Abstract—Finite rate of innovation (FRI) is a powerful re-
construction framework enabling the recovery of sparse Dirac
streams from uniform low-pass filtered samples. An extension of
this framework, called generalised FRI (genFRI), has been
recently proposed for handling cases with arbitrary linear mea-
surement models. In this context, signal reconstruction amounts
to solving a joint constrained optimisation problem, yielding esti-
mates of both the Fourier series coefficients of the Dirac stream
and its so-called annihilating filter, involved in the regularisation
term. This optimisation problem is however highly non convex
and non linear in the data. Moreover, the proposed numerical
solver is computationally intensive and without convergence
guarantee.

In this work, we propose an implicit formulation of the
genFRI problem. To this end, we leverage a novel regularisa-
tion term which does not depend explicitly on the unknown
annihilating filter yet enforces sufficient structure in the solution
for stable recovery. The resulting optimisation problem is still
non convex, but simpler since linear in the data and with less
unknowns. We solve it by means of a provably convergent
proximal gradient descent (PGD) method. Since the proximal step
does not admit a simple closed-form expression, we propose an
inexact PGD method, coined as Cadzow plug-and-play gradient
descent (CPGD). The latter approximates the proximal steps by
means of Cadzow denoising, a well-known denoising algorithm
for CPGD. Through extensive numerical simulations, we demon-
strate the superiority of CPGD against the state-of-the-art in the
case of non uniform time samples.

Index Terms—finite rate of innovation, non bandlimited sam-
ing, Dirac streams, non convex optimisation, Cadzow denoising,
proximal gradient descent, alternating projections.

I. INTRODUCTION

SAMPLING theorems lie at the foundation of modern
digital signal processing as they permit the convenient
navigation between the analogue and digital worlds [1], [2].
The most famous is undoubtedly the Shannon sampling theorem
[3], which states that bandlimited signals can be recovered
exactly from their discrete samples for a sufficient sampling rate.
This major result has had tremendous impact on the field
of signal processing and by extension on many fields of natural
sciences. But this unanimous celebration lead many scientists
to start thinking about sampling theory exclusively in terms of
bandlimitedness, which is only a sufficient condition for a
signal to admit a discrete representation. In fact, sampling
theorems can also be devised for non-bandlimited signals as long
as they possess finitely many degrees of freedom.

This fact was brought to the attention of the signal processing
community in [4], where the authors introduced the finite rate
of innovation (FRI) framework. FRI is concerned with
the sampling of sparse non bandlimited signals such as the
prototypical sparse signal, namely the $T$-periodic stream of
Diracs:

$$x(t) = \sum_{k' \in \mathbb{Z}} \sum_{k=1}^{K} x_k \delta(t - t_k - T k'), \quad \forall t \in \mathbb{R},$$

with $x_k \in \mathbb{C}$ and $t_k \in [0, T]$. In the FRI framework, the sparsity
is measured in terms of its rate of innovation, defined as the
number of degrees of freedom per unit of time. For instance,
the Dirac stream (1) has $2K$ degrees of freedom $\{x_k, t_k\}_{k=1,\ldots,K}$
per period $T$, yielding a finite rate of innovation of $\rho = 2K/T$.
Intuitively, any lossless sampling scheme for (1) must therefore
have a sampling rate at least as large as the rate of innovation
$\rho$ or it will be impossible to fix all degrees of freedom.

Blu et al. described a sampling scheme achieving the second best sampling rate after the critical innovation rate, permitting to
perfectly recover the signal innovations from the knowledge of any $2K + 1$ consecutive Fourier coefficients of $x$ [5]. Unfortunately, this scheme can be very sensitive to
noise perturbations in the collected samples. This is because
the recovery of the innovations $t_k$ relies on the resolution of
a so-called annihilating equation which requires the Toeplitz
matrix built from the Fourier coefficients to be rank deficient.
While this structural constraint is guaranteed to hold in the
case of noiseless recovery of Dirac streams, it can break down
in the presence of noise, inevitable in practical applications.

As a remedy, Blu et al. proposed to denoise the collected
samples prior to solving the annihilating equation. To this end,
they leveraged the well-known Cadzow algorithm [6], which
aims to retrieve the closest rank-deficient Toeplitz matrix to
a high-dimensional embedding of the data via an alternating
projection method. When upgraded with this extra denoising
step, simulations results from Blu et al. revealed that the overall
accuracy of the recovery procedure remains very good for
signal-to-noise ratios (SNR) as low as $5$ dB [5]. While Cadzow
algorithm empirically provides accurate results after a few itera-
tions, convergence in theory has however not been demonstrated
to date, due to the non convex nature of the space of rank-
deficient matrices. Condat and Hirabayashi [7] revisited Cadzow

M. Simeoni and A. Besson are with the Laboratoire de Communications
Audiovisuelles (LCAV), Ecole Polytechnique Federale de Lausanne (EPFL),
CH-1015, Lausanne, Switzerland.

For part of this work, M. Simeoni was also with the Foundations of Cognitive
Solutions group in the IBM Research Laboratory of Zurich.

P. Hurley is with the Centre for Research in Data Science & Mathematics
at Western Sydney University (WSU), Australia.

preprint, under review 1

arXiv:2006.06374v1 [math.OC] 11 Jun 2020
denoising as a structured low-rank approximation (SLRA) problem and proposed a Douglas-Rachford splitting algorithm to solve it [8], with higher accuracy than traditional Cadzow denoising. Unfortunately, the gain in accuracy comes at the price of significantly higher computational cost, the Douglas-Rachford splitting method requiring many more iterations to converge than Cadzow algorithm.

In addition to their somewhat heuristic nature, neither Cadzow denoising nor its upgrade can handle more general types of input measurements as considered in the generalised FRI (genFRI) framework introduced by Pan et al. in [9]. The latter extends FRI to very generic cases where the measurements are related to the unknown Fourier coefficients of signals satisfying the annihilating property by a linear map. In such configurations, both the Fourier coefficients and their corresponding annihilating filter are unknown and must be estimated from the data. Pan et al. proposed to perform this joint estimation task by solving a constrained optimisation problem which recovers the Fourier coefficients, required to minimise a quadratic data-fidelity term, and their corresponding annihilating filter coefficients. The annihilating equation linking the two unknowns is explicitly enforced as a constraint. This optimisation problem is highly non convex and non linear in the data. They suggested to solve it via an iterative alternating minimisation algorithm with multiple random initialisations [9]. The proposed algorithm however comes without convergence guarantees and is computationally intensive.

In this paper, we propose an implicit formulation of the genFRI problem in which only the Fourier coefficients to be annihilated are recovered. This formulation does not rely explicitly on the unknown annihilating filter but rather leverages a structured low-rank regularisation constraint based on a "Toeplitzification" linear operator, guaranteeing non-trivial solutions to the annihilating equation. The resulting optimisation problem is still non convex, but simpler to analyse and solve since it is linear in the data and with less unknowns. We solve the implicit genFRI problem via proximal gradient descent (PGD) [10], [11].

We first consider PGD with exact proximal steps which is shown to converge towards critical points of the implicit genFRI problem. The latter is however impractical since the proximal step involved at each iteration does not have a closed-form expression. We therefore consider an inexact PGD [12], with proximal steps approximated by means of alternating projections. In the case of injective forward matrices, the approximate proximal step is shown to reduce to Cadzow denoising. Such an approach is reminiscent of the plug-and-play (PnP) framework in which proximal operators involved in first-order iterative methods are replaced by generic denoisers [13], [14], [15]. For this reason, we name our reconstruction algorithm Cadzow PnP Gradient Descent (CPGD). We demonstrate that CPGD converges locally towards fixed points of the update equation for injective forward matrices. Through simulations of irregular and noisy time sampling of periodic stream of Diracs we show that CPGD is almost always more accurate and more efficient than the procedure proposed by Pan et al. in [9], sometimes by several orders of magnitude.

The remainder of the paper is organised as follows:

- Preliminary concepts required for the understanding of the further sections are introduced in Section II.
- Section III describes the genFRI problem and details the proposed implicit formulation. The CPGD algorithm is introduced in Section IV.
- Experiments and results are detailed in Section V and concluding remarks are given in Section VI.

Finally, note that all experiments and simulations are fully reproducible using the benchmarking routines provided in our GitHub repository [16].

## II. PRELIMINARIES

In this section we introduce a linear operator, baptised Toeplitzification operator,\(^2\) which transforms a vector into a Toeplitz matrix. This operator will be used in the regularisation term of our implicit genFRI optimisation problem. We then briefly review the method of alternating projections [17] as well as the FRI [4] framework and Cadzow denoising [7].

### A. Toeplitzification Operator

Assume that we are given an arbitrary vector \( x \in \mathbb{C}^N \), \( N = 2M + 1 \), with entries indexed as follows:

\[
x := [x_{-M}, x_{-M+1}, \ldots, x_M, 0] \mathbb{T}.
\]

Then, for any \( P \leq M \), we can embed \( x \) into the space \( T_P \) of Toeplitz matrices of \( \mathbb{C}^{(N-P) \times (P+1)} \) by means of the following Toeplitzification operator:

\[
T_P : \mathbb{C}^N \rightarrow \mathbb{C}^{(N-P) \times (P+1)}
\]

\[
x \mapsto [T_P(x)]_{i,j} := x_{-M+P+i-j},
\]

where \( i = 1, \ldots, N-P \), \( j = 1, \ldots, P + 1 \). Note from (2) that the value of an entry \( [T_P(x)]_{i,j} \) of the matrix \( T_P(x) \) depends only on the distance \( i - j \) between the row and column indexes: \( T_P(x) \) is therefore a Toeplitz matrix and the vector \( x \) is called its generator.

The Toeplitzification operator (2) can be used to implement linear convolutions. Indeed, it can be shown (see Appendix A available as supplementary material) that the multiplication of \( T_P(x) \) with a vector \( u = [u_1, \ldots, u_{P+1}] \mathbb{T} \) returns the valid part\(^3\) of the convolution between the two zero-padded sequences \( \tilde{x} := [\ldots, 0, x_{-M}, \ldots] \in \mathbb{C}^{P+1} \) returns the valid part of the convolution between the two zero-padded sequences \( \tilde{x} := [\ldots, 0, x_{-M}, \ldots] \in \mathbb{C}^{P+1} \) returns the valid part of the convolution between the two zero-padded sequences \( \tilde{x} := [\ldots, 0, x_{-M}, \ldots] \in \mathbb{C}^{P+1} \) returns the valid part of the convolution between the two zero-padded sequences \( \tilde{x} := [\ldots, 0, x_{-M}, \ldots] \in \mathbb{C}^{P+1} \) returns the valid part of the convolution between the two zero-padded sequences \( \tilde{x} := [\ldots, 0, x_{-M}, \ldots] \in \mathbb{C}^{P+1} \)

---

\(^1\)An efficient Python implementation of CPGD is provided on our GitHub repository [16].

\(^2\)The alternative appellation Toeplitzification was used in [7].

\(^3\)See Appendix A for a formal definition of the valid part of a convolution between zero-padded sequences.
B. Inverse Toeplitzification Operator

The inverse Toeplitzification operator is the pseudoinverse of the Toeplitzification operator, mapping a Toeplitz matrix $H \in \mathbb{C}^{(N-P)\times(P+1)}$ to its generator $h \in \mathbb{C}^N$. As we shall prove in Proposition 2, inverse Toeplitzification is achieved by averaging across each diagonal of $T_P(x)$. It is interesting to note that this operation is also leveraged in Cadzow denoising as described in [5], in order to map back the data from its high dimensional Toeplitz embedding. The formal interpretation of this inverse map as the pseudoinverse of the Toeplitzification operator proposed hereafter is nevertheless not discussed in [5], nor anywhere else we may be aware of.

To compute the pseudoinverse of $T_P$, we first need an expression for its adjoint map, detailed in the proposition hereafter.

**Proposition 1** (Adjoint operator of $T_P$). The adjoint operator $T_P^*$ of $T_P$ defined in (2) is given by

$$T_P^* : \mathbb{C}^{(N-P)\times(P+1)} \to \mathbb{C}^N, \quad H \mapsto \sum_{i=k-j-1-P} H_{ik}, \quad j = 1, \ldots, N.$$  \hspace{1cm} (3)

**Proof.** The proof is described in Appendix B available as supplementary material of this manuscript. \hfill \Box

Note that the adjoint map $T_P^*$ proceeds by summing across each diagonal of the input matrix $H$. We are now ready to derive an expression for the (left) pseudoinverse of $T_P$, described in the proposition hereafter.

**Proposition 2** (Pseudoinverse of $T_P$). The pseudoinverse $T_P^{-1} : \mathbb{C}^{(N-P)\times(P+1)} \to \mathbb{C}^N$ of $T_P$ defined in (2) is given by

$$T_P^{-1} = \Gamma^{-1}T_P^*,$$ \hspace{1cm} (4)

where $\Gamma \in \mathbb{C}^{N\times N}$ is a diagonal matrix with diagonal entries given by

$$\Gamma_{ii} = \min(i, P + 1, N + 1 - i), \quad i = 1, \ldots, N.$$ \hspace{1cm} (5)

**Proof.** The proof of this proposition is given in Appendix C available as supplementary material of this manuscript. \hfill \Box

Observe that the composition of $T_P^*$ and $\Gamma^{-1}$ in the expression of the pseudoinverse (4) indeed corresponds to a diagonal averaging since $T_P^*$ first sums across each diagonal of the matrix $H \in \mathbb{C}^{(N-P)\times(P+1)}$ and $\Gamma^{-1}$ then divides the sums by the number of elements on each diagonal.

C. The Method of Alternating Projections

In this section we briefly discuss the method of alternating projections (MAP) [17], central to Cadzow denoising. It is used in computational mathematics to approximate projections onto intersecting sets. In its simplest form proposed by von Neumann in 1933 [18], the MAP performs a cascade of $n$ projection steps onto subsets $\{M_1, \ldots, M_K\}$ of some Hilbert space $\mathcal{H}$, starting from a point $z \in \mathcal{H}$:

$$z = [\Pi_{M_K} \cdots \Pi_{M_1}]^n(z).$$ \hspace{1cm} (6)

In (6), $\Pi_{M_k}$ denotes the orthogonal projection map onto $M_k$, defined for $k = 1, \ldots, K$ as

$$\Pi_{M_k} : \mathcal{H} \to M_k,$$

for some norm $\| \cdot \|$ on $\mathcal{H}$. In the case of closed linear subspaces $\{M_1, \ldots, M_K\}$, von Neumann and Halperin showed that [18], [19], [20]

$$\lim_{n \to \infty} \left\| \Pi_{M_K} \cdots \Pi_{M_1}(z) - \Pi_{\bigcap_{k=1}^K M_k}(z) \right\| = 0, \quad \forall z \in \mathcal{H}.$$ \hspace{1cm} (7)

The MAP equation (6) can therefore be used to approximate the complex projection map $\Pi_{\bigcap_{k=1}^K M_k}$. For closed convex sets $\{M_1, \ldots, M_K\}$, Bregman [17] showed moreover the weak convergence of the MAP towards a point in the intersection $\bigcap_{k=1}^K M_k$. Strong convergence towards the actual projection was achieved by Dysktra’s MAP [21], one of the most popular variant to von Neumann’s original algorithm. In the case of non convex intersecting sets, the convergence of the MAP has only been established locally [22], [23], [24], [25]. For example, Andersson et al. considered in [24] the case of two (potentially non convex) finite-dimensional manifolds $M_1, M_2 \subset \mathcal{H}$ and showed the following local convergence result:

**Theorem 1.** [24, Theorem 1.6] Let $x \in M_1 \cap M_2$ be non-tangential, i.e. the angle between $M_1$ and $M_2$ at $x$ is positive. Then, for $z \in \mathcal{H}$ and $\epsilon > 0$, there exists $\delta \geq 0$ such that, if $\|x - z\| < \delta$,

$$1) \quad \left\| \Pi_{M_1} \Pi_{M_2}(z) - z \right\| \to 0, \quad z \in M_1 \cap M_2,$$

$$2) \quad \left\| z - \Pi_{M_1 \cap M_2}(z) \right\| < \epsilon \|x - \Pi_{M_1 \cap M_2}(z)\|.$$ \hspace{1cm} \hspace{1cm} (8)

Roughly speaking, Theorem 1 states that if the starting point $z$ is close enough to a non-tangential point of $M_1 \cap M_2$ (which as explained in [24] are all but very exceptional points of $M_1 \cap M_2$), then the MAP converges to a point in $M_1 \cap M_2$. Moreover, the error $\|z - \Pi_{M_1 \cap M_2}(z)\|$ can be made arbitrarily small with respect to $\|x - \Pi_{M_1 \cap M_2}(z)\|$. However, Theorem 1 is difficult to apply in practice since the value of $\delta$ guaranteeing a relative error below a given threshold $\epsilon$ is unknown. The MAP is hence often used as a heuristic in non convex settings with no convergence guarantees. This is notably the case of Cadzow denoising, discussed further in Section II-E.

D. FRI in a Nutshell

The classical FRI framework, introduced in [4], aims at estimating the innovations $\{(s_k, t_k), k = 1, \ldots, K\} \subset \mathbb{C} \times [0, T]$, of a $T$-periodic stream of Diracs:

$$x(t) = \sum_{k \in \mathbb{Z}} x_k \delta(t - t_k - Tk'), \quad \forall t \in \mathbb{R}.$$ \hspace{1cm} (9)

In standard FRI, the estimation procedure is divided into two stages. The locations $t_k$ are first estimated by a nonlinear method, and then arranged into a Vandermonde system whose solution yields the Dirac amplitudes [5]. The recovery of the locations $t_k$ relies on the so-called annihilating equation, dating

---

Footnote: \footnote{See [24, Definition 4.2] and [24, Definition 4.3] for a precise definition of the angle between two manifolds and the concept of non-tangentiality.}
from Prony’s work [26], which cancels out the Fourier series coefficients of \( x \) by convolving them with a particular filter, called the annihilating filter. The latter can be obtained numerically by means of Fourier coefficients of \( x \) transform vanishing at roots \( \{ u_k := e^{-j2\pi n / T}, k = 1, \ldots, K \} \):

\[
H(z) = \sum_{k=0}^{K} h_k z^{-k} = \prod_{k=1}^{K} (1 - u_k z^{-1}).
\]  

(8)

For such a filter, we have indeed

\[
(\hat{x} + h)_m = \sum_{k=0}^{K} h_k \hat{x}_{m-k} = \sum_{k=1}^{K} x_k \left( \sum_{k=0}^{K} h_k u_k^{m} \right) u_k^m = 0, \quad m \in \mathbb{Z},
\]

(9)

where \( \hat{x}_m = \sum_{k=0}^{K} x_k u_k^{m} \), \( m \in \mathbb{Z} \), are the Fourier coefficients of \( x \) in (1). Notice that the roots \( u_k \) of the \( z \)-transform \( H(z) \) in (8) of \( h \) are in one-to-one correspondence with the locations \( t_k \). Recovering them amounts to estimating the coefficients \( h = [h_0, \ldots, h_K] \in \mathbb{C}^{K+1} \) of \( h \) from the annihilating equation (9). If for instance we have \( N = 2M + 1 \) consecutive Fourier coefficients of \( x \), e.g. \( x = [\hat{x}_M, \ldots, \hat{x}_1] \in \mathbb{C}^{2M+1} \), we can extract the \( N - K \) equations from (9) corresponding to the convolution indices \( m = M + K, \ldots, M \), and use the Toeplitzification operator\(^5\) defined in (2) to form the following matrix equation:

\[
T_K(x)h = 0_{N-K}, \quad \| h \| \neq 0.
\]

(10)

Observe that any nontrivial element of the nullspace of \( T_K(x) \) is a solution to (10). For \( M \geq K \), it can be shown [5] that \( T_K(x) \in \mathbb{C}^{(N-K) \times (K+1)} \) has rank \( K \) and therefore has a nontrivial nullspace with dimension 1. Up to a multiplicative constant, the annihilating equation (10) admits hence a unique solution. The latter can be obtained numerically by means of total least-squares [5], which computes the eigenvector associated to the smallest\(^6\) eigenvalue of \( T_K(x) \). In the critical case \( M = K \), the matrix \( T_K(x) \) is square, while in the oversampling case \( M > K \) it is rectangular and tall. As explained in [5], oversampling makes the estimation procedure more resilient to potential noise perturbations in the Fourier coefficients. In such cases, Blu et al. recommend moreover to perform Cadzow denoising on the Fourier coefficients \( x \) (see Section II-E) as well as replace (10) by a more general annihilating equation:

\[
T_P(x)\tilde{h} = 0_{N-P}, \quad \| \tilde{h} \| \neq 0.
\]

(11)

with \( K \leq P \leq M \), and \( \tilde{h} \in \mathbb{C}^{P+1} \). Again, it is possible to show that \( T_P(x) \) has rank \( K \), and hence a nontrivial nullspace with dimension \( P + 1 - K \). Solutions to (11) are hence not unique in this case, but all are equally valid for practical purposes. Moreover, the increased nullspace dimension makes Cadzow denoising more efficient at filtering the noise component (see Section II-E hereafter). In practice, the case \( P = M \) has been reported to yield the best empirical performance [5].

\(^5\)Remember the link between the Toeplitzification operator and convolution discussed in Section II-A.

\(^6\)An eigenvalue exactly equal to zero may in practice be impossible to obtain due to numerical inaccuracies.

E. Cadzow Denoising

For strong noise perturbations, the generalised annihilating equation (11) may fail to admit a nontrivial solution. Indeed, noisy generators \( x \) can yield full column rank matrices \( T_P(x) \) with trivial nullspace. As a potential cure, Blu et al. propose to denoise the Fourier coefficients \( x \) prior to solving the annihilating equation. This denoising step attempts to transform \( T_P(x) \) into a Toeplitz matrix with rank at most \( K \), thus guaranteeing the existence of nontrivial solutions to (11). This operation is carried out by means of Cadzow denoising [7], an alternating projection method (see Section II-C) applied heuristically to the subspace \( \mathcal{T}_P \) of Toeplitz matrices and the subset \( \mathcal{H}_K \) of matrices with rank at most \( K \):

\[
\mathcal{H}_K := \left\{ M \in \mathbb{C}^{(N-P) \times (P+1)} \mid \text{rank } M \leq K \right\}.
\]

(12)

Using the notation introduced in Sections II-A, II-B and II-C, Cadzow denoising can be seen as processing the noisy coefficients \( x \) as follows:

\[
\tilde{x} = T_{P}^{*} \left[ \Pi_{\mathcal{T}_P} \Pi_{\mathcal{H}_K} \right]^{-1} T_{P}(x),
\]

(13)

for some suitable \( n \in \mathbb{N} \). The inverse Toeplitzification operator \( T_{P}^{*} \) applied after the \( n \) iterations of the alternating projection method is used to recover the denoised Fourier coefficients \( \tilde{x} \in \mathbb{C}^N \).

Equation (13) allows us to develop an intuitive understanding as to why Cadzow denoising tends to perform better in practice with values of \( P \) close to \( M \) in the generalised annihilating equation (11). Indeed, it is easy to see that the maximal rank of rectangular matrices in \( \mathbb{C}^{(N-P) \times (P+1)} \) ranges in \( [K + 1, M + 1] \) when \( P \) ranges in \( [K, M] \). Consequently, the subset \( \mathcal{H}_K \) of matrices of rank at most \( K \) becomes “smaller and smaller” relatively to the ambient space as \( P \) increases towards \( M \). This implies that the associated projection map \( \Pi_{\mathcal{H}_K} \) is more selective\(^8\) for large values of \( M \), hence making Cadzow’s algorithm better at filtering the noise component.

Note that since \( \mathcal{H}_K \) is a non convex set the convergence of the MAP in (13) is not guaranteed. Nevertheless, experimental results [5, 7] suggest that Cadzow denoising almost always converges after a few iterations (typically \( n \leq 20 \)), which could theoretically\(^9\) be explained by the local convergence result from Theorem 1. We conclude this section by providing closed-form expressions for the projection operators \( \Pi_{\mathcal{T}_P} \) and \( \Pi_{\mathcal{H}_K} \), needed in (13).

1) Projection onto \( \mathcal{T}_P \): As shown in Appendix D of the supplementary material, the orthogonal projection operator onto the subspace \( \mathcal{T}_P \subset \mathbb{C}^{(N-P) \times (P+1)} \) of rectangular Toeplitz matrices can be written in terms of the Toeplitzification operator and its pseudoinverse (see Section 2) as:

\[
\Pi_{\mathcal{T}_P} = T_{P} T_{P}^{*} = T_{P} \Gamma^{-1} T_{P}^{*}.
\]

\(^7\)For \( n, m \in \mathbb{Z}, n < m \), we denote by \( [n, m] \) the integer interval \( [n, m] \cap \mathbb{Z} \).

\(^8\)This can be seen from the closed-form expression of \( \Pi_{\mathcal{H}_K} \) provided in (14).

\(^9\)As explained in Section II-C, the assumptions of Theorem 1 are unfortunately very difficult to verify in practice.
2) Projection onto $\mathcal{H}_K$: The orthogonal projection operator onto the space $\mathcal{H}_K$ of matrices with rank at most $K$ is given by the Eckart-Young-Minsky theorem [27]. The latter states that the projection map

$$\Pi_{\mathcal{H}_K} (X) = \arg \min_{H \in \mathcal{H}_K} \|X - H\|_F, \quad X \in \mathbb{C}^{(N-P)\times(P+1)},$$

can be computed in closed-form as:

$$\Pi_{\mathcal{H}_K} (X) = U \Lambda_K V^H, \quad X \in \mathbb{C}^{(N-P)\times(P+1)},$$

where $X = U \Lambda V^H$ is the singular value decomposition of $X$, and $\Lambda_K$ is the diagonal matrix of sorted singular values truncated to the $K$ strongest ones. Note that the output of the projection map is unique as long as the $K$–th and $(K+1)$–th largest singular values are different. Fortunately, the space of matrices failing to verify this condition is very small –more precisely it is thin, as discussed extensively in [24, Section 2]. In practice moreover, floating-point arithmetic makes it very unlikely that the $K$–th and $(K+1)$–th largest singular values be exactly identical. Thus, the projection map $\Pi_{\mathcal{H}_K}$ can be considered single-valued for practical purposes.

### III. Generalised FRI as an Inverse Problem

#### A. Generalised FRI

In Section II-D, we have described a procedure for recovering the locations $t_k$ from consecutive Fourier coefficients of $x$. The issue of computing these Fourier coefficients from a collection of arbitrary linear measurements $y \in C^L$ of $x$, $L \geq N$ now remains. Blu et al. [5] treated the simple scenario of measurements resulting from regular time sampling with ideal low-pass prefiltering. In such a case, they showed that, for a well chosen prefilter bandwidth, the Fourier coefficients could simply be obtained by applying a discrete Fourier transform to the measurements $y$. For more general measurement types, the situation is more complex, and the Fourier coefficients $x \in \mathbb{C}^N$ must in general be estimated by solving a linear inverse problem:

$$y = Gx + n,$$

where $x \in \mathbb{C}^N$ is the estimate of the Fourier series (typically the noise level), $y$ is additive noise, usually assumed to be a white Gaussian random vector. In [9], Pan et al. proposed the generalised FRI (genFRI) optimisation problem to deal with (15). The latter is a non convex constrained optimisation problem whose objective is to jointly recover the Fourier coefficients $x \in \mathbb{C}^N$ –required to minimise a quadratic data-fidelity term and their corresponding annihilating filter coefficients $h \in \mathbb{C}^{P+1}$. The annihilating equation linking the two unknowns is explicitly enforced as a constraint, yielding an optimisation problem of the form:

$$\min_{h \in \mathbb{C}^{P+1}} \|Gx - y\|_2^2 \quad \text{subject to} \quad \begin{cases} T_P(x)h = 0_{N-P} \\
\langle h, h_0 \rangle = 1 \end{cases},$$

(16)

where $h_0 \in \mathbb{C}^{P+1}$ is generated randomly accorded to the circularly-symmetric complex Gaussian distribution $\mathcal{CN}(0, I_{P+1})$. The normalisation constraint\footnote{In [9], the authors have also considered the more natural normalisation constraint $\|h\| = 1$. They claim however that this normalisation strategy is less successful experimentally.} $(h, h_0) = 1$ is used to exclude trivial solutions to the annihilating equation in (16) [9], [28].

To solve (16), Pan et al. consider the corresponding Lagrangian for a fixed vector $h \in \mathbb{C}^{P+1}$ and show that, when $G$ is full column rank, critical solutions $x \in \mathbb{C}^N$ can be expressed in terms of the annihilating filter $h$ as [9, Appendix A]:

$$x = \xi - (G^H G)^{-1} R_P(h)^H \Omega_n^{-1} R_P(h)^H \xi,$$

(17)

where $\xi : = (G^H G)^{-1} G^H y$, $\Omega_n : = R_P(h)(G^H G)^{-1} R_P(h)^H$ and $R_P$ is the right-dual of the Toeplitzification operator $T_P$, defined as [9, Definition 1]:

$$R_P : \mathbb{C}^{P+1} \to \mathbb{C}^{(N-P)\times N}.$$

By applying (17) to the canonical basis of $\mathbb{C}^N$, it is easy to show that:

$$R_P(h) = \begin{bmatrix} h_0 & h_{P-1} & \cdots & h_0 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & 0 & h_p & h_{P-1} & \cdots & h_0 \end{bmatrix},$$

for $h = [h_0, \ldots, h_P] \in \mathbb{C}^{P+1}$.

Pan et al. then use (17) to saturate (16) with respect to $x$, obtaining an optimisation problem in terms of $h$ only:

$$\min_{h \in \mathbb{C}^{P+1}} \langle \Omega_n^{-1} T_P(\xi)h, T_P(\xi)h \rangle \quad \text{subject to} \quad \langle h, h_0 \rangle = 1.$$  

(19)

Finally, they propose to solve (19) via an heuristic alternating minimisation algorithm. At each iteration, the annihilating filter is updated as the solution to the constrained quadratic minimisation problem:

$$\arg \min_{h \in \mathbb{C}^{P+1}} \left\{ \Omega_n^{-1} T_P(\xi)h, T_P(\xi)h \right\} \quad \text{subject to} \quad \langle h, h_0 \rangle = 1, \quad n \geq 0.$$  

(20)

Observe that (20) corresponds to a relaxation of (19) where the unknown matrix $\Omega_n$ has been replaced by the fixed matrix $\Omega_n$, constructed from the previous estimate $h_0$. Pan et al. show moreover in [9, Appendix B] that the iterates can be computed efficiently by solving, at each iteration, a linear system of size $2(N+1) \times 2(N+1)$:

$$\begin{bmatrix} 0 & T_P(\xi)^H & 0 & 0 \\
0 & -R_P(h_0) & 0 & 0 \\
T_P(\xi)^H & 0 & 0 & \lambda \\
0 & 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 0 \\
\ell \\
v \\
0 \end{bmatrix},$$

(21)

where $\ell \in \mathbb{C}^{N-P}$, $v \in \mathbb{C}^N$ and $\lambda \in \mathbb{C}$ are auxiliary variables. In practice, the algorithm is stopped when the data mismatch $\|Gx_n - y\|_2$ falls below a certain threshold $\epsilon$ (typically the noise level), where $x_n \in \mathbb{C}^N$ is the estimate of the Fourier series coefficients at iteration $n$, obtained by plugging $h_n$ in (17). For greater numerical stability, Pan et al. recommend to compute $x_n$ by solving a linear system with size $(2N-P) \times (2N-P)$, equivalent to (17) [9, Appendix B]:

$$\begin{bmatrix} G^H G & R_P(h_n)^H \\
R_P(h_n) & 0 \end{bmatrix} \begin{bmatrix} x_n \\
\ell \end{bmatrix} = \begin{bmatrix} G^H y \\
0 \end{bmatrix}.$$  

(22)
Note that this heuristic iterative procedure comes without any strong or weak convergence guarantee: it is neither known if the sequence \(\{(h_n, x_n), n \in \mathbb{N}\} \subseteq \mathbb{C}^{P+1} \times \mathbb{C}^N\) converges to a critical point of (16), nor if \(\{(h_n, n \in \mathbb{N}\} \subseteq \mathbb{C}^{P+1}\) converges to a fixed-point of (20). Moreover, the proposed stopping criterion requires knowing the noise level, unknown in practice. When it is unknown, Pan et al. recommend performing the reconstruction for a fixed and arbitrary number of iterations (typically fifty). For optimal performance, they moreover suggest to run the algorithm for multiple random initialisations of \(h_0 \in \mathbb{C}^{P+1}\) (typically fifteen). The overall reconstruction procedure can therefore be computationally intensive, since each iteration has complexity \(O((N+1)^3 + (2N-P)^3) = O(N^3)\) –the cost of solving the linear systems (21) and (22).

**Remark** (Choice of \(P\) and numerical stability). While theoretically well-defined for \(K \leq P \leq M\), the iterative procedure described above was only tested for \(P = K\) in [9]. This conservative choice was most likely motivated by numerical stability considerations: the inversion step in \(\xi = (G^H G)^{-1} G^H y\) can indeed be numerically unstable when \(G^H G\) is ill-conditioned, which is less likely to be the case for tall matrices \(G\) obtained with \(P = K\). We will touch on this again later.

### B. Implicit Generalised FRI

The annihilating equation constraint in (16) can be thought of as regularising the genFRI problem. Indeed, minimising the quadratic term \(\|Gx - y\|_2^2\) alone in the presence of noise would not necessarily yield Fourier coefficients \(x\) with nontrivial annihilating filter, which the annihilating constraint enforces explicitly. Unfortunately, this regularisation also complicates significantly the optimisation procedure. Indeed, it requires the introduction of an extra unknown variable with non linear dependency on the data, namely the annihilating filter \(h\). Moreover, the non linear constraint \(T_P(x) h = 0_{N-P}\) is highly non convex, and state-of-the-art algorithms, such as alternating minimisation or gradient descent [10], may suffer from getting trapped in local minima\(^1\) [29]. To circumvent these issues, we propose the following implicit formulation of the genFRI problem, in which only the Fourier coefficients are recovered:

\[
\min_{x \in \mathbb{C}^N} \|Gx - y\|_2^2 \quad \text{subject to} \quad \begin{cases} \text{rank} T_P(x) \leq K, \\ \|x\|_\Gamma \leq \rho, \end{cases}
\]

where \(K \leq P \leq M, \rho \in [0, +\infty)\), and \(\|x\|_\Gamma\) is the norm induced by the diagonal and positive definite matrix \(\Gamma \in \mathbb{C}^{N \times N}\) in (44):

\[
\|x\|_\Gamma := \sqrt{x^H \Gamma x}, \quad \forall x \in \mathbb{C}^N.
\]

Similarly to (16), the quadratic term \(\|Gx - y\|_2^2\) in (23) is used to guarantee high fidelity of the recovered coefficients to the observed data. Unlike (16), (23) leverages a regularising rank constraint on \(T_P(x)\) which does not explicitly involve the unknown annihilating filter. As already discussed in Section II-E in the context of Cadzow denoising, requiring \(T_P(x)\) to be of rank at most \(K\) is indeed a sufficient condition for the generalised annihilating equation (11) to admit nontrivial solutions. This implicit regularisation greatly simplifies the genFRI problem, since it decouples the problem of estimating the Fourier coefficients from the problem of estimating the annihilating filter. The normalisation constraint \(\|x\|_\Gamma \leq \rho\) enforces finite weighted energy (24) to the recovered Fourier coefficients. As shall be seen in Section IV, it can be relaxed when the forward matrix \(G\) is injective by setting \(\rho = +\infty\). Indeed, it is only used to ensure coercivity in underdetermined cases where the forward matrix \(G\) has a nontrivial null space. Coercivity is indeed a key assumption [30] for the convergence of the proximal gradient descent method envisioned in Section IV-A. We conclude this section by noting that the choice of \(\Gamma\) as weighting matrix in the energy normalisation constraint is arbitrary and purely motivated by computational considerations. Indeed, any choice of positive definite weighting matrix in (24) would have been suitable for the sole purpose of making the objective functional coercive. As explained in Section IV-B however, defining the weighting matrix as \(\Gamma\) greatly simplifies the computations involved at each iteration of the numerical solver proposed in Section IV-A.

**Remark** (On the choice of \(P\)). With similar developments to Section II-E for Cadzow denoising, it is possible to show that the the set of matrices in \(\mathbb{C}^{N-(P-K)\times(N-P)}\) of rank at most \(K\) becomes “smaller and smaller” with respect to the ambient space as \(P\) grows towards \(M\). As a result, the rank constraint in (23) is more selective for values of \(P\) close to \(M\), hence enforcing a stronger regularisation. We can hence expect (23) to perform better in practice for \(P = M\). This is in contrast with the explicit generalised FRI problem (16), whose equality constraint is indeed stringently different for different values of \(P\).

**Remark** (Case \(G = I\)). When \(G = I\), the optimisation problem (23) becomes a simple denoising problem, which could therefore be used as an alternative to Cadzow denoising or its upgrade [7].

### IV. Optimisation Algorithm

#### A. Non Convex Proximal Gradient Descent

The optimisation problem (23) can be rewritten in unconstrained form as:

\[
\min_{x \in \mathbb{C}^N} \|Gx - y\|_2^2 + \iota_{\mathcal{H}_k}(T_P(x)) + \iota_{\mathcal{H}_p}(x),
\]

where \(\mathcal{H}_k\) is the non convex set of matrices with rank lower than \(K\) defined in (12), \(\mathbb{B}_\rho^\Gamma := \{x \in \mathbb{C}^N : \|x\|_\Gamma \leq \rho\}\) is the \(\Gamma\)-ball with radius \(\rho > 0\), and \(\iota_{\mathcal{H}_k} : \mathbb{C}^{(N-P)(P+1)} \to [0, +\infty), \iota_{\mathcal{H}_p} : \mathbb{C}^N \to (0, +\infty)\) are indicator functions with domains \(\mathcal{H}_k\) and \(\mathbb{B}_\rho^\Gamma\), respectively. Observe that the unconstrained optimisation problem (54) can be written as a sum between a convex and differentiable quadratic term

\[
F(x) := \|Gx - y\|_2^2, \quad x \in \mathbb{C}^N,
\]

and a non convex and non differentiable term

\[
H(x) := \iota_{\mathcal{H}_k}(T_P(x)) + \iota_{\mathcal{H}_p}(x), \quad x \in \mathbb{C}^N.
\]

\(^{1}\)This is notably the reason why Pan et al. recommend multiple random initialisations of their algorithm in [9].
It is moreover easy to see that the gradient of $F$

$$\nabla F(x) = 2G^H(Gx - y), \ x \in \mathbb{C}^N, \quad (26)$$

is $\beta$-Lipschitz continuous with respect to the $\Gamma$-norm (24), with Lipschitz constant given by

$$\beta = 2\|G^H G\|_F = \sup \left\{2\|G^H Gx\|_F : x \in \mathbb{C}^N, \|x\|_F = 1 \right\} = \sup \left\{\|1/2 G^H G 1/2 x\|_F : x \in \mathbb{C}^N, \|x\|_F = 1 \right\} = 2\|1/2 G^H G 1/2\|_F. \quad (27)$$

It is hence possible to optimise (54) by means of proximal gradient descent (PGD) [10], an iterative method alternating between gradient and proximal steps according to the following update equation:

$$x_{k+1} \in \text{prox}_H^{\Gamma}(x_k - \tau \nabla F(x_k)), \quad (28)$$

for $k \geq 0$, $x_0 \in \mathbb{C}^N$, $\tau > 0$ and $\text{prox}_H^{\Gamma}$ defined in (29). Given a current estimate $x_k \in \mathbb{C}^N$, the update equation (48) decreases the value of the objective function (54) by selecting a proximal point $x \in \mathbb{C}^N$ to its proximal points with respect to $H$ is called proximal operator, and is defined as [10]

$$\text{prox}_H^{\Gamma}(x) : \begin{cases} \mathbb{C}^N \rightarrow \mathcal{P}(\mathbb{C}^N), \\ x \mapsto \arg \min \frac{1}{2\tau} \|x - z\|^2 + H(z). \end{cases} \quad (29)$$

where $\mathcal{P}(\mathbb{C}^N)$ is the power set of $\mathbb{C}^N$, and $\tau > 0$ controls the relative importance of $H$ with respect to the squared distance to $x$ measured in terms of the $\Gamma$-norm (24). The function $H$ being non convex, the proximal operator (29) will in general return multiple proximal points, which can all be used interchangeably in (48). The convergence of the sequence $\{x_k\}_{k \in \mathbb{N}}$ of PGD iterates (48) towards critical points of (54) is established by the following theorem.

**Theorem 2** (Convergence of PGD for Arbitrary $G$). Assume that $\rho \in [0, +\infty]$ in (54), and $\tau < 1/\beta$ with $\beta$ defined in (48). Then, any limit point $x_*$ of the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by (48) is a local minimum of (54).

**Proof.** The proof of this theorem is adapted from [30, Theorem 1] and given in Appendix E available as supplementary material of this manuscript. \qed

As stated by Theorem 3 hereafter, the convergence of PGD furthermore extends to the case $\rho = +\infty$, at least for injective forward matrices $G$. Setting $\rho = +\infty$ in (23) is equivalent to dropping the energy normalisation constraint, since $\|x\|_F \leq +\infty$ is trivially verified and hence the associated indicator function $\iota_{\mathcal{P}}$ in (54) is always null.

**Theorem 3** (Convergence of PGD for Injective $G$). Assume that $\rho = +\infty$ in (54), $\tau < 1/\beta$ with $\beta$ defined in (48), and $G \in \mathbb{C}^{L \times N}$ in (54) is injective, i.e. $\ker(G) = \{0_N\}$. Then, any

limit point $x_*$ of the sequence $\{x_k\}_{k \in \mathbb{N}}$ generated by (48) is a local minimum of (54).

**Proof.** The proof of this theorem is given in Appendix E available as supplementary material of this manuscript. \qed

A practical implication of Theorem 3 is that, for injective forward matrices $G$, PGD applied to the following relaxed implicit genFRI problem is convergent:

$$\min_{x \in \mathbb{C}^N} \|Gx - y\|^2 + \tau \iota_{H_k}(T_P(x)), \quad (30)$$

where $F(x) := \|Gx - y\|^2$, and $H(x) := \iota_{H_k}(T_P(x))$.

As discussed in Section IV-B, (30) should always be favoured over (54) for injective forward matrices $G$, since solving it via PGD requires less computations at each proximal step.

### B. Cadzow PnP Gradient Descent

As seen in the previous section, PGD requires the computation of the proximal operator (29) at each iteration, which amounts to finding a minimiser to the following non convex optimisation problem:

$$\check{x} \in \arg \min_{x \in \mathbb{C}^N} \left\{\frac{1}{2\tau} \|x - z\|^2 + \iota_{H_k}(T_P(x)) + \iota_{T_{\rho}}(z)\right\}, \quad (31)$$

for some input $x \in \mathbb{C}^N$. Observe that the proximal step (31) can be seen as a generalised projection step, aiming to find a point $\check{x}$ as close as possible in terms of the $\Gamma$-norm\(^{12}\) from $x$ while verifying some convex and non convex constraints specified by the indicator functions. This is formalised by Proposition 3, which shows that solutions to (31) can be identified with those of an orthogonal projection problem:

**Proposition 3.** The proximal operator (29) of $H(x) := \iota_{H_k}(T_P(x)) + \iota_{T_{\rho}}(x)$, for $\rho \in [0, +\infty]$ and $K \leq \rho \leq M$ is given by

$$\text{prox}_H^{\Gamma}(x) = T_P \Pi_{T_{\rho} \cap H_k \cap \mathbb{B}_\rho} T_P(x), \quad \forall x \in \mathbb{C}^N, \quad (32)$$

where $\mathbb{B}_\rho := \{X \in \mathbb{C}^{(N-P) \times (P+1)} : \|X\|_F \leq \rho\}$ and $\Pi_{T_{\rho} \cap H_k \cap \mathbb{B}_\rho}$ is the orthogonal projection operator onto $T_{\rho} \cap H_k \cap \mathbb{B}_\rho$ with respect to the Frobenius norm:

$$\Pi_{T_{\rho} \cap H_k \cap \mathbb{B}_\rho}(X) : \begin{cases} \mathbb{C}^{(N-P) \times (P+1)} \rightarrow \mathcal{P}(\mathbb{C}^{(N-P) \times (P+1)}), \\ X \mapsto \arg \min_{H \in \mathbb{C}^{(N-P) \times (P+1)}} \|X - H\|_F. \end{cases} \quad (33)$$

**Proof.** The proof of this proposition is given in Appendix F available as supplementary material of this manuscript. \qed

Equation (65) provides us with a practical way of computing the proximal step (29) associated to a point $x \in \mathbb{C}^N$. Unfortunately, the orthogonal projection operator $\Pi_{T_{\rho} \cap H_k \cap \mathbb{B}_\rho}$ admits no simple closed-form expression. We therefore propose to approximate it by the method of alternating projections (MAP) (see Section II-C):

$$\Pi_{T_{\rho} \cap H_k \cap \mathbb{B}_\rho} \approx \left[\Pi_{T_{\rho}} \Pi_{H_k} \Pi_{\mathbb{B}_\rho}\right]^6, \quad (33)$$

\(^{12}\)Observe that the weighting matrix $\Gamma$ puts more emphasis on the coefficients that appear more often in the Toeplitz matrix $T_P(x)$.
**Algorithm 1 Cadzow PnP Gradient Descent (CPGD)**

**Require:** $y$, $G$, $T_p$, $x_0$, $K \leq P$, $\tau < 1/\beta$, $n \in \mathbb{N}$, $\rho > 0$  

$k:=0$

repeat

$z_{k+1} := x_k - 2\tau G^H (Gx_k - y)$

if $\rho = +\infty$ then

$x_{k+1} := T_p \left[ \Pi_{T_p} \Pi_{H_k} \right]^n T_p (z_{k+1})$

else

$x_{k+1} := T_p \left[ \Pi_{T_p} \Pi_{H_k} \Pi_{\mathbb{B}_n} \right]^n T_p (z_{k+1})$

end if

$k \leftarrow k + 1$

until a stopping criterion is satisfied

return $x^{(k)}$

for some $n \in \mathbb{N}$. Observe that when $\rho = +\infty$ (which is possible for injective matrices $G$, see Theorem 3) we have $\mathbb{B}_n = \text{Id}$ and since the right-hand side of (33) simplifies to $\left[ \Pi_{T_p} \Pi_{H_k} \right]^n$. Note that since $H_k$ is non convex, the convergence as $n$ grows to infinity of the product $\left[ \Pi_{T_p} \Pi_{H_k} \Pi_{\mathbb{B}_n} \right]^n$ towards the actual projection map $\Pi_{T_p \cap H_k \cap \mathbb{B}_n}$ is not guaranteed in general (see discussion in Section II-C). For the specific case $\rho = +\infty$ however, it is possible to apply Theorem 1 to show the local convergence of the MAP (33):

**Corollary 1.** Let $Z \in H_k \cap T_p$ be a non tangential point [24, Definition 4.3]. Then, for $X \in \mathbb{C}^{(N-P) \times (P+1)}$ and $\epsilon > 0$, there exists $\delta > 0$ such that, if $\|X - Z\|_F \leq \delta$,

1. $\left[ \Pi_{T_p} \Pi_{H_k} \right]^n (X) \rightarrow X \in H_k \cap T_p$,
2. $\|X - \Pi_{H_k \cap T_p}(X)\|_F \leq \epsilon$.

Proof. Similarly to the proof of [31, Theorem 7], Corollary 1 is obtained by applying Theorem 1 to the manifolds $M_1 = R_k$ of matrices with rank exactly $K$ which is dense in $H_k$ [24, Proposition 2.1]– and $M_2 = T_p$. For more details, see the proof of [31, Theorem 7], which discusses the local convergence of the MAP for $H_k \cap H_p$ where $H_p$ denotes the space of rectangular Hankel matrices. Since Hankel matrices are just reflected Toeplitz matrices, the analysis extends to the case of Toeplitz matrices. □

Roughly speaking, Corollary 1 states that, if applied to a matrix $X$ close enough to a non tangential point of $T_p \cap H_k$ (as which discussed in [24] for the case of Hankel matrices are all but very exceptional matrices of $T_p \cap H_k$), the MAP (33) converges to a point in $T_p \cap H_k$. Moreover, the error $\|X - \Pi_{H_k \cap T_p}(X)\|_F$ can be made arbitrarily small with respect to $\|X - \Pi_{H_k \cap T_p}(X)\|_F$. While difficult to verify in practice, the local convergence result Corollary 1 reassures us on the well-foundedness of approximation (33).

Plugging (33) into (65) finally yields the following approximate proximal step:

$$\text{prox}_{\Gamma_f^H}(x) = T_p \left[ \Pi_{T_p} \Pi_{H_k} \Pi_{\mathbb{B}_n} \right]^n T_p (x), \quad \forall x \in \mathbb{C}^N,$$  

(34)

for some $n \geq 0$. The PGD algorithm with proximal step (34) is provided in Algorithm 1. Observe that when $\rho = +\infty$, (34) reduces to Cadzow denoising (13). The effect of heuristic (33) is hence to replace the proximal step in the PGD iterations by a generic denoising step. Such an approach is reminiscent of the **plug-and-play (PnP)** framework [15], [14] from image processing, which leverages generic denoisers to approximate complex projection or proximal operators [13]. For this reason, we baptise our algorithm **Cadzow PnP Gradient Descent (CPGD)**. In the next section, we study the convergence of Algorithm 1.

C. **Local Fixed-Point Convergence of CPGD**

In Section IV-A, we established Theorems 2 and 3 which show the convergence of PGD towards critical points of (54). However, such results required the computation of exact proximal steps (29) in the PGD iterations, and do not apply to CPGD which leverages the inexact proximal step (34). Convergence of PGD in non convex setups with exact proximal steps was studied in [12], [32]. The results established in both papers require the proximal step approximation errors incurred at each iteration to be decreasing and summable, which may not necessarily be the case for the MAP approximation (33). It is nevertheless possible to demonstrate that the iterations of CPGD are locally **contractive**, and therefore locally convergent towards a fixed point using the Banach contraction principle. Such a result is stated in Theorem 4 hereafter.

**Theorem 4** (CPGD is a Local Contraction). Let $R_k \subset \mathbb{C}^{(N-P) \times (P+1)}$ be the set of matrices of rank exactly $K \leq P \leq \lceil N/2 \rceil$, and $U_{\tau,n} : \mathbb{C}^N \rightarrow \mathbb{C}^N$ the update CPGD map

$$U_{\tau,n}(x) := H_n (x - \tau VF(x)), \quad x \in \mathbb{C}^N,$$  

(35)

with $H_n(x) := T_p \left[ \Pi_{T_p} \Pi_{H_k} \Pi_{\mathbb{B}_n} \right]^n T_p (x)$. Let $G \in \mathbb{C}^{L \times N}$ be injective, and $\Gamma$ be the diagonal and positive definite matrix defined in (44). Define

$$\alpha := 2\lambda_{\min}(\Gamma^{1/2} G^H G \Gamma^{-1/2}),$$  

$$\beta := 2\lambda_{\max}(\Gamma^{1/2} G^H G \Gamma^{-1/2}),$$

where $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimum and maximum eigenvalues of a matrix $M$ respectively.

Then, $U_{\tau,n}$ is locally well-defined (single-valued) and Lipschitz continuous with respect to the $\Gamma$-norm

$$\|U_{\tau,n}(x) - U_{\tau,n}(z)\|_F \leq L_{\tau} \|x - z\|_F,$$  

for all $x, z \in \mathbb{C}^N$ such that $T_p(x)$ and $T_p(z)$ are in some neighbourhood of some matrix $R \in R_k$. The Lipschitz constant $L_{\tau}$ is given by

$$L_{\tau} = \max \{ |1 - \tau \alpha|, |1 - \tau \beta| \}.$$  

Moreover, $U_{\tau,n}$ is contractive, i.e. $0 \leq L_{\tau} < 1$, for $0 < \tau < 2/\beta$, and $L_{\tau}$ is minimised for $\tau = 2/(\alpha + \beta)$.

Proof. The proof of this theorem is given in Appendix G as supplementary material of this manuscript. □

The following corollary shows the local convergence of CPGD towards a fixed-point of the update map (35):

**Corollary 2** (CPGD Converges Locally). With the same notations as in Theorem 4, assume that all CPGD iterates \( \{x_k\}_{k \in \mathbb{N}} \) are such that

$$\{T_p(x_{k+1}), T_p(x_k)\} \subset U_k, \quad \forall k \in \mathbb{N},$$  

(36)
for some neighbourhood $\mathcal{U}_k$ of some matrix $R_k \in \mathcal{R}_K$. Assume further that $0 < \tau < 2/\beta$. Then, $x_k \xrightarrow{k \to \infty} x_*$, where $x_* \in \mathbb{C}^N$ is a fixed-point of $U_{\tau,n}$, i.e. $U_{\tau,n}(x_*) = x_*$. Moreover, we have $\|x_* - x_k\| \leq \frac{L}{1 - L^\tau} \|x_1 - x_0\|, \quad \forall k \geq 1. \quad (37)$

**Proof.** The proof of this theorem is given in Appendix H available as supplementary material of this manuscript. \[\square\]

**Remark (Speed of Convergence).** From Theorem 4 and (2), we see that the sequence $(x_k)_{k \in \mathbb{N}}$ converges the fastest when $L_\tau$ is minimised, i.e. $\tau = 2/(\alpha + \beta)$.

**Remark (Fixed Points vs. Critical Points).** Note that Corollary 2 is a much weaker result than Theorems 2 and 3. Indeed, Corollary 2 only shows the local convergence of CPGD towards fixed points of $U_{\tau,n}$, which may not necessarily be critical points of the optimisation problem (54). Theorems 2 and 3 on the other hand, show the global convergence of PGD with exact proximal step towards critical points of (54). This is however the price to pay for computing the proximal step (31) efficiently in practice.

**Remark (Geometric Interpretation of Condition (36)).** Roughly speaking, Corollary 2 guarantees the convergence of CPGD towards a fixed point of the update map (35), provided that the forward matrix $G$ is injective, and that any two consecutive lifted estimates $T_P(x_k), T_P(x_{k+1})$, are in a common neighbourhood $\mathcal{U}_k$ of some matrix $R_k \in \mathcal{R}_K$. Note that this is much less stringent than requiring the entire lifted path $\{T_P(x_k)\}_{k \in \mathbb{N}}$ to belong to some neighbourhood $\mathcal{U}$ of some fixed matrix $R \in \mathcal{R}_K$. Indeed, condition (36) allows the lifted estimates to travel from one neighbourhood of the manifold $\mathcal{R}_K$ to another, provided that every visited neighbourhood contains at least two consecutive lifted estimates (see Figure 1 for an illustration). This condition, although difficult to verify in practice, seems however likely to hold for $\rho = +\infty$, small enough step sizes, large enough $n$ and $x_0 = 0_N$. Indeed, in such a case, we have:

- $T_P(x_0) \in \mathcal{H}_K$ is in some neighbourhood of $\mathcal{R}_K$ since $\mathcal{R}_K$ is dense in $\mathcal{H}_K$.
- For $n$ large enough, $T_P(x_k)$ is very likely to be in some neighbourhood of $\mathcal{R}_K$, since the denoising step in the update map (35) makes $T_P(x_k)$ close to be in the intersection $\mathcal{H}_K \cap T_P$ (see Corollary 1).
- For a small enough step size $\tau$, $T_P(x_k)$ and $T_P(x_{k+1})$ are likely to belong to the same neighbourhood of $\mathcal{R}_K$.

**V. EXPERIMENTS AND RESULTS**

In this section we validate the CPGD method numerically, considering as a testbed the scenario of irregular time sampling from [9, Section IV.A]. We assess both the reconstruction accuracy and the computational complexity of the method, and compare it to the state-of-the-art.

**Remark (Reproducibility).** Special care has been taken into making the experiments and simulations of this section fully reproducible. To reproduce the results, the reader is referred to the routines provided in our GitHub repository [16].

**A. Reconstruction Accuracy**

We define a 1-periodic stream of $K = 9$ Diracs (see Fig. 2):

$$x(t) = \sum_{m \in \mathbb{Z}} \sum_{k=1}^K x_k \delta(t - t_k - m), \quad \forall t \in \mathbb{R}, \quad (38)$$

where the amplitudes $x_k \in \mathbb{R}_+$ and locations $t_k \in [0,1]$ are random, with log-normal and uniform\(^{13}\) distributions respectively. We then generate $L = 73$ noisy samples as

$$y_l = \sum_{m=-M}^M \hat{x}_m e^{j2\pi m\theta_l} + \epsilon_l, \quad l = 1, \ldots, L, \quad (39)$$

where $\hat{x}_m = \sum_{k=1}^K x_k \exp(-j2\pi mt_k)$, $m = -M, \ldots, M$, are the Fourier coefficients of the Dirac stream $x, \theta_l \in [0,1]$ are chosen uniformly\(^{14}\) at random, and $\epsilon_l \in \mathbb{R}$ are independent realisations of a Gaussian random variable\(^{15}\) $N(0, \sigma^2)$, for $l \in [1, L]$.

As explained in [9, Section IV.A], the measurements $y_l$ correspond to noisy samples of the low-pass filtered Dirac stream $x$ at irregular times $\theta_l$ (see Fig. 2), where the low-pass filter is chosen as an ideal low-pass filter with bandwidth $2M + 1 \leq L$. Using the formalism of Section III, we can rewrite (39) in vector notation as

$$y = Gx + \epsilon, \quad (40)$$

where $y = [y_1, \ldots, y_L] \in \mathbb{C}^L, \quad x = [\hat{x}_M, \ldots, \hat{x}_M] \in \mathbb{C}^{N = 2M+1}, \quad \epsilon = [\epsilon_1, \ldots, \epsilon_L] \in \mathbb{C}^L, \quad \text{and} \quad G \in \mathbb{C}^{L \times N}$.

$$G = \begin{bmatrix}
  e^{-j2\pi M\theta_0} & \cdots & 1 & \cdots & e^{j2\pi M\theta_0} \\
  e^{-j2\pi M\theta_1} & \cdots & 1 & \cdots & e^{j2\pi M\theta_1} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  e^{-j2\pi M\theta_{L-1}} & \cdots & 1 & \cdots & e^{j2\pi M\theta_{L-1}} \\
  e^{-j2\pi M\theta_L} & \cdots & 1 & \cdots & e^{j2\pi M\theta_L}
\end{bmatrix}$$

Note that from the periodicity of complex exponentials, it is possible to flip the columns of $G$ so as to rewrite it as a Vandermonde matrix [9]. This shows that $G$ is injective – since $2M + 1 \leq L$ and the irregular time samples are all distinct. From the samples $y$ and the data model (40), we consider recovering the Fourier coefficients $x \in \mathbb{C}^N$ by means of three algorithms:
The CPGD algorithm 1 with $P = M$, $\rho = +\infty$ (since $G$ is injective) and step size $\tau = 1.5/\beta$ (where $\beta$ is set as described in Theorem 4).

The state-of-the-art algorithm of Pan et al. [9], described in Section III-A and referred to hereafter as GenFRI. For smooth integration, the Python 3 implementation of GenFRI provided by Pan et al. on their official Github repository [33] was included in our own algorithmic interface. Since the noise level is assumed to be unknown, we set –as recommended in [9, Section III-C]– the number of inner iterations and random initialisations to their default values, 50 and 15 respectively.

The baseline method, referred to hereafter as LS-Cadzow, which consists in applying Cadzow denoising to the least-squares estimate of the Fourier coefficients

$$
\begin{align*}
\chi_{LS} &= \operatorname*{argmin}_{x \in \mathbb{C}^N} \| Gx - y \|_2^2, \\
\chi_{LS-Cadzow} &= T_P \left[ \Pi_{\mathcal{T},P} \Pi_{\mathcal{H}_K} \right]^N T_P (\chi_{LS}).
\end{align*}
$$

We solve the least-squares optimisation problem in (41) by means of the `lstsq` function in the Python 3 package `scipy` [34], with cut-off ratio $\text{cond} = 10^{-4}$.

For CPGD, we fix the maximum number of iterations to 500 and consider that convergence is reached if the iterate norm is changed by less than 0.01% between two iterations. For Cadzow denoising, we fix the number of iterations to 10 for both LS-Cadzow and CPGD. The reconstruction accuracy is assessed by matching the true Dirac locations $t_k$ to the recovered ones, denoted by $\omega_k$, for $k$ between 1 and $K$. To do so, we proceed as explained in Section II-D and infer the Dirac locations $\omega_k$ from the z-transform roots of the annihilating filter associated to the Fourier coefficients estimated by each method. Then, we solve the following matching problem by means of the Hungarian algorithm [35]

$$
\min_{j_1, \ldots, j_K \in \{1, \ldots, K\}} \left\{ \frac{1}{K} \sum_{k=1}^{K} d(t_k, \omega_{j_k}) \right\},
$$

where $d(t, \omega) = \min\{|t - \omega|, 1 - |t - \omega|\}$, $\forall t, \omega \in [0, 1]$, is the canonical distance on the periodised interval $[0, 1]$. Finally, we report the average positioning error, corresponding to the value of the cost function $\sum_{k=1}^{K} d(t_k, \omega_{i_k})/K$ for the indices $\{i_1, \ldots, i_K\}$ solutions to the matching problem (42). This metric is computed for 192 noise realisations, various cutoff frequencies $M = \beta K$ with the oversampling factor $\beta \in \{1, 2, 3, 4\}$ (see Figs. 2a, 2b, 2c and 2d respectively) and various noise levels

$$
\sigma = \max_{k=1, \ldots, K} |x_k| \times \exp \left( \frac{-\text{PSNR}}{10} \right),
$$

where the peak signal to noise ratio PSNR ranges from $-30$ dB to 30 dB. The conditioning numbers of the matrix $G^{\beta}G$ for the

\[\text{See } [7, \text{Fig. 2}]] for additional details on the procedure used to recover the Dirac locations from the annihilating filter coefficients.\]

\[\text{The Hungarian algorithm is implemented in the } \texttt{linear_sum_assignment} \text{ function from the Python 3 package } \texttt{scipy} \text{ [34].}\]
Table I: Conditioning number of $G^H G$ for various values of the oversampling parameter $\beta$.

| $\beta$ | 1   | 2   | 3   | 4   |
|---------|-----|-----|-----|-----|
| $\kappa(G^H G)$ | 5.89 | 1.7 x 10^2 | 1.9 x 10^5 | 1.55 x 10^16 |

We can draw from the results the following:

- Figs. 3a, 5a, 5b and 5c reveal that without oversampling in the Fourier domain (i.e. $\beta = 1$ and a minimal cutoff frequency of $M = K = 9$), the three algorithms CPGD, GenFRI and LS-Cadzow perform similarly throughout the entire PSNR range. The average positioning error – almost indistinguishable for the three algorithms – goes from approximately 10% of the period for PSNRs of $\leq 30$ dB to 1% of the period for PSNRs of 30 dB.

- Figs. 3b, 5d, 5e and 5f reveal that with an oversampling

  of $\beta = 2$ –yielding a cutoff frequency of $M = 18$– the three algorithms CPGD, GenFRI and LS-Cadzow start behaving differently for PSNRs larger than 0 dB: CPGD has the lowest positioning error, followed by LS-Cadzow and finally GenFRI. The inter-quartile regions of the positioning errors distribution are however overlapping, which means that the differences in performance are not statistically significant. For PSNRs larger than 0 dB, all algorithms have a lower positioning error than in the case $\beta = 1$. For high PSNRs, this improvement can be as high as one and a half order of magnitude.

- Figs. 3c, 5g, 5h and 5i reveal that with an oversampling

  of $\beta = 3$ –yielding a cutoff frequency of $M = 27$, the three algorithms CPGD, GenFRI and LS-Cadzow again behave differently for PSNRs larger than 0 dB: CPGD has the lowest positioning error, followed by GenFRI and finally LS-Cadzow. For high PSNRs, the differences in performance among the three algorithms become statistically significant: the inter-quartile regions of the positioning error’s empirical distribution do not overlap anymore. CPGD moreover reaches a positioning error as low as 0.01% of the period, which is up to two orders of magnitude smaller than the minimal positioning

  error.
error of GenFRI or LS-Cadzow in this scenario. For PSNRs greater than 10 dB, CPGD improves its positioning error with respect to the case $\beta = 2$ by a bit less than half an order of magnitude. This is not the case for GenFRI and LS-Cadzow which both underperform with respect to the case $\beta = 2$ and even with respect to the case $\beta = 1$ for LS-Cadzow. This can be explained by the large conditioning number of the matrix $G^H G$ in this case (see Table I), affecting the numerical stability of both algorithms (see remark in Section III-A).

- Figs. 3a, 5j, 5k and 5l reveal that an oversampling of $\beta = 4$ —yielding a critical bandwidth of $2M + 1 = 73$ equal to the number of measurements $L$, CPGD is superior to GenFRI which is itself superior to the baseline method LS-Cadzow in nearly all cases, with the exception of very low PSNRs ($\sim -30$ dB), where the three methods have comparable reconstruction accuracy. For PSNRs larger than $-20$ dB, the differences in performance are statistically significant. CPGD is more accurate than GenFRI and LS-Cadzow by a few orders of magnitude (from 1 to 3 orders of magnitude for PSNRs larger than $-10$ dB), reaching a minimal positioning error as low as 0.005% of the period. For PSNRs greater than $-20$ dB, CPGD improves its positioning error with respect to all previous cases $\beta \in \{1, 2, 3\}$. Again, this is not the case for GenFRI and LS-Cadzow which both perform as good as or worse than the case $\beta = 1$. This can be explained by the (very) large conditioning number of the matrix $G^H G$ in this case (see Table I), which severely affects the numerical stability of both algorithms.

In conclusion, in these simulations CPGD is better at leveraging oversampling in the Fourier domain to improve the reconstruction accuracy by several orders of magnitude with respect to the non oversampled case. In particular, CPGD performs best when the bandwidth of the low-pass filter is chosen as large as the number of measurements. As explained in Section II-E, Cadzow denoising exhibits a similar behaviour. This similarity is not fortuitous: both algorithms leverage a similar rank constraint which becomes more and more selective as the oversampling parameter increases. In contrast, GenFRI and LS-Cadzow are negatively affected by large oversampling parameters, due to numerical stability issues. For the critical bandwidth $2M + 1 = L$, CPGD notably outperforms GenFRI and LS-Cadzow by one to three orders of magnitude, and this even for PSNRs as low as $-20$ dB.

### B. Computational Complexity

As explained in Section III-A, solving the linear systems (21) and (22) are the most expensive operations performed at each iteration of GenFRI, yielding an overall computational complexity of $O(N^3)$. For CPGD, the computational cost of each iteration is dominated by the successive projections onto $\mathcal{H}_K$ in the approximate proximal step, which are computed via a SVD —see Algorithm 1 and (14). At a cursory glance, it may seem that the overall complexity of CPGD is somewhat comparable to the one of GenFRI, since computing the SVD of a matrix with size $(N-P) \times (P+1)$ has in general a computational complexity of $O((N-P)^2(P+1) + (P+1)^3)$ [36] which reduces to $O(N^3)$ when $P = M$. In practice however, projecting onto $\mathcal{H}_K$ does not require to perform a complete SVD since only the $K$ strongest eigenvalues and their associated eigenvectors are needed. This truncated SVD can be performed very efficiently by means of the implicitly restarted Arnoldi method (IRAM) [37], or the implicitly restarted Lanczos method for Hermitian matrices. When $K \ll P + 1$, such methods are obviously much more efficient than a wasteful standard SVD. IRAM is moreover a matrix-free method [38]: it does not need the processed matrix to be stored in memory but simply requires an algorithm for performing matrix/vector products. In our context, since the truncated SVDs are exclusively performed on Toeplitz matrices, such matrix/vector products can be efficiently implemented by means of FFTs thanks to the convenient links between Toeplitz matrices and convolutions outlined in Section II-A.

The CPGD implementation provided in our Github repository [16] leverages all these computational tricks. In Fig. 4, we show that our implementation of CPGD is considerably faster than GenFRI. Fig. 4 reports the computation times of CPGD and GenFRI for $K = 9$ and bandwidth $N = 2\beta M + 1 = L$ with $\beta$ ranging from 1 to 300. To save computational time, the reconstruction times were scaled from the execution time of a single iteration of CPGD and GenFRI, assuming a typical number of iterations of 100 and $15 \times 50 = 750$ respectively. We observe that CPGD is always faster than GenFRI, sometimes by two orders of magnitudes. Moreover, regressions performed in log-log scale reveal that CPGD scales as $N^{2.11}$ while GenFRI scales as $N^{2.79}$. Note that this difference in scaling behaviour is overlooked by the complexity analysis above.

### VI. CONCLUSION

We propose an implicit version of the generalised finite-rate-of-innovation (genFRI) problem for the recovery of the Fourier series coefficients of sparse Dirac streams with arbitrary linear sensing. This formulation relies on a novel regularisation term which enforces the annihilation of the recovered Fourier series

![Figure 4: Reconstruction times for CPGD and GenFRI for various bandwidth sizes $N$ and $K = 9$. The reported times are for a MacBook Pro (16-inch, 2019), Intel Core i7 (6C/12T) @ 2.6GHz with 32GB RAM. These results can be reproduced using the Python script `reproduce_execution_times.py` located in the directory `/benchmarking/` of our GitHub repository [16].](image)
coefficients without explicitly involving the unknown annihilating filter. The resulting non convex optimisation problem is consequently simpler and linear in the data. To solve it, we suggest a proximal gradient descent (PGD) algorithm which we prove converges towards a critical point of the objective function. We further introduce an inexact PGD method, coined Cadzow plug-and-play gradient descent (CPGD), where the intractable proximal steps involved in PGD are approximated by means of alternating projections, akin to the popular Cadzow denoising algorithm. We outline the resemblance of CPGD to PnP methods used in image processing and prove its local fixed-point convergence under relatively weak assumptions. Considering the traditional irregular time sampling testbed, we demonstrate empirically that CPGD outperforms by several orders of magnitude the state-of-the-art GenFRI algorithm, both in terms of accuracy and reconstruction time.

For future work, we plan on investigating acceleration techniques for CPGD, such as approximate sketching-based eigenvalue decomposition methods [39], [40], more computationally efficient for large-scale problems. Applications of CPGD to acoustics and radio astronomy will also be investigated.

REFERENCES
[1] M. Vetterli, J. Kovačević, and V. K. Goyal, Foundations of signal processing. Cambridge University Press, 2014.
[2] B. Rimoldi, Principles of Digital Communication: A Top-Down Approach. Cambridge University Press, 2016.
[3] C. E. Shannon, “A mathematical theory of communication,” ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, pp. 3–55, 2001.
[4] M. Vetterli, P. Marziliano, and T. Blu, “Sampling signals with finite rate of innovation,” IEEE Trans. Signal Process., vol. 50, no. 6, pp. 1417–1428, 2002.
[5] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, “Sparse sampling of signal innovations,” IEEE Signal Process. Mag., vol. 25, no. 2, pp. 31–40, 2008.
[6] A. Cadzow, “Signal enhancement—a composite property mapping algorithm,” IEEE Trans. Acoust., vol. 36, no. 1, pp. 49–62, 1988.
[7] L. Condat, A. Hirabayashi, L. Condat, and A. Hirabayashi, “Cadzow denoising upgraded: A new projection method for the recovery of discrete pulses from noisy linear measurements,” 2015.
[8] P. L. Combettes and J.-C. Pesquet, “Proximal-Jetting methods in signal processing,” in Fixed-point algorithms for inverse problems in science and engineering. Springer, 2011, pp. 185–212.
[9] H. Pan, T. Blu, and M. Vetterli, “Towards generalized FRI sampling with an application to source resolution in radioastronomy,” IEEE Trans. Signal Process., vol. 65, no. 4, pp. 821–835, 2017.
[10] N. Parikh, S. Boyd et al., “Proximal algorithms,” Foundations and Trends in Optimization, vol. 1, no. 3, pp. 127–239, 2014.
[11] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM journal on imaging sciences, vol. 2, no. 1, pp. 183–202, 2009.
[12] B. Gu, D. Wang, Z. Huo, and H. Huang, “Inexact proximal gradient methods for non-convex and non-smooth optimization,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[13] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCan, and M. Unser, “Context-based projected gradient descent for consistent ct image reconstruction,” IEEE transactions on medical imaging, vol. 37, no. 6, pp. 1440–1453, 2018.
[14] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, “Plug-and-play priors for model based reconstruction,” in 2013 IEEE Global Conference on Signal and Information Processing. IEEE, 2013, pp. 945–948.
[15] E. K. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, “Plug-and-play methods provably converge with properly trained denoisers,” arXiv preprint arXiv:1905.05406, 2019.
[16] M. Simeoni, A. Besson, P. Hurley, and M. Vetterli, “Pyoneer: continuous recovery of non-bandlimited periodic signals with finite rates of innovation,” https://github.com/matthieumeo/pyoneer, accessed: 2020-04-16.
[17] R. Escalante and M. Raydan, Alternating projection methods. SIAM, 2011.
[18] J. von Neumann, “The geometry of orthogonal spaces, functional operators-vol. ii,” Annals of Math. Studies, vol. 22, 1950.
[19] H. Bauschke, F. Deutsch, H. Hundal, and S.-H. Park, “Accelerating the convergence of the method of alternating projections,” Transactions of the American Mathematical Society, vol. 355, no. 9, pp. 3433–3461, 2003.
[20] I. Halperin, “The product of projection operators,” Acta Sci. Math.(Szeged), vol. 23, no. 1, pp. 96–99, 1962.
[21] H. H. Bauschke and J. M. Borwein, “Dykstra’s alternating projection algorithm for two sets,” Journal of Approximation Theory, vol. 79, no. 3, pp. 418–443, 1994.
[22] A. S. Lewis, D. R. Luke, and J. Malick, “Local linear convergence for alternating and averaged nonconvex projections,” Foundations of Computational Mathematics, vol. 9, no. 4, pp. 485–513, 2009.
[23] A. S. Lewis and J. Malick, “Alternating projections on manifolds,” Mathematics of Operations Research, vol. 33, no. 1, pp. 216–234, 2008.
[24] F. Andersson and M. Carlsson, “Alternating projections on low-dimensional manifolds.”
[25] M. K. Tam, “Alternating projection methods failure in the absence of convexity.”
[26] C. Riche de Prony, “Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastique et sur celles de la force expansive de la vapeur de l’eau et de la vapeur de l’alcool, à différentes températures,” J. de l’École Polytechnique, vol. 1, pp. 24–76, 1795.
[27] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” Psychometrika, vol. 1, no. 3, pp. 211–218, 1936.
[28] H. Pan, M. Simeoni, P. Hurley, T. Blu, and M. Vetterli, “Leap: Looking beyond pixels with continuous-space estimation of point sources,” Astronomy & Astrophysics, vol. 608, p. A136, 2017.
[29] T. F. Chan and C.-K. Wong, “Convergence of the alternating minimization algorithm for blind deconvolution,” Linear Algebra and its Applications, vol. 316, no. 1-3, pp. 259–285, 2000.
[30] H. Li and Z. Lin, “Accelerated proximal gradient methods for nonconvex programming,” in Advances in neural information processing systems, 2015, pp. 379–387.
[31] F. Andersson, M. Carlsson, and P.-A. Ivert, “A fast alternating projection method for complex frequency estimation,” arXiv preprint arXiv:1107.2028, 2011.
[32] Q. Yao, J. T. Kwok, F. Gao, W. Chen, and T.-Y. Liu, “Efficient inexact proximal gradient algorithm for nonconvex problems,” arXiv preprint arXiv:1612.09609, 2016.
[33] H. Pan, T. Blu, M. Vetterli, and F. Dümweg, “Unified algorithmic framework for reconstructing signals with finite rate of innovation,” https://github.com/hanjiepan/FRI_pkg, accessed: 2019-10-18.
[34] E. Jones, T. Oliphant, P. Peterson et al., “SciPy: Open source scientific tools for Python,” http://www.scipy.org/, 2001–Now, accessed: 2019-10-18.
[35] H. W. Kuhn, “The hungarian method for the assignment problem,” Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.
[36] G. H. Golub and C. F. Van Loan, Matrix computations. JHU press, 2012, vol. 3.
[37] R. B. Lehoucq and D. C. Sorensen, “Deflation techniques for an implicitly restarted Arnoldi iteration,” SIAM Journal on Matrix Analysis and Applications, vol. 17, no. 4, pp. 789–821, 1996.
[38] S. Diamond and S. Boyd, “Matrix-free convex optimization modeling,” in Optimization and its applications in control and data sciences. Springer, 2016, pp. 221–264.
[39] D. P. Woodruff et al., “Sketching as a tool for numerical linear algebra,” Foundations and Trends in Theoretical Computer Science, vol. 10, no. 1–2, pp. 1–157, 2014.
[40] P. Indyk, A. Vakilian, and Y. Yuan, “Learning-based low-rank approximations,” in Advances in Neural Information Processing Systems, 2019, pp. 7400–7410.
[41] A. B. Taylor, J. M. Hendrickx, and F. Glineur, “Exact worst-case convergence rates of the proximal gradient method for composite convex minimization,” Journal of Optimization Theory and Applications, vol. 178, no. 2, pp. 455–476, 2018.
[42] A. Jung, “A fixed-point of view on gradient methods for big data,” Frontiers in Applied Mathematics and Statistics, vol. 3, p. 18, 2017.
[43] F. R. Deutsch, Best approximation in inner product spaces. Springer Science & Business Media, 2012.
[44] D. R. Luke, “Finding best approximation pairs relative to a convex function.”
APPENDIX A
THE TOEPLITZIFICATION OPERATOR AND CONVOLUTIONS

Consider the Toeplitzification operator defined in (2). When multiplied with a vector \( u = [u_1, \ldots, u_{P+1}] \in \mathbb{C}^{P+1} \), the matrix \( T_P(x) \) returns the valid part of the convolution between the two zero-padded sequences:

\[
\tilde{x} := [\ldots, 0, x_{-M}, \ldots, x_0, \ldots, x_M, 0, \ldots] \in \mathbb{C}^Z
\]

and

\[
\tilde{u} := [\ldots, 0, u_1, \ldots, u_{P+1}, 0, \ldots] \in \mathbb{C}^Z.
\]

Indeed,

\[
(\tilde{x} \ast \tilde{u})[k] := \sum_{j \in \mathbb{Z}} \tilde{x}_{j} \tilde{u}_{j-k} = \sum_{j=1}^{P+1} \tilde{x}_{j} u_{j-k}, \quad k \in \mathbb{Z}.
\]

The valid part corresponds to the indices \( k \in \mathbb{Z} \) for which all the terms in the summation are nonzero. This is the case when

\[
k \in \{-M + P + 1, \ldots, M + 1\}.
\]

Consider \( i = k + M - P \) we get that the valid part of the convolution is given by

\[
(\tilde{x} \ast \tilde{u})[i - M + P] = \sum_{j=1}^{P+1} x_{-M+j} u_j,
\]

which corresponds precisely to \( T_P(x)u \).

Using similar arguments, it is easy to show that the multiplication of \( T_P(x)^H \in \mathbb{C}^{(P+1) \times (N-P)} \) with a vector \( v = [v_1, \ldots, v_{N-P}]^T \in \mathbb{C}^{N-P} \) returns the valid part of the cross-correlation

\[
(\tilde{x} \ast \tilde{v})[k] := \sum_{j \in \mathbb{Z}} \tilde{x}_{j} \tilde{v}_{j+k}, \quad k \in \mathbb{Z},
\]

between \( \tilde{x} \) and the zero-padded sequence

\[
\tilde{v} := [\ldots, 0, v_1, \ldots, v_{N-P}, 0, \ldots] \in \mathbb{C}^Z.
\]

This time however, the valid part corresponds to the indices:

\[
k \in \{M + 1 - P, \ldots, M + 1\}.
\]

APPENDIX B
PROOF OF PROPOSITION 1

Consider a matrix \( H \in \mathbb{C}^{(N-P) \times (P+1)} \) and the following Frobenius inner product

\[
\langle T_P(x), H \rangle_F = \text{tr} \left( T_P^H(x)H \right)
\]

\[
= \sum_{i=1}^{N-P} \sum_{k=1}^{P+1} T_P(x)_i k H_{ik}
\]

\[
= \sum_{k=1}^{N-P} \sum_{i=k+P}^{N-1} x_{-M+i} \left( \sum_{j=k+s-P} H_{ik} \right)
\]

\[
(43)
\]

The term \( \sum_{i=k+s-P}^{N-1} H_{ik} \) sums the elements of \( H \) along lines with equation \( i = k + (s - P) \). These lines have slope 1 and intercept \( b = s - P \). Notice that these lines have nonnull intersection with the lattice \( \mathbb{Z}^2 \) for \( b \in [-P, N - P - 1] \). Indeed, the two extreme cases occur when the lines hit the points \((1, N - P)\) and \((P + 1, 1)\). This happens respectively when \( 1 + b = N - P \) and \( b = N - P - 1 \) and \( P + 1 + b = 1 \) \( b = -P \). Since \( s \in [0, N - 1] \) the intercept \( b \) varies indeed in the range \([-P, N - P - 1]\) and each term in the summation is nonnull. The summation \( \sum_{i=k+s-P}^{N-1} H_{ik} \) corresponds then to summing across each diagonal of \( H \). We finally get:

\[
\langle T_P(x), H \rangle_F = \langle x, T_P^*(H) \rangle,
\]

with

\[
T_P^* : \mathbb{C}^{(N-P) \times (P+1)} \rightarrow \mathbb{C}^N
\]

\[
H \mapsto H_{ik}, \quad j = 1, \ldots, N.
\]

APPENDIX C
PROOF OF PROPOSITION 2

From (3) and the definition of \( T_P \), it is straightforward to observe that the operator \( \Gamma = T_P^* T_P : \mathbb{C}^N \rightarrow \mathbb{C}^N \) is a diagonal matrix, with diagonal entries given by:

\[
\Gamma_{i,i} = \begin{cases} i & \text{for } i \leq P \\ P + 1 & \text{for } P < i \leq N - P \\ N + 1 - i & \text{for } N - P < i \leq N \end{cases}
\]

(44)

The operator \( T_P^\dagger = \Gamma^{-1} T_P^* \) is hence a left inverse for \( T_P \):

\[
T_P T_P^\dagger = \Gamma^{-1} T_P^* T_P = (T_P^* T_P)^{-1} T_P^* T_P = I_N.
\]

Moreover, the latter is actually the pseudoinverse of \( T_P \). Indeed, we have trivially:

\[
T_P T_P^\dagger T_P = T_P, \quad T_P^\dagger T_P T_P^\dagger = T_P^\dagger, \quad (T_P T_P^\dagger)^* = T_P T_P^\dagger.
\]

Finally, we have

\[
(T_P T_P^\dagger)^* = T_P^\dagger \Gamma^{-1} T_P^* = T_P^\dagger T_P^\dagger.
\]

\(19\) For \( n, m \in \mathbb{Z}, n < m \), we denote by \([n, m]\) the 1D lattice \([n, m] \cap \mathbb{Z}\).
since $\Gamma$ is diagonal and hence symmetric. $T_P^\dagger$ verifies thus the definition of the pseudoinverse of $T_P$.  

\[\Box\]

**Appendix D**

**Projecton on the Subspace of Toeplitz Matrices**

The operator $T_P$ is actually a surjection onto the subspace $T_P \subset \mathbb{C}^{(N-P)(P+1)}$ of rectangular Toeplitz matrices with size $(N - P) \times (P + 1)$. Indeed, it is easy to see that every such matrix can be written as in (2) for some generator $x \in \mathbb{C}^N$. Moreover, we have from (45) that $T_P^\dagger T_P = I_N$ and hence from [1, Theorem 2.29], $T_P^\dagger T_P^\dagger$ is a projection operator onto the range $T_P$ of $T_P$. Since $T_P^\dagger T_P^\dagger$ is moreover self-adjoint from (46), it is actually an orthogonal projection operator, which achieves the proof.

\[\Box\]

**Appendix E**

**Proofs of Theorems 2 and 3**

The proofs of Theorems 2 and 3 rely on the following lemma, adapted from [30, Theorem 1], which establishes the convergence of PGD in a general setup:

**Lemma 1.** Consider the norm $\|x\| := \sqrt{\langle x, x \rangle}$, $x \in \mathbb{R}^n$, induced by some inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$. Consider moreover the general problem:

$$\min_{x \in \mathbb{R}^n} \Phi(x) := F(x) + H(x),$$

(47)

where $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $H : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are potentially non convex functions such that:

(A) $F$ is a proper function, i.e. its domain is non empty, differentiable and with Lipschitz continuous gradient for some Lipschitz constant $0 \leq \beta < +\infty$, $\|\nabla F(x) - \nabla F(y)\| \leq \beta \|x - y\|$, $\forall x, y \in \mathbb{R}^n$.

(B) $H$ is a proper and lower semicontinuous (lWSC) function (see supplementary material of [30] for a definition), potentially non smooth.

(C) $\Phi = F + H$ is coercive, i.e. $\Phi$ is bounded from below and

$$\lim_{\|x\| \to +\infty} \Phi(x) = +\infty.$$

Then, the iterates $\{x_k\}_{k \in \mathbb{N}}$ generated by the proximal gradient descent applied to (47):

$$x_{k+1} \in \text{prox}_{\tau H}(x_k - \tau \nabla F(x_k)), \quad k \geq 0,$$

(48)

with $\tau < 1/\beta$ and $x_0 \in \mathbb{R}^n$, are bounded. Moreover, any limit point $x_* \in \{x_k\}_{k \in \mathbb{N}}$ is a local minimum of $\Phi$.

**Proof.** The lemma is easily shown by specifying the proof of [30, Theorem 1] for the non-accelerated case. For the sake of completeness, it is provided hereafter.

From the definition of the proximal operator,

$$\text{prox}_{\tau H}(x) : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n), \quad \begin{cases} x \mapsto \arg \min_{z \in \mathbb{R}^n} \frac{1}{2\tau} \|x - z\|^2 + H(z), \end{cases}$$

we can reinterpret (48) as

$$x_{k+1} \in \arg \min_{z \in \mathbb{R}^n} \frac{1}{2\tau} \|x - z\|^2 + \langle \nabla F(x_k), z - x_k \rangle + H(z).$$

(49)

We have hence

$$\frac{1}{2\tau} \|x_{k+1} - x_k\|^2 + \langle \nabla F(x_k), x_{k+1} - x_k \rangle + H(x_{k+1}) \leq H(x_k).$$

From the Lipschitz continuity of $\nabla F$ we have moreover

$$\Phi(x_{k+1}) \leq H(x_{k+1}) + F(x_k) + \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{\beta}{2} \|x_{k+1} - x_k\|^2 \leq H(x_k) - \frac{1}{2\tau} \|x_{k+1} - x_k\|^2 - \langle \nabla F(x_k), x_{k+1} - x_k \rangle + F(x_k) + \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{\beta}{2} \|x_{k+1} - x_k\|^2 \leq \Phi(x_k) - \frac{1}{2\tau} - \frac{\beta}{2} \|x_{k+1} - x_k\|^2. \quad (50)$$

Since $\tau < 1/\beta$, we have hence $(1/(2\tau - \beta/2)) \geq 0$ and

$$\Phi(x_{k+1}) \leq \Phi(x_k) \leq \Phi(x_0), \quad \forall k \geq 1.$$

The sequence $\{\Phi(x_k)\}_{k \in \mathbb{N}}$ is hence bounded and since $\Phi$ is coercive so is $\{x_k\}_{k \in \mathbb{N}}$. The sequence $\{x_k\}_{k \in \mathbb{N}}$ admits hence limit points. Moreover, since $\Phi(x_k)$ is decreasing and bounded from below, it takes the same value $\Phi_* \in \mathbb{R}$ at all of these limit points. Summing (50), we obtain hence:

$$\left(\frac{1}{2\tau} - \frac{\beta}{2}\right) \sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 \leq \Phi(x_0) - \Phi_* < +\infty.$$

(51)

Since $\tau < 1/\beta$, we have necessarily $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 < +\infty$, which yields

$$\lim_{k \to +\infty} \|x_{k+1} - x_k\| = 0. \quad (51)$$

From the optimality condition (49) and Items 1 and 3 of Proposition 1 of the supplementary material of [30], we have moreover

$$0_n \in \nabla F(x_k) + \frac{1}{\tau}(x_{k+1} - x_k) + \partial H(x_{k+1}) = \partial \Phi(x_{k+1}) - \nabla F(x_k) + \frac{1}{\tau}(x_{k+1} - x_k), \quad (52)$$

where $\partial H : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ and $\partial \Phi : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ denote the (set-valued) subdifferential operators of $H$ and $\Phi$ respectively (see Definition 2 of the supplementary material of [30]).

Equation (52) can moreover be rewritten as

$$\nabla F(x_{k+1}) - \nabla F(x_k) - \frac{1}{\tau}(x_{k+1} - x_k) \in \partial \Phi(x_{k+1}).$$

Furthermore, from the Lipschitz continuity of $F$, we have

$$\|\nabla F(x_{k+1}) - \nabla F(x_k) - \frac{1}{\tau}(x_{k+1} - x_k)\| \leq \left(\beta + \frac{1}{\tau}\right) \|x_{k+1} - x_k\|, \quad (53)$$

and hence from (51):

$$\lim_{\|x\| \to +\infty} \|\nabla F(x_{k+1}) - \nabla F(x_k) - \frac{1}{\tau}(x_{k+1} - x_k)\| = 0. \quad (53)$$
Let \( \{ x_k \}_{k \in \mathbb{N}} \) be a convergent subsequence of \( \{ x_k \}_{k \in \mathbb{N}} \), with limit \( x^* \). Then, we have from (53) and Item 2 of Proposition 1 of the supplementary material of [30]:

\[
0_n = \lim_{j \to +\infty} \partial \Phi(x_k) = \partial \Phi(x^*),
\]

which completes the proof. \( \square \)

We now show Theorems 2 and 3 by applying Lemma 1 to the implicit genFRI problem in unconstrained form (54)

\[
\min_{x \in \mathbb{C}^N} \|Gx - y\|^2_2 + \nu_H \langle T_P(x)\rangle + \nu_{\rho} \langle x \rangle.
\]  

(54)

To do so, we must first convert (54) into an optimisation problem of the form (47), defined over \( \mathbb{R}^N \) for some \( n \in \mathbb{N} \). We achieve this by proceeding as in [2, Section 7.8] and identifying \( \mathbb{C}^N \) with \( \mathbb{R}^{2N} \) (respectively \( \mathbb{C}^L \) with \( \mathbb{R}^{2L} \)) in the canonical way

\[
x \in \mathbb{C}^N \leftrightarrow \tilde{x} := \begin{bmatrix} \Re(x) \\ \Im(x) \end{bmatrix} \in \mathbb{R}^{2N},
\]

where \( \Re \) and \( \Im \) denote the real and imaginary parts respectively. Such an identification makes the canonical inner products and norms on \( \mathbb{C}^N \) and \( \mathbb{R}^{2N} \) (respectively \( \mathbb{C}^L \) and \( \mathbb{R}^{2L} \)) consistent with one another, i.e. for all \( x, z \in \mathbb{C}^N \), we have

\[
\langle x, z \rangle_{\mathbb{C}^N} = z^H x = \Re(z)^T \Re(x) + \Im(z)^T \Im(x) = \tilde{x}^T \tilde{z} \in \mathbb{R}^{2N},
\]

and

\[
\|x\|_{\mathbb{C}^N} = \sqrt{x^H x} = \sqrt{\|\Re(x)\|_{\mathbb{R}^{2N}}^2 + \|\Im(x)\|_{\mathbb{R}^{2N}}^2} = \|x\|_{\mathbb{R}^{2N}}.
\]

Still following [2, Section 7.8], we moreover identify the linear map \( G : \mathbb{C}^N \to \mathbb{C}^L \) with a linear map \( \hat{G} : \mathbb{R}^{2N} \to \mathbb{R}^{2L} \) with matrix representation:

\[
\hat{G} := \begin{bmatrix} \Re(G) & -\Im(G) \\ \Im(G) & \Re(G) \end{bmatrix} \in \mathbb{R}^{2L \times 2N}.
\]

Again, it is easy to show that the two operators are consistent, in the sense that

\[
\hat{G}x = \tilde{G}\tilde{x}, \quad \text{and} \quad \hat{G}^H x = \tilde{G}^T \tilde{x}, \quad \forall x \in \mathbb{C}^N.
\]

Similarly, the Toeplitzification operator \( T_P : \mathbb{C}^N \to \mathbb{C}^{(N-P)(P+1)} \) is identified with the linear operator \( \hat{T}_P : \mathbb{R}^{2N} \to \mathbb{R}^{2L} \) defined as

\[
\hat{T}_P(\tilde{x}) := T_P(\Re(x)) + j T_P(\Im(x)), \quad \forall x \in \mathbb{C}^N,
\]

where \( j \) is the complex 2-root of unity. From the linearity of \( T_P \), this definition yields indeed \( T_P(x) = \hat{T}_P(\tilde{x}) \). Finally, the \( \Gamma \)-ball \( \mathbb{B}_{\rho}^\Gamma \subset \mathbb{C}^N \) is identified with

\[
\mathbb{B}_{\rho}^\Gamma = \left\{ \tilde{x} \in \mathbb{R}^{2N} : \|\tilde{x}\|_{\Gamma} \leq \rho \right\},
\]

where \( \tilde{\Gamma} \in \mathbb{R}^{2N \times 2N} \) is a positive definite and diagonal matrix defined as

\[
\tilde{\Gamma} := \begin{bmatrix} \Gamma & 0_{N \times N} \\ 0_{N \times N} & \Gamma \end{bmatrix}.
\]

Again, we trivially have \( \|\tilde{x}\|_{\Gamma} = \|x\|_{\Gamma} \) and hence \( x \in \mathbb{B}_{\rho}^\Gamma \Leftrightarrow \tilde{x} \in \mathbb{B}_{\rho}^\Gamma \) for all \( x \in \mathbb{C}^N \).

In summary, the optimisation problem (54) is hence equivalent to the following optimisation problem with search space \( \mathbb{R}^{2N} \):

\[
\min_{\tilde{x} \in \mathbb{R}^{2N}} \|\hat{G}\tilde{x} - \hat{y}\|^2_{2L} + \nu_H \langle \hat{T}_P(\tilde{x})\rangle + \nu_{\rho} \langle \tilde{x} \rangle.
\]

(55)

Letting \( \hat{F}(\tilde{x}) := \|\hat{G}\tilde{x} - \hat{y}\|^2_{2L} \) and \( \hat{H}(\tilde{x}) := \nu_H \langle \hat{T}_P(\tilde{x})\rangle + \nu_{\rho} \langle \tilde{x} \rangle \), we have \( \hat{F} : \mathbb{R}^{2N} \to \mathbb{R}_+ \) and \( \hat{H} : \mathbb{R}^{2N} \to [0, +\infty) \), so that (55) is indeed of the form (47). We must now verify assumptions (A), (B) and (C) of Lemma 1:

(A) \( \hat{F} \) is proper, differentiable and \( \nabla \hat{F} \) Lipschitz continuous. \( \hat{F} \) is proper since

\[
\nabla \hat{F}(\tilde{x}) = 2\hat{G}^T (\hat{G}\tilde{x} - \hat{y}) = \nabla F(x), \quad \tilde{x} \in \mathbb{R}^{2N}.
\]

(56)

The gradient (56) is moreover \( \beta \)-Lipschitz continuous with respect to the norm \( \| \cdot \|_{\tilde{\Gamma}} \) on \( \mathbb{R}^{2N} \), and its Lipschitz constant is given by:

\[
\beta = 2\|\hat{G}^T \hat{G}\|_{\tilde{\Gamma}} = \sup \left\{ 2\|\hat{G}^T \hat{G}\|_{\tilde{\Gamma}} : \tilde{x} \in \mathbb{R}^{2N}, \|\tilde{x}\|_{\Gamma} = 1 \right\} = \sup \left\{ 2\|\hat{G}^T \hat{G}\|_{\tilde{\Gamma}} : x \in \mathbb{C}^N, \|x\|_{\Gamma} = 1 \right\} = \beta < +\infty.
\]

(57)

(B) \( \hat{H} \) is proper and lower semicontinuous. \( \hat{H} \) is proper since for all \( \rho > 0 \), and \( K \in \mathbb{N} \),

\[
\hat{H}(\mathbf{0}_{2N}) = \nu_H (0_{(N-P)(P+1)}) + \nu_{\rho} (\mathbf{0}_{2N}) = 0 < +\infty.
\]

The indicator functions are moreover lower semicontinuous since the sets \( \mathbb{H}_0 \) and \( \mathbb{F}_{\rho}^\Gamma \) are both closed. Since \( T_P \) is a bounded linear operator, it is continuous and hence \( \hat{H} \) is indeed lower semicontinuous as composition between continuous and lower semicontinuous functions. \( \square \)

(C) \( \Phi = \hat{F} + \hat{H} \) is coercive. It is easy to see that \( \Phi = \hat{F} + \hat{H} \geq 0 \). To show that \( \Phi \) is coercive, it is hence sufficient to show that

\[
\lim_{\|\tilde{x}\|_{\tilde{\Gamma}} \to +\infty} \Phi(\tilde{x}) = +\infty.
\]

To this end, we distinguish two cases, which correspond respectively to the assumptions of Theorems 2 and 3:

1. \( \rho \in [0, +\infty) \): in this case \( \Phi \) is trivially coercive since

\[
t_{\beta,\rho}(\tilde{x}) = +\infty, \quad \forall \|\tilde{x}\|_{\tilde{\Gamma}} \geq \rho.
\]

2. \( \rho = +\infty \) and \( G \) injective: When \( \rho = +\infty \), the term \( t_{\beta,\rho} \) is always null and \( \Phi \) simplifies to

\[
\Phi(\tilde{x}) = \|\hat{G}\tilde{x} - \hat{y}\|^2_{2L} + \nu_H \langle \hat{T}_P(\tilde{x})\rangle, \quad \tilde{x} \in \mathbb{R}^{2N}.
\]

From [2, Section 7.8], we have moreover that

\[
\det(\hat{G}^T \hat{G}) = |\det(G^T G)|^2 \neq 0,
\]

(58)
since \( G \) is injective by assumption. From the reverse triangle inequality, we have hence
\[
\| \hat{G} \hat{x} - \hat{y} \|_{2, \mathbb{L}} \geq \sigma_{\min} \| \hat{x} \|_{2, \mathbb{E}} - \| \hat{y} \|_{2, \mathbb{L}}, \quad \forall \hat{x} \in \mathbb{R}^N.
\]
where \( \sigma_{\min} = \sqrt{\lambda_{\min}(\hat{G}^T \hat{G})} > 0 \) is the square root of the eigenvalue of \( \hat{G}^T \hat{G} \) with lowest magnitude, which is non null from (58). From the equivalence of norms in finite dimensions, there exist moreover \( c_1, c_2 > 0 \) such that
\[
c_1 \| \hat{x} \|_\Gamma \leq \| \hat{x} \|_{2, \mathbb{L}} \leq c_2 \| \hat{x} \|_\Gamma, \quad \forall \hat{x} \in \mathbb{R}^N.
\]
This yields
\[
\| \hat{G} \hat{x} - \hat{y} \|_{2, \mathbb{L}} \geq \sigma_{\min} c_1 \| \hat{x} \|_\Gamma - \| \hat{y} \|_{2, \mathbb{L}}, \quad \forall \hat{x} \in \mathbb{R}^N,
\]
and hence
\[
\lim_{\| \hat{x} \|_\Gamma \to +\infty} \| \hat{G} \hat{x} - \hat{y} \|_{2, \mathbb{L}} \geq \lim_{\| \hat{x} \|_\Gamma \to +\infty} \sigma_{\min} c_1 \| \hat{x} \|_\Gamma = +\infty,
\]
which shows that \( \hat{\Phi} \) is indeed coercive.

\[\square\]

We can hence apply Lemma 1, to show that the iterates \( \{\hat{x}_k\}_{k \in \mathbb{N}} \subset \mathbb{R}^N \) generated by PGD applied to (55):
\[
\hat{x}_{k+1} = \text{prox}_{\tau H} \left( \hat{x}_k - \tau \nabla \hat{F}(\hat{x}_k) \right), \quad (59)
\]
with \( \tau < 1/\beta \) and \( \hat{x}_0 \in \mathbb{R}^N \), are bounded. Moreover, any limit point \( \hat{x}_* \) of \( \{\hat{x}_k\}_{k \in \mathbb{N}} \) is a critical point of (55), i.e. \( \Omega_{2N} \subset \partial \Phi(\hat{x}_*) \).

Observe finally, that the iterations (59) can be rewritten in complex form as
\[
x_{k+1} = \text{prox}_{\tau H} \left( x_k - \tau \nabla F(x_k) \right), \quad (60)
\]
with \( \tau < 1/\beta \) and \( x_0 \in \mathbb{C}^N \), and where we have used (56), (57) and
\[
\text{prox}_{\tau H}(\hat{x}) = \arg \min_{\hat{x} \in \mathbb{R}^N} \frac{1}{2\tau} \| \hat{x} - \hat{z} \|_2^2 + \hat{H}(\hat{z}),
\]
which follows trivially from the previous identifications. By identification and equivalence between the real and complex optimisation problems (55) and (54), we can hence conclude that limit points of the iterates \( \{x_k\}_{k \in \mathbb{N}} \subset \mathbb{C}^N \) generated by (60) are critical points of (54), which achieves the proof.

\[\square\]

**APPENDIX F**

**PROOF OF PROPOSITION 3**

Recall the definition of the proximal set associated to a point \( x \in \mathbb{C}^N \):
\[
\text{prox}_{\tau H}(x) = \arg \min_{z \in \mathbb{C}^N} \left\{ \frac{1}{2\tau} \| x - z \|_2^2 + \tau \mathcal{H}_k(T_p(z)) + \tau \mathcal{B}_p(z) \right\}, \quad (61)
\]
When mapped via the Toeplitzification operator \( T_p \), the proximal set (61) becomes
\[
T_p \left( \text{prox}_{\tau H}(x) \right) = \left\{ \right. \left. \begin{array}{l}
T_p(\hat{x}), \ x = \text{prox}_{\tau H}(x) \\
\hat{x} \in \mathbb{T}_p, \ T_p^*(\hat{x}) = \text{prox}_{\tau H}(x) \end{array} \right\} = \left\{ \right. \left. \begin{array}{l}
\arg \min_{z \in \mathbb{T}_p \cap \mathcal{H}_k} \left\{ \frac{1}{2\tau} \| T_p(z) - x \|_2^2 + \tau \mathcal{H}_k (z) + \tau \mathcal{B}_p \left( T_p^*(z) \right) \right\} \\
\arg \min_{z \in \mathbb{T}_p \cap \mathcal{H}_k} \left\{ \frac{1}{2\tau} \| T_p(z) - x \|_2^2 + \tau \mathcal{B}_p \left( T_p^*(z) \right) \right\} \right\}, \quad (62)
\]
where we have used the fact that \( T_p^* T_p(z) = z \) for all \( z \in \mathbb{C}^N \). We have moreover, \( \forall Z \in \mathbb{T}_p \):
\[
\| T_p^*(z) - x \|_2 = \| \Gamma \Gamma^H (z - T_p(x), Z)_F \|_F = \| \Gamma \Gamma^H (Z, T_p(x), Z)_F \|_F = \| Z \|_F + \| T_p(x) \|_2\| Z \|_F = Z - T_p(x) \|_2 \| Z \|_F = \| Z - T_p(x) \|_F, \quad (63)
\]
where we have used the fact that \( \Gamma^* = \Gamma H = T_p \Gamma P \) and \( T_p \Gamma^{-1} P \Gamma = \Pi_{T_p} \) (see Appendices C and D). With similar arguments, we have \( \forall Z \in \mathbb{T}_p \):
\[
\| T_p^*(z) \|_2 = \| \Gamma \Gamma^H (z, T_p(z))_2 \|_F \leq \rho \Rightarrow \sqrt{\| (z, z)_F \|_F} \leq \rho \Rightarrow \| Z \|_F \leq \rho, \quad \rho \leq \frac{1}{\tau}, \quad (64)
\]
so that
\[
\mathcal{B}_p \left( T_p^*(z) \right) = \mathcal{B}_p (z), \quad \forall Z \in \mathbb{T}_p, \quad (65)
\]
where \( \mathcal{B}_p := \left\{ Z \in \mathbb{C}^{(N - P)(P + 1)} : \| Z \|_F \leq \rho \right\} \). Plugging (63) and (64) into (62) hence yields
\[
T_p \left( \text{prox}_{\tau H}(x) \right) = \left\{ \right. \left. \begin{array}{l}
\arg \min_{z \in \mathbb{T}_p \cap \mathcal{H}_k} \left\{ \frac{1}{2\tau} \| T_p(z) - x \|_2^2 + \tau \mathcal{B}_p \left( T_p^*(z) \right) \right\} \\
\arg \min_{z \in \mathbb{T}_p \cap \mathcal{H}_k} \left\{ \frac{1}{2\tau} \| Z - T_p(x) \|_2^2 + \tau \mathcal{B}_p \left( z \right) \right\} \right\} = \left\{ \right. \left. \begin{array}{l}
\Pi_{T_p \cap \mathcal{H}_k} \cap \mathcal{B}_p, \quad (62)
\end{array} \right\}.
\]

Recall the definition of the proximal set associated to a point \( x \in \mathbb{C}^N \):
Using the fact that $T_P^T T_P = I_N$ we can finally rewrite (65) as
\[
\text{prox}_{\mathcal{H}}^T(x) = T_P^{T} \Pi_{T \cap \mathcal{H} \cap \overline{\mathcal{B}^r}} T_P(x),
\]
which completes the proof. \qed

**Appendix G**

**Proof of Theorem 4**

The proof of Theorem 4 relies on the four lemmas hereafter. The first lemma shows that gradient descent is Lipschitz continuous, and exhibits step size ranges for which it is also a contraction. This is a famous result in optimisation [41], [42], traditionally stated in terms of the $\ell_2$ canonical norm. Lemma 2 in contrast assumes the $\Gamma$-norm as underlying norm, since the latter is more natural for our particular problem.

**Lemma 2** (Contractive Gradient Descent). Let $G \in \mathbb{C}^{L \times N}$ be injective, and $\Gamma$ be the diagonal and definite positive matrix defined in (44). Define
\[
\alpha := 2 \lambda_{\text{min}} \left( \Gamma^{1/2} G^H \Gamma^{-1/2} \right), \quad (66)
\]
\[
\beta := 2 \lambda_{\text{max}} \left( \Gamma^{1/2} G^H \Gamma^{-1/2} \right), \quad (67)
\]
where $\lambda_{\text{min}}(M)$ and $\lambda_{\text{max}}(M)$ denote the minimum and maximum eigenvalue of a matrix $M$ respectively. Let $\tau \in \mathbb{R}$ be a positive constant and consider the linear map
\[
D_\tau : \mathbb{C}^N \to \mathbb{C}^N, \quad x \mapsto x - 2 \tau \Gamma^H (G x - y), \quad (68)
\]
for some $y \in \mathbb{C}^L$. Then, $D_\tau$ is Lipschitz continuous with respect to the norm induced by $\Gamma$:
\[
\| D_\tau(x) - D_\tau(z) \|_\Gamma \leq L_\tau \| x - z \|_\Gamma, \quad \forall x, z \in \mathbb{C}^N,
\]
with Lipschitz constant
\[
L_\tau = \max \{|1 - \tau \alpha|, |1 - \tau \beta|\}. \quad (69)
\]
Moreover, $D_\tau$ is contractive, i.e. $0 \leq L_\tau < 1$, for $0 < \tau < 2/\beta$, and $L_\tau$ is minimised for $\tau = 2/(\alpha + \beta)$.

**Proof.** We have
\[
\| D_\tau(x) - D_\tau(z) \|_\Gamma = \| (I_N - 2 \tau \Gamma^H G) (x - z) \|_\Gamma \\
\leq \| I_N - 2 \tau \Gamma^H G \|_\Gamma \| x - z \|_\Gamma, \\
= L_\tau \| x - z \|_\Gamma,
\]
where the Lipschitz constant $L_\tau := \| I_N - 2 \tau \Gamma^H G \|_\Gamma > 0$ is the operator norm of $I_N - 2 \tau \Gamma^H G$ induced by the $\Gamma$-norm on $\mathbb{C}^N$:
\[
\| I_N - 2 \tau \Gamma^H G \|_\Gamma = \sup_{\| x \|_\Gamma = 1} \| (I_N - 2 \tau \Gamma^H G) x \|_\Gamma \\
= \sup_{\| x \|_\Gamma = 1} \| \Gamma^{1/2} (I_N - 2 \tau \Gamma^H G) x \|_2 \\
= \sup_{\| x \|_2 = 1} \| \Gamma^{1/2} (I_N - 2 \tau \Gamma^H G) \Gamma^{-1/2} x \|_2 \\
= \| I_N - 2 \tau \Gamma^{1/2} \Gamma^H \Gamma^{-1/2} \|_\Gamma. \quad (70)
\]

Note that since $G$ is injective, $G^H G$ is positive definite and hence we easily get [42] that the eigenvalues of $I_N - 2 \tau \Gamma^{1/2} \Gamma^H \Gamma^{-1/2}$ are contained in the interval $[1 - \tau \beta, 1 - \tau \alpha]$, where $\beta \geq \alpha > 0$ are defined in (66) and (67) respectively. Its spectral norm is hence given by:
\[
\| I_N - 2 \tau \Gamma^{1/2} \Gamma^H \Gamma^{-1/2} \|_2 = \max \{|1 - \tau \alpha|, |1 - \tau \beta|\},
\]
which proves (69). Finally, the restriction on $\tau$ for $L_\tau$ to be smaller than one follows from basic algebra, and is discussed in [41]. \qed

The second lemma states that in a Hilbert space, orthogonal projection maps onto closed convex sets are non-expansive. This is a known result of approximation theory [43], [44].

**Lemma 3** (Non-Expansiveness of Closed Convex Projections). Let $\mathcal{H}$ be some Hilbert space with some inner-product norm $\| \cdot \|$ and $C \subset \mathcal{H}$ a closed, convex set. Then the orthogonal projection map onto $C$, defined as
\[
\Pi_C(x) = \arg \min_{z \in C} \| x - z \|, \quad \forall x \in \mathcal{H},
\]
is non-expansive, i.e.
\[
\| \Pi_C(x) - \Pi_C(z) \| \leq \| x - z \|, \quad \forall x, z \in \mathcal{H}.
\]

**Proof.** Lemma 3 is proven in [43, Theorem 5.5]. \qed

The third lemma states that the singular value projection map $\Pi_{H_k}$ is locally non-expansive in every neighbourhood of the manifold of matrices with rank exactly $k$.

**Lemma 4** (Local Non-Expansiveness of the Singular Value Projection). Let $\mathbb{C}^{m \times n}$ be the space of complex-valued rectangular matrices of size $m \times n$, and $\mathcal{H}_k \subset \mathbb{C}^{m \times n}$, $R_k \subset \mathbb{C}^{m \times n}$ the sets of matrices with rank at most and exactly $k \leq \max\{m,n\}$ respectively. Denote further by $\Pi_{H_k}$ the orthogonal projection map onto $\mathcal{H}_k$ given in (14). Then, for every $R \in R_k$, the map $\Pi_{H_k}$ is well-defined (single-valued) and locally non-expansive
\[
\| \Pi_{H_k}(X) - \Pi_{H_k}(Z) \|_F \leq \| X - Z \|_F, \quad \forall X, Y \in U,
\]
for some neighbourhood $U \ni R$.

**Proof.** Since $R_k$ is dense in $H_k$ [24, Proposition 2.1], we have $\Pi_{H_k} = \Pi_{R_k}$ in a neighbourhood $W$ of every $R \in R_k$ (see [23, Example 2.3] for a detailed proof of this fact). Moreover, [45, Lemma 3] tells us that, for every $R \in R_k$, $\Pi_{R_k}$ is, in a neighbourhood $U \ni R$ such that $U \subset W$, well-defined (single-valued), continuous and differentiable, with gradient given by: $\nabla \Pi_{R_k}(R) = \Pi_{T_{R_k}}(R)$ where $T_{R_k}(R) \subset \mathbb{C}^{m \times n}$ is the tangent plane of the manifold $R_k$ in $R$ [see [23, Example 2.2]]. Since $\Pi_{R_k}(R)$ is by definition a linear subspace of $\mathbb{C}^{m \times n}$, the orthogonal projection operator $\Pi_{R_k}(R)$ is bounded with unit spectral norm. The map $\Pi_{R_k} = \Pi_{H_k}$ is consequently 1-Lipschitz continuous (i.e. non-expansive) with respect to the Frobenius norm in the neighbourhood $U$ of $R \in R_k$. \qed

The last lemma finally, makes use of Lemmas 3 and 4 to show that the denoising operator $H_2(x) = T_P^T [\Pi_{T_{R_k}}(R_k) \Pi_{R_k}] T_P(x)$ is locally non-expansive with respect to the $\Gamma$-norm:
Lemma 5 (Local Non-Expansiveness of Denoisers). Let $\mathbb{C}^{(N-P)\times(P+1)}$ be the space of complex-valued rectangular matrices of size $(N - P) \times (P + 1)$, $P \leq \lceil N/2 \rceil$, and $\mathcal{H}_K \subset \mathbb{C}^{(N-P)\times(P+1)}$, $\mathcal{R}_K \subset \mathbb{C}^{(N-P)\times(P+1)}$ the sets of matrices with rank at most and exactly $K \leq P$ respectively. Let

$$H_n(x) := T_p^\dagger [\Pi_{\mathcal{H}_K} \Pi_{\mathcal{R}_K}] \Pi_{\mathcal{R}_K}^\dagger T_p(x), \quad \forall x \in \mathbb{C}^N,$$

be the approximate proximal operator (34). Then, $H_n$ is locally well-defined (single-valued) and non-expansive with respect to the $\Gamma$-norm

$$\|H_n(x) - H_n(z)\|_\Gamma \leq \|x - z\|_\Gamma,$$

for all $x, z \in \mathbb{C}^N$ such that $T_p(x), T_p(z)$ are in some neighbourhood of some matrix $R \in \mathcal{R}_K$.

Proof. First, we have, for all $x, z \in \mathbb{C}^N$:

$$\|H_n(x) - H_n(z)\|_\Gamma = \left\| T_p^\dagger [D_n T_p(x) - D_n T_p(z)] \right\|_F,$$

where $D_n = [\Pi_{\mathcal{H}_K} \Pi_{\mathcal{R}_K}]^n$. Notice that for $X \in T_p$, we have

$$\left\| T_p^\dagger(X) \right\|_F^2 = (TT_p(X) T_p^\dagger(X))_2 = (T^{-1} T_p^\dagger(X))_2 = (X, T_p^\dagger(X))_F = (X, X)_F = \|X\|_F^2. $$

Since the range of $D_n$ is $T_p$, (71) becomes:

$$\|H_n(x) - H_n(z)\|_\Gamma = \left\| D_n T_p(x) - D_n T_p(z) \right\|_F.$$

Assuming now that $T_p(x)$ and $T_p(z)$ are in some neighbourhood of some point $R \in \mathcal{R}_K$, we can invoke Lemmas 3 and 4 recursively to obtain:

$$\|D_n T_p(x) - D_n T_p(z)\|_F \leq \|T_p(x) - T_p(z)\|_F = \|x - z\|_F,$$

where we have used: $\|T_p(x)\|_F^2 = \langle T_p(x), T_p(x) \rangle = \langle T_p^\dagger T_p(x), x \rangle = \|x\|_F^2, \quad \forall x \in \mathbb{C}^N$. We finally get

$$\|H_n(x) - H_n(z)\|_\Gamma \leq \|x - z\|_\Gamma,$$

for all $x, z \in \mathbb{C}^N$ such that $T_p(x), T_p(z)$ are in some neighbourhood of some matrix $R \in \mathcal{R}_K$. □

We are now ready to show Theorem 4. Let

$$U_{r,n}(x) := H_n(x - \tau \nabla F(x)) = H_n(D_r(x)), \quad x \in \mathbb{C}^N.$$

Then, for every $x, z \in \mathbb{C}^N$ such that $T_p(x), T_p(z)$ are in some neighbourhood of some matrix $R \in \mathcal{R}_K$, $U_{r,n}$ is locally Lipschitz continuous as composition between two (locally) Lipschitz continuous functions $H_n$ and $D_r$, see Lemmas 5 and 2 respectively. Moreover, the Lipschitz constant is the product of the Lipschitz constants of $H_n$ and $D_r$, 1 and $L_r$ in (69) respectively. We have therefore

$$\|U_{r,n}(x) - U_{r,n}(z)\|_\Gamma \leq L_r \|x - z\|_\Gamma,$$

for all $x, z \in \mathbb{C}^N$ such that $T_p(x), T_p(z)$ are in some neighbourhood of some matrix $R \in \mathcal{R}_K$. Finally, the restriction on $\tau$ for $L_r$ to be smaller than one results from Lemma 2. □

APPENDIX H

PROOF OF COROLLARY 2

First, we note that from Theorem 4, we have under the assumptions of the corollary that

$$\|x_k - x_k\|_\Gamma \leq \|U_{r,n}(x_k) - U_{r,n}(x_k)\|_\Gamma \leq L_r \|x_k - x_k\|_\Gamma,$$

for all $k \geq 1$ and hence by induction

$$\|x_k - x_k\|_\Gamma \leq L_r \|x_0 - x_0\|_\Gamma \quad \forall k \geq 1. \quad (72)$$

By assumption $0 < \tau < 2/\beta$ and therefore $0 < L_r < 1$. We deduce hence from (72) that $(x_k)_{k \in \mathbb{N}}$ is a Cauchy sequence. Let $j, k \in \mathbb{N}$ with $j > k$:

$$\|x_j - x_k\|_\Gamma \leq \sum_{m=k}^{j-1} \|x_{m+1} - x_m\|_\Gamma \leq \sum_{m=k}^{j-1} L_r^m \|x_1 - x_0\|_\Gamma \leq \frac{L_r^k}{1 - L_r} \|x_1 - x_0\|_\Gamma.$$

For every $\epsilon > 0$, we can choose a $J \in \mathbb{N}$ such that

$$L_r^j < \frac{\epsilon}{\|x_1 - x_0\|_\Gamma},$$

and hence for all $j > k > J$

$$\|x_j - x_k\|_\Gamma < \epsilon.$$

The sequence $(x_k)_{k \in \mathbb{N}}$ is hence a Cauchy sequence, and since $\mathbb{C}^N$ is complete, it converges towards a limit point $x_* \in \mathbb{C}^N$. We have moreover, since $U_{r,n}$ is continuous

$$x_* = \lim_{n \to \infty} x_k = \lim_{n \to \infty} U_{r,n}(x_{k-1}) = U_{r,n} \left( \lim_{n \to \infty} x_{k-1} \right) = U_{r,n}(x_*),$$

and hence $x_*$ is a fixed-point of $U_{r,n}$. Note moreover that, from (73) we get

$$\|x_* - x_k\|_\Gamma = \lim_{j \to \infty} \|x_j - x_k\|_\Gamma \leq \lim_{j \to \infty} \|x_1 - x_0\|_\Gamma \frac{L_r^j - 1}{L_r^j} \sum_{m=0}^{j-1} L_r^m \leq \|x_1 - x_0\|_\Gamma \frac{L_r^k}{1 - L_r} \|x_1 - x_0\|_\Gamma,$$

which proves (37) of Corollary 2. □
Figure 5: Actual vs. recovered Dirac locations for LS-Cadzow, CPGD and GenFRI, various oversampling parameters $\beta \in \{1, 2, 3, 4\}$, a PSNR in $\{-30, 0, 30\}$ dB. For each case and each source (denoted with the same colours as in Fig. 2), the markers and horizontal lines represent respectively the median and inter-quartile region of the estimated locations’ empirical distribution obtained from 192 noise realisations. The closer a marker is from the line $y = x$ (in dark grey), the better the recovery.