen improvement is possible by re-imagining DFT in the context of real-world data with appropriate curation protocols. We argue that our proposed transformation DFT21 would revolutionize the digital world in terms of accuracy, reliability, and information retrievability from raw-data.
DFT21: Discrete Fourier Transform in the 21st century

Prasanta Pal, Shataneek Banerjee, Amardip Ghosh, David R. Vago, & Judson A. Brewer

Abstract—Knowingly or unknowingly, digital-data is an integral part of our day-to-day lives. Realistically, there is probably not a single day when we do not encounter some form of digital-data. Typically, data originates from diverse sources in various formats out of which time-series is a special kind of data that captures the information about the time-evolution of a system under observation. However, capturing the temporal-information in the context of data-analysis is a highly non-trivial challenge. Discrete Fourier-Transform is one of the most widely used methods that capture the very essence of time-series data. While this nearly 200-year-old mathematical transform, survived the test of time, however, the nature of real-world data sources violates some of the intrinsic properties presumed to be present to be able to be processed by DFT. Adhoc noise and outliers fundamentally alter the true signature of the frequency domain behavior of the signal of interest and as a result, the frequency-domain representation gets corrupted as well. We demonstrate that the application of traditional digital filters as is, may not often reveal an accurate description of the pristine time-series characteristics of the system under study. In this work, we analyze the issues of DFT with real-world data as well as propose a method to address it by taking advantage of insights from modern data-science techniques and particularly our previous work SOCKS. Our results reveal that a dramatic, never-before-seen improvement is possible by re-imagining DFT in the context of real-world data with appropriate curation protocols. We argue that our proposed transformation DFT21 would revolutionize the digital world in terms of accuracy, reliability, and information retrievability from raw-data.

Index Terms—Signal processing, data analysis, Fourier transform, DFT, FFT, data cleaning, noise reduction, signal recovery, Digital data

I. Introduction

It is the pulse-dynamics of our heartbeat that keep our lives moving on a moment-to-moment basis. The behavior of such dynamical systems can be measured as a time-series of an observable. Time-series analysis [1]–[4] is at the heart of time-dependent data analysis. Various methods have been developed [5], [6] over the past few centuries to address this very intricate problem. Out of all the established methods, the well-cel-
value e.g. Gaussian-filter ($G$), Wiener-filter ($W$), Median-filter ($M$), TotalVariation-filter ($J$), etc. While these traditional filters have certain utilities for smoothing and presentation purposes, however apparent visually appealing smoothness of filtered signals should not be misrepresented as accurate recovery of the underlying pristinity that intrinsically describes the nature of the observable under study.

In the context of time-series analysis, it is to be noted that, FFT does not by itself facilitate any underlying signal recovery process. Rather, it simply transforms measured-signal into the frequency domain often followed by some form of bandpass-filtering, while remaining completely agnostic about the true underlying signature of the raw-data.

Being data-driven is one of the most unbiased ways to learn about physical systems. However, if the data itself is not pre-processed to fit into basic assumptions under which it is processed, the outcomes are inaccurate. We argue that these inaccuracies are not only a margin of error issue but rather something with a dangerous side-effect when used in mission-critical systems like medical-devices or aviation-systems. Some of the basics assumptions around FFT are that the raw-data fed as input to the transform has a simple linear structure in its spectral signature and that there is some measure of equilibrium and ergodicity intrinsically present in the data. However, this kind of presumption is no better than wishing for mere good-luck to perform rigorous, quantitative scientific analysis. We argue that optimal signal-recovery from raw-data through appropriate curation procedure should precede traditional FFT operations. In this work, we not only developed a systematic method to obtain scientifically more accurate outcomes from FFT operations but also propose to extend our methods to build a data-discovery-driven new generation of data-curation and filtering pipelines. By taking advantage of very high level but close to metal intrinsically multi-threaded and natively-supported by the programming language interfaces like (C++ >= -std-11) and Rust, etc. We anticipate a future revolution in data-filtering techniques grounded on fundamental data-discovery principle. The discovery step would prepare the data to be ready for processing by idealized traditional filters. The primary insight here is, to make an initial $O(N)$ data discovery step (e.g. noise, outlier identification) and curation step throughout the entire domain of the data followed by traditional processing techniques.

In the previous generations of filters, this step got overlooked possibly to avoid computational costs. But fortunately, we can afford that overhead now due to the lower overall cost of computation both in personal and enterprise settings. Also, intrinsic multi-threaded processing capabilities along with availability of multi-core, modern hardware and programming language interfaces provides us with a new realm of accurate filtering. We claim that, this would lead to a paradigm shift in the development of the next generation of advanced, more robust data-filters.

II. Method

The method primarily consists of three steps, a) data source simulation b) traditional data-filteration c) FFT21 data curation followed by comparison of results in signal and frequency domains.

A. Data source simulation

In order to demonstrate the signal-recovery process of arbitrarily contaminated data we choose a simple harmonic function as in Eq. 1 with $\omega = 3$ to generate a set number of $N \approx 200$ simulated sample points ($S(f(t), N)$) in the time-domain range $0 \leq t \leq 2\pi$ that we treat as raw input-data ($n$). As a simple illustration to demonstrate the proof of principle case study of data-curation, we generated pristine, uncontaminated data using the following harmonic function as in Eq. 1.

$$f(t) = \sin(\omega t)$$

For the sake of simplicity, we choose $\omega = \omega_0 = 3$ and call it the resonance frequency in this context. The basic reasoning for the simple choice of $f(t)$ in Eq. 1 is that, a simple analytical harmonic function is easy to sample, make visual observations for features (or differences w.r.t reference signal), the frequency-spectrum is precisely known (e.g. $\omega_0$). Any contamination of this function (or signal) followed by filtering (or de-contamination) should recover $f(t)$ precisely. A signal contaminant is added to the pristine function $f(t)$ in the form of a Weiner process (or for all practical purposes Gaussian random number generator) as demonstrated in Eq. 2.

$$\tilde{f}(t) = f(t) + W_t$$

$\tilde{f}(t)$ is then sampled ($S(\tilde{f}(t), N)$) with the same number of sample points ($N \approx 200$) as that of $f(t)$. $S(f(t), N)$ and $S(\tilde{f}(t), N)$ are plotted in Fig. 1 and Fig. 2 respectively. Next step is to recover the underlying pristine signal from $S(\tilde{f}(t), N)$ using traditional filters to see how much restoration is possible.
Fig. 1: Plot of the harmonic function $f(t) = \sin(\omega t)$ for $\omega = 3$

Fig. 2: Plot of the contaminated harmonic function from Eq. 2 for $\omega = 3$

Fig. 3: Comparison of the effect of various standard de-noising filters (a) Gaussian, (b) Wiener, (c) median (d) TotalVariation filter respectively. The kernel-size $\mathcal{K} = 5$ for (a)-(c) is and $\mathcal{K} = 0.3$ for TotalVariation filter.

Fig. 4: As a matter of fact, it performs much worse in comparison to small kernel size filter.

B. Traditional filtering

Traditional data filters have been widely used in the context of signal smoothing and recovery. Some common filters ($\mathcal{F}_k$) are Wiener-filter ($\mathcal{W}$), Gaussian-filter ($\mathcal{G}$), Median-filter ($\mathcal{M}$), and TotalVariation-filter ($\mathcal{T}$). It is reasonable to see how the application of filters affects the recovery process for various filter parameters e.g. Kernel size ($\mathcal{K}$).

Fig. 3 is a demonstration of the effect of various de-noising filters on contaminated signal (simulated). Fig. 4 is a demonstration of the effect of various de-noising filters with relatively bigger filter kernel size applied on contaminated signal (simulated). Ideally, for a given plot label, the filtered data should collapse with the pristine signal after the application of a given filter $\mathcal{F}_k$. However, it is clear from Fig. 3 (a)-(d), that, although some filters perform better (e.g. median-filter) than others, however, a near collapse is almost impossible to achieve despite comparing with various filters. The situation does not improve if we have a bigger kernel size as shown in

Simultaneously for all the frequency amplitude values, despite testing most of the commonly available filters in the literature the signal recovery is a difficult task. Even if we change the parameters of the filters from smaller to bigger parametric values, the situation does not improve if not get worse. This calls for a dramatic paradigm.
shift and complete rethinking of the problem of getting very accurate FFT outcomes. Our core insight is that the computationally driven data-discovery centric data-curation is the missing link.

C. Data curation

In most practical scenarios of time-series data analysis, we are typically interested in a given frequency-band of interest called bandwidth \( B_{\omega}^{\omega'} \). We argue that, the complimentary bandwidth \( B' \) (e.g., orthogonal to this range) composed of low-pass \( B_{\omega}^{\omega'} \) and high-pass \( B'_{\omega} \) can be curated using some optimal curation-protocol \(^{15}\) (say \( C^{\omega'}) \) such that, the true \(^{15}\) bandwidth \( B_{\omega}^{\omega'} \) is optimally recovered from the raw-signal \((,\)all\( )\.

Fig. 5: Visual demonstration of the differences between the spectrum of original raw-signal, noise-contaminated signal, and filtered-signal. Data point labels, \( L_1, L_1', L_1'' \) \( (L \in \{a, b, c, d\}) \) indicate the zero frequency Fourier amplitudes of pristine-signal, contaminated signal and filtered-signal with respect to a given filter type respectively. Highlighting windows \( W_{1-d}^a, W_{1-d}^b, \) & \( W_{1-d}^c \) represent three representative data points.

\(^{14}\)The details of which is not relevant to this work

\(^{15}\)or as close as possible to the truth

\(^{16}\)typically \( \omega = \frac{\pi}{\#} \)
of simplicity of our discussion, we would curate only the high-frequency components as that is where most random noise lies.

The fundamental basis of our proposed methods is first, to scan the entire data set to identify local-fluctuations of values with respect to a specified length (ℒ) of data-records and a pre-defined threshold-parameter (λ) in a recursive fashion for various scales of ℒ. ℒ defines the length-scale (or time-scale or equivalently frequency band) of interest.

The details of the basic idea behind the identification of the points of fluctuations and basic curation steps are outlined in Figs. 3 & 4 in ref. [52] although a future customization of the core-algorithms and parameters would be needed to carve-out a module specific to the DFT problem described here but the details of the module is beyond the scope of this work. In the present context, we primarily demonstrate that, it is as a matter of principle, possible to optimally recover the underlying harmonic signal perturbed by simple noise processes, that is significantly accurate in its Fourier-Transformed space as well as in original signal-space.

III. Results

Fig. 6 demonstrates that the FFT of the curated signal using FFT21 is almost non-discernible from the pristine counterpart. This is dramatic compared to those with traditional filters tested even with different scale parameters as demonstrated in Fig 3 (small-scale) & Fig. 4 (bigger-scale). From Fig 7 we observe that the restored signal makes a visually non-discernible overlap with the pristine frequency amplitudes at almost all sample points except for a few points highlighted by the yellow rectangle. All the error bars are approximately within the symbol sizes.

Table I demonstrates the percent deviation of various

<table>
<thead>
<tr>
<th>sl.</th>
<th>Filter</th>
<th>ω₀</th>
<th>ω’₀</th>
<th>1</th>
<th>1’</th>
<th>2</th>
<th>2’</th>
<th>ω₀ = 3</th>
<th>ω₀’</th>
<th>4</th>
<th>4’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cont.</td>
<td>1658.8</td>
<td>∞</td>
<td>421.9</td>
<td>∞</td>
<td>-1.5</td>
<td>∆ω₀</td>
<td>400.7</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ℜ</td>
<td>1626.1</td>
<td>∞</td>
<td>419.8</td>
<td>∞</td>
<td>-3.7</td>
<td>0</td>
<td>373.7</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ℳ</td>
<td>524.6</td>
<td>∞</td>
<td>241.1</td>
<td>∞</td>
<td>-1.5</td>
<td>0</td>
<td>10.7</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Ⅎ</td>
<td>1903.7</td>
<td>∞</td>
<td>507</td>
<td>∞</td>
<td>-5.9</td>
<td>0</td>
<td>391.6</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>™</td>
<td>2191.7</td>
<td>∞</td>
<td>746.4</td>
<td>∞</td>
<td>-41.5</td>
<td>0</td>
<td>443.9</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>FFT21</td>
<td>331.3</td>
<td>∞</td>
<td>91.8</td>
<td>∞</td>
<td>-0.3</td>
<td>0</td>
<td>66.9</td>
<td>∞</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* some arbitrary % deviation from ω₀ depending on contamination level

frequency (ω) amplitudes (absolute value) from its pristine counterpart at a given value frequency value. The primed symbol (ω’) column for a given value of ω stands for its theoretical counterpart since, any value of ω except
under pristine conditions should be theoretically zero and as a results any deviation from that would tend towards $\infty$. The fact that the measured values of non-$\omega_0$ frequencies are non-zero is attributed to finite sampling-size (e.g. $N = 200$) but this does not affect the arguments we are building to establish the effects of various filters on the frequency bands.

There are few surprises in col-7 of table II a) None of the traditional filters actually improved the recovery of the pristine signal amplitudes! Although median-filter is the only one that doesn’t make it worse, the other filters actually reduces the amplitude (e.g. degraded outcome). This is contrary to the popular belief that smooth looking filtered data points are actually better in information content, b) filters like $G$ although makes the output signal look much smoother, e.g. Fig. 3 (b), but in reality it actually degrades the signal amplitude (while reducing the non-resonant frequency components as well) as apparent from row-5 of table II.

IV. Discussion, limitations and future directions

Signal recovery from noisy data is a very important but challenging problem that manifests ubiquitously in almost all natural systems. FFT, although one of the most powerful mathematical transform, relevant in nearly all signal processing contexts could be improved dramatically by carefully curating the data such that, it is transformed into more appropriate and aligned data-sets to fit into the core underlying requirements of FFT. Merely feeding data into FFT processors doesn’t guarantee accuracy or reliability of outcomes for real-world signals that is almost always contaminated one way or the other. It is very important to note that, our work no way questions the validity of the FFT method itself, but rather illustrates the pitfalls of using un-curated raw-data (or even traditionally filtered data), to feed into mathematical transforms while presuming that the underlying hypothesis behind those transforms are naturally satisfied by the raw-data. Given, the immense significance that FFT plays in the world of medical-diagnosis, it is important to be aware of the limitations described here to avoid any un-verified complacency that traditional filtering necessarily impore diagnostic outcomes that are grounded on FFT. General details behind the fundamental principles behind this method is described in ref. [52] and in future we would create a dedicated module to address FFT21 specific parameters and algorithmic twists to establish a dedicated software module to make more widely available through commercial softwares like Matlab [53], GNU Octave [55], Mathematica [57], etc.

Also, as a proof of principle, we described a simple case-study as an illustrative example. Although, the perturbations on the pristine signal from Eq. 1 is generated randomly in an unbiased fashion, real world signals are much more complex than the simple illustrative model described here but that discussion is beyond the scope of this work. Future work would include case-studies of more non-trivial, realistic signals as well as more complex simulations to capture realistic conditions.

One interesting observation is that, even the historically well established signal processing filters like Wiener, Gaussian, Median and TotalVariation are better suited only for signal-smoothing operations rather than signal-recovery processes as evident form col-7 of table II. Traditional filters fail to recover even a vanilla (or simple) category of underlying harmonic signal perturbed with a simplistic noise-model. One of our major insight to mitigate those limitations is the observation that, filtering process first exclude and curate the identified outlier data points before traditional processing can take place rather than blindly applying it to all measured raw data-points.

One of the known limitation of this works is that, we created a relatively simple signal and noise model to demonstrate the principles which should not be treated as representative of real-world complex signals like those found in realistic EKG, EEG etc. However, we’ll leave that to the future work to simple it simple and rather keep the focus on the extreme importance of raw-data pre-processing method before FFT or other transforms are applied.

V. Conclusion

In conclusion, we argue that, the very important Discrete Fourier-Transform can be dramatically improved by pre-processing and curating raw-data to make it eligible for traditional processing so that the underlying assumptions behind DFT are minimally violated. We have demonstrated that outcomes of traditional mathematical transforms can be significantly enhanced in the context of real-world data if the data itself is curated to be consistent with the underlying assumptions of those transforms (e.g. DFT). The curation of the real-world data indicates that curated raw-signal has a close to pristine frequency amplitude (and phase as well) than that from the un-curated raw-data. As the world becomes more

17 or any other transforms

18 with respect to some arbitrary measure
and more data-driven the accuracy of the data (uncurated or curated) fed to the transforms, determines the accuracy of the predictions made from those datasets. A philosophical shift is the need of the moment to develop scalable data-curation strategies as opposed to mere data-collection strategies which currently dominate the field in the context of being data-driven. It is to be noted that, FFT21 is not a new FFT method but rather a pre-processing method to make the raw-data eligible for further processing with FFT. This pre-processing followed by the regular FFT is what makes a complete FFT21 operation.

This work is far from complete. In future, we would develop formal case studies to determine curation strategies based on actual nature of the data (e.g. time-series vs. matrix-like) as well as targeted methods specific to the particular types of transforms like Radon-transform [59], [60], Bessel-transform [61], [62], Hankel-transform [63], [64] etc.

The argument we developed here is not limited to transform space scenarios. At the heart of data curation is entropy and dimensional reduction. The case study with FFT is a merely a mechanism to highlight the fact that curation is an essential step before processing with raw source level data. This work calls for a radical change in our approach to readily generalize the utility of pure mathematical expressions in real-world scenarios.

References