Xampling at the Rate of Innovation

Tomer Michaeli, *Student Member, IEEE*, and Yonina C. Eldar, *Senior Member, IEEE*

Abstract—We address the problem of recovering signals from samples taken at their rate of innovation. Our only assumption is that the sampling system is such that the parameters defining the signal can be stably determined from the samples, a condition that lies at the heart of every sampling theorem. Consequently, our analysis subsumes previously studied nonlinear acquisition devices and nonlinear signal classes. In particular, we do not restrict attention to memoryless nonlinear distortions or to union-of-subspace models. This allows treatment of various finite-rate-of-innovation (FRI) signals that were not previously studied, including, for example, continuous phase modulation transmissions. Our strategy relies on minimizing the error between the measured samples and those corresponding to our signal estimate. This least-squares (LS) objective is generally nonconvex and might possess many local minima. Nevertheless, we prove that under the stability hypothesis, any optimization method designed to trap a stationary point of the LS criterion necessarily converges to the true solution. We demonstrate our approach in the context of recovering pulse streams in settings that were not previously treated. Furthermore, in situations for which other algorithms are applicable, we show that our method is often preferable in terms of noise robustness.

Index Terms—Finite rate of innovation, generalized sampling, iterative recovery, nonlinear distortion, Xampling.

I. INTRODUCTION

SAMPLING theory is concerned with recovery of continuous-time signals from their samples. Being an under determined problem, sampling theorems often rely on the assumption that the signal to be recovered belongs to some predefined class of functions. The "richness" of this class dictates a minimal sampling rate required for perfect reconstruction. For example, the well known Shannon sampling theorem [1] states that any signal $x(t)$ that is π/T-bandlimited can be perfectly recovered from its pointwise uniformly-spaced samples if the sampling interval does not exceed T. Similarly, if $x(t)$ is known to belong to the class of spline functions with knots at $t = nT$, $n \in \mathbb{Z}$, then it can be recovered from pointwise uniform samples with interval T [2].

Until recently, much of the sampling literature treated linear acquisition devices and linear signal priors, that is, families of signals that form subspaces of L_2 [3]. These include shift-invariant spaces [4], of which the bandlimited and spline priors are

special cases, and their generalizations [5]. Reconstruction in SI spaces from nonuniform pointwise samples was treated in [6]. Recovery from linear measurements in arbitrary subspaces was studied from an abstract Hilbert space viewpoint in [7]–[9]. The appeal of subspace models and linear sampling stems from the fact that they result in linear recovery algorithms that are often easy to implement. However, many real-world signal classes do not conform to the subspace model and practical samplers often introduce nonlinear distortions [10].

One notable line of work deviating from these settings treats *nonlinear sampling of linear models*. The first contributions in this direction can be attributed to [11] and [12], which studied reconstruction of bandlimited signals from companding (namely, applying a memoryless nonlinear distortion) and subsequent bandlimiting. These results were later extended to stochastic processes [13] and to more general spaces [14]. In [10], the authors generalized these developments to the setting in which the linear part of the acquisition device does not necessarily match the signal prior. A simpler iterative algorithm, consisting of linear time-invariant (LTI) filtering operations, was recently developed in [15] for the same setting.

Another, rather parallel, deviation from the widely studied linear setting treats *linear sampling of nonlinear models*. Notable in this respect is the study initiated in [16] of sampling finite rate of innovation (FRI) signals. Theses signal classes correspond to families of functions defined by a finite number ρ of parameters per time unit, a quantity referred to as their rate of innovation. Much of the recent attention attracted by this field emerges from the observation that several commonly-encountered FRI signals can be perfectly recovered from samples taken at their rate of innovation. Specifically, in [16], it was demonstrated how periodic and finite-duration streams of Diracs, nonuniform splines and piecewise polynomials can be recovered from uniformly-spaced samples taken at the rate of innovation with either a sinc or a Gaussian kernel. Extensions to certain infinite-duration signals as well as more general classes of sampling kernels appeared in [17], though at the cost of an increase in the sampling rate beyond the rate of innovation. A family of finite-duration sampling kernels was presented in [18] and demonstrated to substantially improve recovery stability. A robust multichannel sampling scheme was recently proposed in [19]. Finally, the authors of [20] studied sampling of a class of semi-periodic functions at the minimal possible rate, using a filter-bank of properly chosen filters.

All the works mentioned above for linear sampling of nonlinear models focused on signals that can be represented as weighted combinations of shifted pulses. These signal classes correspond to unions of subspaces [21]. Another important family within the union-of-subspace category is the set of multiband signals. As shown in [22], when using point-wise samples, the minimal sampling rate required for perfect recovery of these signals is twice their Landau rate, defined as

twice the length of the support in the frequency domain. A low-rate multi-coset sampling method for multi-band signals was proposed in [22]. A more practical multichannel sampling system was later developed [23] and implemented on a board [24]. An important feature of these systems is that the low-rate samples can be used directly to perform digital processing operations, without requiring reconstruction of the analog signal or its high-rate samples as an initial step. This is the key in the recently introduced Xampling paradigm for sampling signals that lie in a union of subspaces [25], [26].

Both lines of work treating nonlinear sampling of linear models and linear sampling of nonlinear models lack the full generality required for deployment in a wide range of practical systems. In particular, common to all nonlinear sampling works is the assumption that the nonlinearity is memoryless, such as in the Wiener–Hammerstein model treated in [10]. However, this is not the case in many real-world applications. An exception is [27], which treats Volterra systems, but only focuses on bandlimited signals and point-wise samples. Similarly, all nonlinear models treated in the literature correspond to unions of subspaces, with the vast majority focusing on pulse streams. These do not include, for example, FRI signals such as continuous-phase modulation (CPM) transmissions. Furthermore, even within the restricted category of pulse streams, solutions are only available for a few special cases of signal structures and sampling devices. These solutions are very unstable in certain situations [28]. An iterative algorithm for reconstructing signals lying in unions of subspaces from linear samples was proposed in [29]. The disadvantage of this technique, though, is that it requires, in each iteration, computing the orthogonal projection of the current signal estimate onto the set of all feasible signals. For most interesting signal models, this necessitates solving a nonconvex optimization problem, which does not admit a closed form solution and for which there is no guarantee that standard optimization techniques will find its solution.

In this paper, we address the problem of reconstructing arbitrary FRI signals from possibly nonlinear measurements obtained at the rate of innovation. The only assumption we make on the sampling mechanism and signal prior is that the parameters defining the signal can be stably recovered from the samples. This assumption must be made by any practical sampling theorem that attempts to recover the signal parameters, whether explicitly or implicitly. Our approach is based on minimization of the error norm between the given set of samples and those of our signal estimate. Our main result is that under the stability assumption, this least-squares (LS) criterion possesses a unique stationary point. Consequently, any optimization algorithm designed to trap a stationary point, will necessarily converge to the true parameters. In particular, we show that the steepest-descent and quasi-Newton methods can be used to recover the signal parameters.

Our approach holds several important advantages. First, it is suited to a family of problems, which supersedes those treated by existing techniques. In particular, we do not assume that the sampling mechanism is linear or that the class of feasible signals forms a union of subspaces. Second, it provides a unified framework for recovering signals from samples taken at their rate of innovation. Thus, rather than tailoring a different algorithm for every possible combination of sampling method and

signal prior, we can apply the same optimization technique to recover the signal parameters. Lastly, our method directly extracts the parameters defining the signal, which are the quantities of interest in most applications, thus conveniently allowing for further digital processing. For example, the parameters can correspond to transmitted symbols in a communication setting, reflector locations in ultrasound imaging [18], and more. These properties all align with the Xampling methodology [26] and even broaden it to beyond the standard linear sampling and union-of-subspace settings.

It is important to note that our approach requires that all feasible signals can be stably recovered from the samples. Thus, even if a specific signal can theoretically be stably recovered, our method is not guaranteed to succeed when there exist other feasible signals which cannot be stably reconstructed. We demonstrate this limitation in the context of a concrete example in Section VI-C.

The paper is organized as follows. In Section II, we describe the problem setting and assumptions. In Section III, we derive a lower bound on the minimal sampling rate required for perfect recovery with a given sampling system. Next, in Section IV, we describe and prove the validity of a general strategy for recovering signals from samples taken at the minimal rate. Two practical iterative methods are then studied in detail in Section V. Finally, we demonstrate our approach in the context of finite-duration and periodic pulse-stream recovery in Section VI and in the context of CPM receivers in Section VII. We show that our method can cope with sampling systems beyond those previously studied. Furthermore, we demonstrate that in time-delay settings for which other algorithms are applicable, our method is often more robust to noise.

II. PROBLEM SETTING

We denote scalars by lowercase letters, vectors by bold lowercase letters and matrices by bold uppercase letters (e.g., $a \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^N$ and $\boldsymbol{A} \in \mathbb{R}^{M \times N}$). The adjoint of a linear operator S is denoted S^* and its null space and range space are written as $\mathcal{N}(S)$ and $\mathcal{R}(S)$ respectively. If h is a function from some Hilbert space $\mathcal{H}_1$ to another Hilbert space $\mathcal{H}_2$, then its Fréchet derivative at x_0 is a continuous linear operator $(\partial h/\partial x)|_{x_0} : \mathcal{H}_1 \to \mathcal{H}_2$ such that

$$\lim_{\boldsymbol{\delta} \to 0} \frac{\left\| h(x_0 + \delta) - h(x_0) - \frac{\partial h}{\partial x}\big|_{x_0} \delta \right\|_{\mathcal{H}_2}}{\|\delta\|_{\mathcal{H}_1}} = 0 \quad (1)$$

where the limit is with respect to the norm defined on $\mathcal{H}_1$.

A. Signal Model

The signal classes we treat are those that are determined by a finite number of parameters per time unit. The τ-local rate of innovation of a signal $x(t)$, denoted ρ_τ, is the minimal number of parameters defining any length-τ segment of $x(t)$, divided by τ. An FRI signal is one for which ρ_τ is finite, at least for large enough τ.

Perhaps the simplest class of FRI signals corresponds to functions that can be expressed as

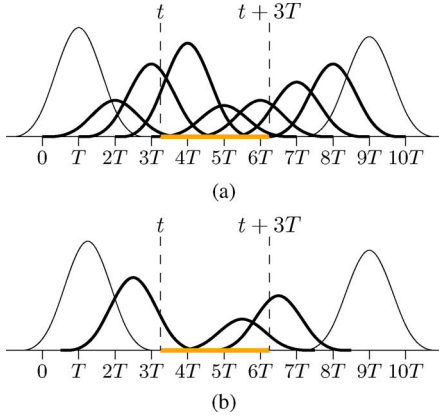$$x(t) = \sum_{m \in \mathbb{Z}} a_m g(t - mT) \quad (2)$$

Fig. 1. Streams of shifted versions of a pulse $g(t)$, supported on $[-2T, 2T]$. Bold pulses are those that affect the observation segment $[t, t + 3T]$. (a) Fixed pulse positions (2), spaced T seconds apart. Here, the segment $[t, t + 3T]$ is affected by 7 pulses so that $\rho_{3T} = 7/(3T)$. (b) Unknown pulse positions (4) with minimal separation T. Here, the rate of innovation is $\rho_{3T} = 2 \times 7/(3T) = 14/(3T)$. Note that the specific segment $[t, t+3T]$ is affected only by 3 pulses so that there are $(2 \times 3)/(3T) = 2/T$ parameters per time unit at that location.

with some arbitrary sequence $\{a_m\} \in \ell_2$, where $g(t)$ is a given pulse in L_2 and $T > 0$ is a given scalar. This set of signals is a linear subspace of L_2, which is often termed a *shift-invariant* (SI) space [4]. The subspace of π/T-bandlimited signals is a special case of (2), with $g(t) = \text{sinc}(t/T)$. Similarly, (2) can represent the space of spline functions (by letting $g(t)$ be a B-spline function) and communication signals such as pulse-amplitude modulation (PAM) and quadrature amplitude modulation (QAM). If the support of $g(t)$ is contained in $[t_a, t_b]$, then any interval of the form $[t, t + \tau]$, where $\tau > 0$, is affected by no more than $\lceil (t_b - t_a + \tau)/T \rceil$ coefficients from the sequence $\{a_m\}$. This is demonstrated in Fig. 1(a). Thus, the τ-local rate of innovation of signals of the form (2) is

$$\rho_\tau = \frac{1}{\tau} \left\lceil \frac{t_b - t_a + \tau}{T} \right\rceil. \tag{3}$$

The asymptotic rate of innovation in this case, which can be found by taking τ to infinity, is $1/T$. We note that, according to our definition, if $g(t)$ is not compactly supported then the rate of innovation is infinite. Thus, for example, bandlimited signals (which correspond to $g(t) = \text{sinc}(t/T)$) are not considered FRI in this paper.

A more complicated model results when the location of the pulses are unknown *a priori*, as often happens in channel sounding scenarios. In these cases,

$$x(t) = \sum_{m \in \mathbb{Z}} a_m g(t - t_m), \tag{4}$$

where both $\{a_m\}$ and $\{t_m\}$ are unknown parameters. This class of signals is not a linear subspace, and is therefore much harder to handle. If we fix the time-delays $\{t_m\}$ and vary only the amplitudes $\{a_m\}$ then we get a subspace. But different choices of time-delays result in different subspaces so that overall (4) corresponds to a *union of subspaces*. Assuming that the minimal separation between any two time delays is T, this model is determined by (at most) twice the number of parameters defining (2) per time unit, as demonstrated in Fig. 1(b). Therefore, the associated τ-local rate of innovation is twice ρ_τ of (3) and the asymptotic rate is $2/T$.

The model (4) and several of its variants have received the largest amount of attention in the FRI literature.[1] However, other interesting FRI signal classes exist. As an example, suppose that L transmissions of the form (2) are modulated, each with a different carrier frequency, to yield

$$x(t) = \sum_{\ell=1}^{L} \sum_{m \in \mathbb{Z}} a_{\ell,m} g(t - mT) \sin(\omega_\ell t). \tag{5}$$

Here, $\{a_{\ell,m}\}_{m \in \mathbb{Z}}$ is the data transmitted by the ℓth user on the carrier frequency ω_ℓ. This model generalizes the family of multiband signals treated in [22] and [23], which corresponds to the case in which $g(t) = \text{sinc}(t/T)$. It is readily seen that if $\text{supp}\{g\} \in [t_a, t_b]$ then any segment $[t, t + \tau]$ of $x(t)$ is affected by at most $L \lceil (t_b - t_a + \tau)/T \rceil$ of the coefficients $\{a_{\ell,m}\}$. With the addition of the L unknown frequencies, we find that the τ-local rate of innovation of signals of the form (5) is

$$\rho_\tau = \frac{L}{\tau} \left(1 + \left\lceil \frac{t_b - t_a + \tau}{T} \right\rceil \right). \tag{6}$$

Note that the asymptotic rate of innovation, which is given by L/T in this setting, is not affected by the fact that we do not know the L carrier frequencies. This is because as we increase the observation period, their effect becomes negligible. The set of signals of the form (6) is a union of subspaces, where the frequencies $\{\omega_\ell\}$ determine the subspace and the amplitudes $\{a_{\ell,m}\}$ determine the position within the subspace.

To the best of our knowledge, only union-of-subspace settings were treated within the FRI literature. However, FRI signals do not have to conform to the union-of-subspace model. An example is continuous-phase modulation (CPM) transmissions. These include continuous phase frequency shift keying (CPFSK) and minimum shift keying (MSK), tamed frequency modulation (TFM), Gaussian MSK (GMSK), and more. Here, the transmitted signal takes on the form

$$x(t) = \cos \left(\omega_0 t + 2\pi h \int_{-\infty}^{t} \sum_{m \in \mathbb{Z}} a_m g(\tau - mT) d\tau \right) \tag{7}$$

where ω_0 is a fixed carrier frequency, $a_m \in \{\pm 1, \pm 3, \ldots, \pm(Q-1)\}$ are the message symbols, h is the modulation index (usually a rational number), and $g(t)$ is a pulse shape that is supported on $[0, LT]$ for some integer $L > 0$ and satisfies $\int_0^{LT} g(t)dt = 0.5$. The rate of innovation of CPM signals can be determined by expressing (7) as

$$x(t) = \cos \left(\omega_0 t + \sum_{m \in \mathbb{Z}} \tilde{a}_m \tilde{g}(t - mT) \right) \tag{8}$$

where

$$\tilde{a}_m = \sum_{n=-\infty}^{m} a_n \tag{9}$$

and

$$\tilde{g}(t) = 2\pi h \int_{-\infty}^{t} (g(\tau) - g(\tau - T)) d\tau. \tag{10}$$

[1] In fact, the original definition of FRI signals, given in [16], was limited only to functions of the form (4).

Since knowing $\{a_m\}$ is equivalent to knowing $\{\tilde{a}_m\}$ (up to initial boundary condition) and $\tilde{g}(t)$ is supported on $[0, (L+1)T]$, the number of coefficients affecting $x(t)$ on any interval $[t, t+\tau]$ is the same as in (2) with $t_a = 0$ and $t_b = (L+1)T$. Consequently, the rate of innovation of CPM signals is

$$\rho_\tau = \frac{1}{\tau}\left(\left\lceil \frac{\tau}{T} \right\rceil + L + 1\right) \tag{11}$$

and their asymptotic rate is $1/T$.

Finally, we note that there are union-of-subspace models that do not correspond to FRI signals. As an example, consider the set of signals

$$x(t) = \sum_{m \in \mathbb{Z}} a_m g_m(t) \tag{12}$$

where the only knowledge we have about the pulses $\{g_m(t)\}$ is that they decay exponentially as $t \to \pm\infty$. Clearly, every possible choice of pulse shapes corresponds to a subspace. Nevertheless, for each m, the number of parameters required for describing $g_m(t)$ is infinite.

Any arbitrary length-τ segment of an FRI signal is determined by at most $K = \lceil \tau \rho_\tau \rceil$ parameters. Therefore, it is reasonable to expect that a properly designed set of K measurements should suffice to identify the parameters of the segment. As discussed in the introduction, this is often the case, implying that many FRI signals can be perfectly recovered from samples taken at their rate of innovation.

Without loss of generality, we focus in this paper on the recovery of an arbitrary segment from an FRI signal. From an abstract viewpoint, any such segment is a vector in some Hilbert space $\mathcal{H}$, which is known to lie within the set

$$\mathcal{X} = \{x = h(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{A}\} \tag{13}$$

where $\mathcal{A}$ is an open set in $\mathbb{R}^K$ and $h : \mathcal{A} \to \mathcal{H}$ is some given function. For example, for any integer $M > 0$, the segment $[T + t_b, MT + t_a]$ from (2) is affected only by the pulses with indices $m = 1, \ldots, M$. Consequently, this segment corresponds to the parameter vector $\boldsymbol{\theta} = (a_1 \quad \cdots \quad a_M)^T$ and to the function $h : \mathbb{R}^M \to L_2([T + t_b, MT + t_a])$ given by

$$h : (a_1 \quad \cdots \quad a_M)^T \mapsto \sum_{m=1}^M a_m g(t - mT). \tag{14}$$

Note that, since the signal prior corresponds to a subspace in this case, the function h is linear. In the channel sounding model (4), however, this is no longer true. Specifically, with a minimal separation of T seconds between any two of the time delays $\{t_m\}$, the segment $[T + t_b, MT + t_a]$ from (4) is affected by no more than M pulses. Indexing these pulses as $m = 1, \ldots, M$, this setting corresponds to the $2M$-dimensional parameter vector $\boldsymbol{\theta} = (t_1 \quad \cdots \quad t_M \quad a_1 \quad \cdots \quad a_M)^T$ and to the nonlinear function $h : \mathbb{R}^{2M} \to L_2([T + t_b, MT + t_a])$ given by

$$h : (t_1 \quad \cdots \quad t_M \quad a_1 \quad \cdots \quad a_M)^T \mapsto \sum_{m=1}^M a_m g(t - t_m). \tag{15}$$

We will assume in the sequel that h is Fréchet differentiable with respect to the parameter vector $\boldsymbol{\theta}$. This demand is not very restrictive and is satisfied in most practical scenarios. In particular, if the pulse shape $g(t)$ is in L_2, then the models (2), (5) and (8) are all Fréchet differentiable with respect to their parameters on any finite-duration interval. If, in addition, $g(t)$ is differentiable and its derivative $g'(t)$ is in L_2, then the model (4) is also Fréchet differentiable. For example, the Fréchet derivative of h of (15) at $\boldsymbol{\theta}_0 = (t_1 \quad \cdots \quad t_M \quad a_1 \quad \cdots \quad a_M)^T$ is the linear operator $(\partial h/\partial \boldsymbol{\theta})|_{\boldsymbol{\theta}_0} : \mathbb{R}^{2M} \to L_2([T + t_b, MT + t_a])$ defined by[2]

$$(\partial h/\partial \boldsymbol{\theta})|_{\boldsymbol{\theta}_0} \boldsymbol{b} = -a_1 g'(t - t_1)b_1 - \cdots - a_M g'(t - t_M)b_M$$
$$+ g(t - t_1)b_{M+1} + \cdots + g(t - t_M)b_{2M}. \tag{16}$$

In addition to the recovery of x, it is often of interest to identify the parameters $\boldsymbol{\theta}$ defining it. This goal, of course, cannot be achieved if the parametrization of the set $\mathcal{X}$ is redundant in the sense that there exist parameters $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ such that $h(\boldsymbol{\theta}_1) = h(\boldsymbol{\theta}_2)$. To be able to recover $\boldsymbol{\theta}$ in a stable manner, we require the slightly stronger condition that

$$\alpha_h \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\mathbb{R}^K} \leq \|h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_2)\|_{\mathcal{H}} \tag{17}$$

for some constant $\alpha_h > 0$ and for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{A}$. As we discuss in Section II-B below, some of the aforementioned signal models do not comply with this requirement unless the feasible set $\mathcal{A}$ is chosen appropriately.

No further assumptions on the structure of $\mathcal{X}$, beyond (17), are needed for our analysis. Nevertheless, a few remarks are in place regarding the implication of this condition in the widely studied union-of-subspace setting.

B. Implication to Union-of-Subspace Models

Suppose that $\boldsymbol{\theta}$ can be partitioned as[3] $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathrm{N}} \quad \boldsymbol{\theta}^{\mathrm{L}})$, where the parameters $\boldsymbol{\theta}^{\mathrm{N}}$ determine a subspace $\mathcal{A}_{\boldsymbol{\theta}^{\mathrm{N}}}$ in $\mathcal{H}$ and the parameters $\boldsymbol{\theta}^{\mathrm{L}}$ determine a vector within $\mathcal{A}_{\boldsymbol{\theta}^{\mathrm{N}}}$. This setting includes as special cases (4), in which $\boldsymbol{\theta}^{\mathrm{N}}$ comprises the time shifts $\{t_\ell\}$ and $\boldsymbol{\theta}^{\mathrm{L}}$ the amplitudes $\{a_\ell\}$, and (5), in which $\boldsymbol{\theta}^{\mathrm{N}}$ comprises the frequencies $\{\omega_\ell\}$ and $\boldsymbol{\theta}^{\mathrm{L}}$ the sequences $\{a_{\ell,m}\}$.

In this situation, condition (17) implies that $\boldsymbol{\theta}^{\mathrm{L}}$ *must be bounded away from zero* for every signal $x \in \mathcal{X}$. Indeed, otherwise we could choose $\boldsymbol{\theta}_1^{\mathrm{L}} = \boldsymbol{\theta}_2^{\mathrm{L}} = \mathbf{0}$ and $\boldsymbol{\theta}_1^{\mathrm{N}} \neq \boldsymbol{\theta}_2^{\mathrm{N}}$ so that $h(\boldsymbol{\theta}_1) = h(\boldsymbol{\theta}_2) = 0$ despite the fact that $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$.

Condition (17) also imposes limitations on the parameters $\boldsymbol{\theta}^{\mathrm{N}}$. Specifically, assume that the parametrization is such that the subspace $\mathcal{A}_{\boldsymbol{\theta}^{\mathrm{N}}}$ is not affected by permutation of the elements of $\boldsymbol{\theta}^{\mathrm{N}}$. This is the case, for instance, in the channel sounding application (4) and in the multiband setting (5) where $\boldsymbol{\theta}^{\mathrm{N}}$ comprises the time delays $\{t_\ell\}$ and frequencies $\{\omega_\ell\}$, respectively. This permutation-invariance implies that if two elements of the vector $\boldsymbol{\theta}^{\mathrm{N}}$ are equal, then there exist multiple choices for the parameters $\boldsymbol{\theta}^{\mathrm{L}}$ yielding the same signal. Therefore, condition (17)

---

[2]Fréchet differentiability is guaranteed in this setting by the fact that the Gateaux (namely directional) derivative of h at $\boldsymbol{\theta}_0$ in the direction $\Delta_{\boldsymbol{\theta}}$ is a bounded linear function of $\Delta_{\boldsymbol{\theta}}$.

[3]The superscripts "N" and "L" stand for nonlinear and linear respectively, intending as a reminder that h is linear in $\boldsymbol{\theta}^{\mathrm{L}}$ and nonlinear in $\boldsymbol{\theta}^{\mathrm{N}}$.

is clearly violated in this case. We thus conclude that in a permutation-invariant parametrization, the elements of $\boldsymbol{\theta}^{\mathrm{N}}$ *must be bounded away from each other*.

Finally, condition (17) imposes restrictions on the maximal possible distance $\|\boldsymbol{\theta}_1^{\mathrm{N}} - \boldsymbol{\theta}_2^{\mathrm{N}}\|$ for any two vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{A}$. More concretely, suppose that the function $h(\boldsymbol{\theta})$ is such that $\|h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_2)\|$ cannot be made arbitrarily large by varying only the subvectors $\boldsymbol{\theta}_1^{\mathrm{N}}$ and $\boldsymbol{\theta}_2^{\mathrm{N}}$ of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. This always happens, for example, in the channel sounding setting (4) with a finitely-supported pulse $g(t)$ because the pulses $g(t - t_1)$ and $g(t - t_2)$ cease to overlap when the distance $|t_2 - t_1|$ exceeds the pulse's width. In this setting, condition (17) cannot be satisfied unless the distance $\|\boldsymbol{\theta}_1^{\mathrm{N}} - \boldsymbol{\theta}_2^{\mathrm{N}}\|$ is bounded. In other words, $\boldsymbol{\theta}^{\mathrm{N}}$ *must be restricted to a bounded set*. Therefore, in model (4), for instance, the time delays $\{t_m\}$ must all lie in some bounded interval. Perhaps a more appealing alternative is to require that t_1 lie in some bounded interval and that there exist an upper bound on the separation between any two consecutive time-delays.

To conclude, in the union-of-subspace setting the feasible set $\mathcal{A}$ must be such that elements of $\boldsymbol{\theta}^{\mathrm{L}}$ are bounded away from zero, the vector $\boldsymbol{\theta}^{\mathrm{N}}$ is restricted to a bounded set in $\mathbb{R}^K$ and its elements are sufficiently separated. This can be achieved in the model (4), for example, by requiring that

$$a_m > a_0, \quad T_{\min} < t_m - t_{m-1} < T_{\max} \tag{18}$$

for every $m = 1, \ldots, M$, where $a_0 > 0$ is a lower-bound on the amplitude, $0 < T_{\min} < T_{\max} < \infty$ constitute a lower- and an upper-bound on the separation between consecutive time-delays and t_0 is an arbitrary constant.

C. Sampling Method

Our goal is to recover x by observing N generalized samples $\boldsymbol{c} = (c_1, \ldots, c_N)^T$ obtained as

$$\boldsymbol{c} = S(x) \tag{19}$$

where $S : \mathcal{H} \to \mathbb{R}^N$ is some (possibly nonlinear) Fréchet differentiable operator. This representation is more general than the widely used linear setting, in which

$$c_n = \langle x, s_n \rangle, \quad n = 1, \ldots, N \tag{20}$$

for some set of vectors $\{s_n\}_{n=1}^N$ in $\mathcal{H}$. In particular, (19) may account for nonlinear distortions introduced by the sampling device. For example, S can represent the samples

$$c_n = f(\langle x, s_n \rangle), \quad n = 1, \ldots, N \tag{21}$$

where $f(\cdot)$ is a nonlinear sensor response.

We say that a sampling operator S is *stable with respect to* $\mathcal{X}$ if there exist constants $0 < \alpha_s \leq \beta_s < \infty$ such that

$$\alpha_s \|x_2 - x_1\|_{\mathcal{H}} \leq \|S(x_1) - S(x_2)\|_{\mathbb{R}^N} \leq \beta_s \|x_2 - x_1\|_{\mathcal{H}} \tag{22}$$

for all $x_1, x_2 \in \mathcal{X}$. This definition is the same as that used in [21] apart from the fact that here the set $\mathcal{X}$ is not necessarily a union of subspaces and the operator S is not necessarily linear. The left-hand inequality ensures that if two signals x_1 and x_2 are sufficiently different from one another, then their samples $S(x_1)$ and $S(x_2)$ are different as well. In particular, it implies that two

different signals $x_1, x_2 \in \mathcal{X}$ cannot produce the same set of samples, so that there is a unique recovery $x \in \mathcal{X}$ associated with every valid set of samples $\boldsymbol{c} = S(x) \in \mathbb{R}^N$.

Conditions (22) and (17) lie at the heart of any practical sampling theorem, whether implicitly or not. It is out of the scope of this paper to survey the situations in which these conditions are satisfied, as this is rather problem-specific. The interested reader may refer to [10] for an analysis of the SI model (2) with nonlinear samples (21), to [29] for linear sampling of several union-of-subspace models and for [30] for a general theory for the stability of FRI models. In the sequel we show that these two conditions dictate a minimal sampling rate below which perfect recovery cannot be guaranteed. More interestingly, we will also show that when (22) and (17) hold, perfect recovery can be attained at this minimal sampling rate by using a wide family of iterative algorithms.

III. MINIMAL SAMPLING RATE

To be able to devise a general reconstruction strategy for signals in $\mathcal{X}$ that were sampled by S, we first determine the minimal number of samples N required for perfect recovery. Interestingly, conditions (22) and (17) implicitly impose a limitation on N.

Proposition 1: Suppose that the function $h : \mathcal{A} \to \mathcal{H}$ satisfies (17) and that the operator $S : \mathcal{H} \to \mathbb{R}^N$ satisfies (22). Then

$$N \geq K + \max_{x_1 \in \mathcal{X}} \dim\left(\mathcal{N}\left(\left(\left.\frac{\partial S}{\partial x}\right|_{x_1}\right)^*\right)\right). \tag{23}$$

Before providing a proof, we note that Proposition 1 shows that the minimal number of samples N required for perfect recovery is the number of parameters K defining x. In other words, stable recovery is impossible when sampling below the rate of innovation. While very intuitive and stated in every FRI sampling paper, we believe that this result was not formally proved before for the general signal model and acquisition mechanism discussed in this paper.

Proposition 1 further shows that sampling at the rate of innovation is insufficient if the null space of $(\partial S/\partial x)^*$ is nonempty at some $x \in \mathcal{X}$. When S is a linear operator and $\mathcal{X}$ is a subspace, spanned by vectors $\{x_k\}_{k=1}^K$, this condition implies that the vectors $\{Sx_k\}_{k=1}^K$ should be linearly independent. In other words, the $N \times K$ matrix whose (n, k) entry is $\langle s_n, x_k \rangle$, should have an empty nullspace, where $\{s_n\}_{n=1}^N$ are the sampling vectors of (20). If S is linear but $\mathcal{X}$ is not contained in any finite-dimensional subspace, then sampling at the rate of innovation necessitates that the sampling vectors $\{s_n\}_{n=1}^N$ be linearly independent. Indeed, if $\{s_n\}_{n=1}^N$ are linearly dependent, then there exists an index j such that $s_j = \sum_{n \neq j} a_n s_n$ for some coefficients $\{a_n\}_{n \neq j}$. Consequently, the sample c_j can be expressed in terms of the other samples as $c_j = \langle x, s_j \rangle = \sum_{n \neq j} \bar{a}_n \langle x, s_n \rangle = \sum_{n \neq j} \bar{a}_n c_n$ and thus can be disregarded.

As another example, suppose that one of the measurements produced by the sensing device, say c_1, is the energy $0.5\|x\|^2$ of x. In this case $(\partial c_1/\partial x)|_{x_1} = x_1$. Consequently, from Proposition 1, sampling at the minimal rate is impossible if the set of signals $\mathcal{X}$ contains the signal $x_1 = 0$. The intuition here follows from the observation that small perturbations in x around the signal $x_1 = 0$ do not show in c_1. Therefore, if the input to

our sampling device happens to be $x = 0$ in this setting, then sampling is unavoidably unstable, as the left-hand side of condition (22) is violated.

Proof: Since $h(\boldsymbol{\theta})$ and $S(x)$ are Fréchet differentiable, it follows that the function $\hat{\boldsymbol{c}}(\boldsymbol{\theta}) = S(h(\boldsymbol{\theta}))$ is Fréchet differentiable as well. We will start by showing that its derivative $\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}$, which is an $N \times K$ matrix, has an empty null space.

By definition, the Fréchet derivative $\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}$ at $\boldsymbol{\theta}_1$ satisfies

$$\lim_{\boldsymbol{\delta} \to 0} \frac{\left\| \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1 + \boldsymbol{\delta}) - \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1) - \frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \boldsymbol{\delta} \right\|_{\mathbb{R}^N}}{\|\boldsymbol{\delta}\|_{\mathbb{R}^K}} = 0. \quad (24)$$

In particular, for any nonzero $\boldsymbol{a} \in \mathbb{R}^K$,

$$\lim_{t \to 0} \frac{\left\| \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1 + t\boldsymbol{a}) - \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1) - t \frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \boldsymbol{a} \right\|_{\mathbb{R}^N}}{\|t\boldsymbol{a}\|_{\mathbb{R}^K}} = 0 \quad (25)$$

where t is a scalar variable. Now, assume that $\boldsymbol{a} \in \mathcal{N}(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1})$. Then (25) implies that

$$\lim_{t \to 0} \frac{\|\hat{\boldsymbol{c}}(\boldsymbol{\theta}_1 + t\boldsymbol{a}) - \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1)\|_{\mathbb{R}^N}}{\|t\boldsymbol{a}\|_{\mathbb{R}^K}} = 0. \quad (26)$$

However, (17) and (22) imply that

$$\begin{aligned} &\frac{\|\hat{\boldsymbol{c}}(\boldsymbol{\theta}_1 + t\boldsymbol{a}) - \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1)\|_{\mathbb{R}^N}}{\|t\boldsymbol{a}\|_{\mathbb{R}^K}} \\ &= \frac{\|S(h(\boldsymbol{\theta}_1 + t\boldsymbol{a})) - S(h(\boldsymbol{\theta}_1))\|_{\mathbb{R}^N}}{\|t\boldsymbol{a}\|_{\mathbb{R}^K}} \\ &\geq \alpha_s \frac{\|h(\boldsymbol{\theta}_1 + t\boldsymbol{a}) - h(\boldsymbol{\theta}_1)\|_{\mathbb{R}^N}}{\|t\boldsymbol{a}\|_{\mathbb{R}^K}} \\ &\geq \alpha_s \alpha_h > 0 \end{aligned} \quad (27)$$

for every $t \neq 0$. This contradicts (26) and therefore demonstrates that $\mathcal{N}(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1}) = \{\mathbf{0}\}$, which implies that $\dim(\mathcal{R}(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1})) = K$.

Next, note that $\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1} = (\partial S/\partial x|_{h(\boldsymbol{\theta}_1)})(\partial h/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1})$ so that $\mathcal{R}(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1}) \subseteq \mathcal{R}(\partial S/\partial x|_{h(\boldsymbol{\theta}_1)}) = \mathcal{N}((\partial S/\partial x|_{h(\boldsymbol{\theta}_1)})^*)^{\perp}$. Therefore,

$$\begin{aligned} K &= \dim\left(\mathcal{R}\left(\frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \right) \right) \\ &\leq \dim\left(\mathcal{N}\left(\left(\frac{\partial S}{\partial x}\Big|_{h(\boldsymbol{\theta}_1)} \right)^* \right)^{\perp} \right) \\ &= N - \dim\left(\mathcal{N}\left(\left(\frac{\partial S}{\partial x}\Big|_{h(\boldsymbol{\theta}_1)} \right)^* \right) \right). \end{aligned} \quad (28)$$

Since (28) holds for every $\boldsymbol{\theta}_1 \in \mathcal{A}$, it holds for the $\boldsymbol{\theta}_1$ minimizing the right-hand side, completing the proof. ∎

Throughout the rest of the paper we focus on the case in which $N = K$ samples of $x(t)$ are obtained with an operator S satisfying

$$\mathcal{N}\left(\left(\frac{\partial S}{\partial x}\Big|_{x_1} \right)^* \right) = \{\mathbf{0}\}, \quad \forall x_1 \in \mathcal{X}. \quad (29)$$

This corresponds to sampling at the rate of innovation.

IV. LEAST SQUARES RECOVERY

Suppose we want to recover a signal $x = h(\boldsymbol{\theta}_0) \in \mathcal{H}$ from its samples $\boldsymbol{c} = S(x)$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^K$ is an unknown parameter vector and $S : \mathcal{H} \to \mathbb{R}^K$ is a given sampling operator. To address this problem, it is natural to seek the minimizer of the function

$$\varepsilon(\boldsymbol{\theta}) = \frac{1}{2}\|S(h(\boldsymbol{\theta})) - \boldsymbol{c}\|_{\mathbb{R}^K}^2 = \frac{1}{2}\|\hat{\boldsymbol{c}}(\boldsymbol{\theta}) - \boldsymbol{c}\|_{\mathbb{R}^K}^2, \quad (30)$$

where we defined $\hat{\boldsymbol{c}}(\boldsymbol{\theta}) = S(h(\boldsymbol{\theta}))$. The reasoning behind this choice follows from the following observation

Proposition 2: Suppose that the function $h : \mathbb{R}^K \to \mathcal{H}$ satisfies (17) and that the operator $S : \mathcal{H} \to \mathbb{R}^K$ satisfies (22). Then $\boldsymbol{\theta}_0$ is the unique global minimizer of $\varepsilon(\boldsymbol{\theta})$.

Proof: Clearly, $\varepsilon(\boldsymbol{\theta}) \geq 0$ for every $\boldsymbol{\theta} \in \mathbb{R}^K$ and $\varepsilon(\boldsymbol{\theta}_0) = 0$, so that $\boldsymbol{\theta}_0$ is a global minimizer of $\varepsilon(\boldsymbol{\theta})$. This minimizer is unique since, due to (17) and (22), $\varepsilon(\boldsymbol{\theta}) \geq \alpha_s \alpha_h \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\mathbb{R}^K}$ so that $\varepsilon(\boldsymbol{\theta}) > 0$ for every $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. ∎

The LS criterion (30) is also plausible when the samples $\boldsymbol{c}$ correspond to a perturbation of the true sample vector by white Gaussian noise. In this case, the minimizer of (30) is a maximum-likelihood estimate of $\boldsymbol{\theta}$ from $\boldsymbol{c}$.

Unfortunately, the function $\varepsilon(\boldsymbol{\theta})$ is generally nonconvex and might possess many local minima. It therefore seems that standard optimization techniques may fail in finding its global minimizer $\boldsymbol{\theta}_0$. However, as we show next, when sampling at the rate of innovation, assumptions (17) and (22) guarantee that $\boldsymbol{\theta}_0$ is the unique stationary point of $\varepsilon(\boldsymbol{\theta})$. Thus, any algorithm designed to trap a stationary point, necessarily converges to the true parameter vector $\boldsymbol{\theta}_0$. The proof of this result follows that of [10, Theorem 6], which treats the special case of SI signals and memoryless nonlinear samples.

Theorem 1: Suppose that the function $h : \mathbb{R}^K \to \mathcal{H}$ satisfies (17), the operator $S : \mathcal{H} \to \mathbb{R}^K$ satisfies (22) and its Fréchet derivative $\partial S/\partial x$ satisfies (29). Then $\nabla \varepsilon(\boldsymbol{\theta}_1) = \mathbf{0}$ only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$.

Proof: The gradient $\nabla \varepsilon(\boldsymbol{\theta}_1)$ is given by

$$\nabla \varepsilon(\boldsymbol{\theta}_1) = \left(\frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \right)^* (\hat{\boldsymbol{c}}(\boldsymbol{\theta}_1) - \boldsymbol{c}). \quad (31)$$

We showed in the proof of Proposition 1 that $\mathcal{R}(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1}) = \mathbb{R}^K$. Since here $\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_1}$ is a $K \times K$ matrix, it follows that

$$\mathcal{N}\left(\left(\frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \right)^* \right) = \mathcal{R}\left(\frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_1} \right)^{\perp} = \{\mathbf{0}\} \quad (32)$$

so that $\nabla \varepsilon(\boldsymbol{\theta}_1) = 0$ only if $\hat{\boldsymbol{c}}(\boldsymbol{\theta}_1) - \boldsymbol{c} = \mathbf{0}$. This, by Proposition 2, happens only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$, completing the proof. ∎

The importance of Theorem 1 lies in the fact that it provides a unified mechanism for recovering FRI signals from samples taken at the rate of innovation. Namely, rather than developing a different algorithm for every choice of signal family and sampling method, we can employ the same general-purpose optimization technique to find the stationary point of (30). Furthermore, this strategy is also advantageous over the iterative approach of [29], as it avoids the need for projecting the signal estimate onto $\mathcal{X}$ in each iteration, an operation that possesses no closed form solution for most FRI signal classes.
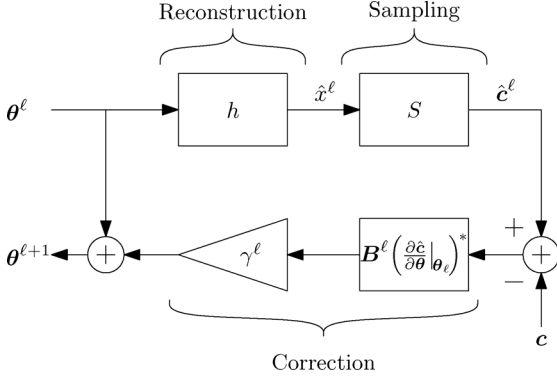
Fig. 2. Schematic interpretation of one iteration of (35).

V. ITERATIVE RECOVERY

There are numerous optimization algorithms that can be used to find the stationary point of the objective function $\varepsilon(\boldsymbol{\theta})$ over $\mathcal{A}$. For simplicity, we focus here on unconstrained optimization methods, namely those that can be applied when $\mathcal{A} = \mathbb{R}^K$. This does not limit the generality of the discussion since if $\mathcal{A} \neq \mathbb{R}^K$, then the constrained problem $\min_{\boldsymbol{\theta} \in \mathcal{A}} \varepsilon(\boldsymbol{\theta})$ can be transformed into the unconstrained problem $\min_{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^K} \varepsilon(p(\tilde{\boldsymbol{\theta}}))$, where $p : \mathbb{R}^K \to \mathcal{A}$ is one-to-one and onto. The latter problem possesses a unique stationary point $\tilde{\boldsymbol{\theta}}_0 = p^{-1}(\boldsymbol{\theta}_0)$. Therefore, once $\tilde{\boldsymbol{\theta}}_0$ is determined, the desired solution can be computed as $\boldsymbol{\theta}_0 = p(\tilde{\boldsymbol{\theta}}_0)$. For example, the model (4) with the set $\mathcal{A}$ of constraints defined by (18) can be handled by defining

$$\tilde{\theta}_m^{\mathrm{L}} = \ln(a_m - a_0), \quad \tilde{\theta}_m^{\mathrm{N}} = \tan\left(\pi \frac{t_m - t_{m-1} - \bar{T}}{\Delta}\right) \quad (33)$$

where $\bar{T} = (T_{\max} + T_{\min})/2$ and $\Delta = T_{\max} - T_{\min}$, so that

$$a_m = e^{\tilde{\theta}_m^{\mathrm{L}}} + a_0, \quad t_m = t_0 + m\bar{T} + \frac{\Delta}{\pi}\sum_{i=1}^{m} \arctan\left(\tilde{\theta}_i^{\mathrm{N}}\right). \quad (34)$$

With this choice, the set $\mathcal{X}$ of all feasible signals is obtained by varying $\tilde{\boldsymbol{\theta}}^{\mathrm{L}}$ and $\tilde{\boldsymbol{\theta}}^{\mathrm{N}}$ over the entire space $\mathbb{R}^M$ and not over some subset of $\mathbb{R}^M$.

Most unconstrained optimization methods start with an initial guess $\boldsymbol{\theta}^0$ and perform iterations of the form

$$\boldsymbol{\theta}^{\ell+1} = \boldsymbol{\theta}^{\ell} - \gamma^{\ell} \boldsymbol{B}^{\ell} \nabla \varepsilon(\boldsymbol{\theta}^{\ell}) \quad (35)$$

where γ^{ℓ} is a scalar step size obtained by means of a one dimensional search and $\boldsymbol{B}^{\ell}$ is a positive definite matrix. Due to the structure of $\nabla \varepsilon(\boldsymbol{\theta}^{\ell})$ in our case (see (31)), the iterations (35) can be given a simple interpretation, as shown in Fig. 2. Specifically, at the ℓth iteration, the current estimate $\boldsymbol{\theta}^{\ell}$ of the parameters $\boldsymbol{\theta}$ is used to construct our estimate $\hat{x}^{\ell}$ of the signal x by applying the function h. This estimate is then sampled using the operator S to obtain an estimated sample vector $\hat{\boldsymbol{c}}^{\ell}$. Finally, the difference between $\hat{\boldsymbol{c}}^{\ell}$ and the true set of samples $\boldsymbol{c}$ is multiplied by a correction matrix and added to $\boldsymbol{\theta}^{\ell}$ to yield the updated estimate $\boldsymbol{\theta}^{\ell+1}$ of the parameter vector $\boldsymbol{\theta}$.

In our setting, the objective function $\varepsilon(\boldsymbol{\theta})$ is bounded from below. The iterations (35) are therefore guaranteed to converge to a stationary point of $\varepsilon(\boldsymbol{\theta})$ if γ^{ℓ} is chosen to satisfy the Wolfe conditions [31], $\boldsymbol{B}^{\ell}$ is chosen such that

$$\frac{\left\langle \boldsymbol{B}^{\ell} \nabla \varepsilon(\boldsymbol{\theta}^{\ell}), \nabla \varepsilon(\boldsymbol{\theta}^{\ell}) \right\rangle_{\mathbb{R}^K}}{\left\| \boldsymbol{B}^{\ell} \nabla \varepsilon(\boldsymbol{\theta}^{\ell}) \right\|_{\mathbb{R}^K} \left\| \nabla \varepsilon(\boldsymbol{\theta}^{\ell}) \right\|_{\mathbb{R}^K}} > \delta \quad (36)$$

for some constant $\delta > 0$ independent of ℓ, and the gradient $\nabla \varepsilon(\boldsymbol{\theta})$ is Lipschitz continuous in an environment of the level-set $\mathcal{N} = \{\boldsymbol{\theta} : \varepsilon(\boldsymbol{\theta}) \leq \varepsilon(\boldsymbol{\theta}^0)\}$ [31].

A step size satisfying the Wolfe conditions can be found by using the backtracking method [31], as presented in Algorithm 1. Condition (36) is trivially satisfied with $\boldsymbol{B}^{\ell} = \boldsymbol{I}$, which corresponds to the steepest descent method. As we show in Appendix A, this condition is also satisfied with $\boldsymbol{B}^{\ell} = ((\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}^{\ell}})^* (\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}^{\ell}}))^{-1}$ if

$$\|h(\boldsymbol{\theta}_1) - h(\boldsymbol{\theta}_2)\|_{\mathcal{H}} \leq \beta_h \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\mathbb{R}^K} \quad (37)$$

for some $\beta_h < \infty$ and for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{N}$. This choice corresponds to the Gauss-Newton method, which typically converges much faster than steepest descent. Finally, we show in Appendix B that a sufficient condition for $\nabla \varepsilon(\boldsymbol{\theta})$ to be Lipschitz continuous over $\mathcal{N}$ is that the derivative of h be Lipschitz continuous there, namely that

$$\left\| \frac{\partial h}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_1} - \frac{\partial h}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}_2} \right\| \leq \beta_{h'} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\mathbb{R}^K} \quad (38)$$

for some $\beta_{h'} < \infty$ and for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{N}$. The analyses in Appendices A and B follow closely those in the proof of [10, Theorem 7]. To summarize, we have the following result.

Theorem 2: Suppose that the function $h : \mathbb{R}^K \to \mathcal{H}$ satisfies (17), its Fréchet derivative $\partial h/\partial \boldsymbol{\theta}$ satisfies (38), the operator $S : \mathcal{H} \to \mathbb{R}^K$ satisfies (22) and its Fréchet derivative $\partial S/\partial x$ satisfies (29). Consider the iterations (35), where the step size γ^{ℓ} is obtained via Algorithm 1. Then, each of the following options guarantees that $\boldsymbol{\theta}^{\ell} \to \boldsymbol{\theta}_0$:

- $\boldsymbol{B}^{\ell} = \boldsymbol{I}$;
- $\boldsymbol{B}^{\ell} = ((\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}^{\ell}})^* (\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}^{\ell}}))^{-1}$ and condition (37) holds.

Algorithm 1: Backtracking line search

set $\boldsymbol{g}^{\ell} = \nabla \varepsilon(\boldsymbol{\theta}^{\ell})$, $\boldsymbol{d}^{\ell} = -\boldsymbol{B}^{\ell}\boldsymbol{g}^{\ell}$, $\delta = 1$ and $\rho, \eta \in (0,1)$
while $\varepsilon(\boldsymbol{\theta}^{\ell} + \delta \boldsymbol{d}^{\ell}) > \varepsilon(\boldsymbol{\theta}^{\ell}) + \eta \delta \langle \boldsymbol{d}^{\ell}, \boldsymbol{g}^{\ell} \rangle_{\mathbb{R}^K}$ **do**
$\qquad \delta \leftarrow \rho\delta$
end while
return $\gamma^{\ell} = \delta$

VI. APPLICATION TO CHANNEL SOUNDING

We now demonstrate our approach in the channel sounding setting (4). Specifically, suppose that

$$x(t) = \sum_{m=1}^{M} a_m g(t - t_m), \quad t \in [0, \tau] \quad (39)$$

where $g(t)$ is a known pulse shape, $\{a_m\}_{m=1}^M$ are unknown amplitudes, and $\{t_m\}_{m=1}^M$ are unknown time-delays. As explained in Section II-B, the parameter vector

$$\boldsymbol{\theta} = (\, t_1 \quad \cdots \quad t_M \quad a_1 \quad \cdots \quad a_M \,)^T \qquad (40)$$

cannot be stably recovered unless the amplitudes all surpass a certain threshold and the pulses are well separated yet confined to a bounded interval. We therefore adopt the assumptions (18) and transform the optimization problem into an unconstrained one by using the parameter vector $\tilde{\boldsymbol{\theta}} = p^{-1}(\boldsymbol{\theta})$ described in (33), with the transformation $\boldsymbol{\theta} = p(\tilde{\boldsymbol{\theta}})$ of (34). Our goal is to recover the signal parameters from the samples (21), where $\{s_n(t)\}_{n=1}^N$ are sampling kernels in $L_2([0,\tau])$ and $f(\,\cdot\,)$ is a nonlinear response function.

As discussed in the introduction, when $f(\,\cdot\,)$ is the identity operator, there are several combinations of pulse shapes $g(t)$ and sampling kernels $\{s_n(t)\}$ that can be treated via existing algorithms in a stable manner, such as [18] and [19]. However, none of the existing techniques is applicable when $f(\,\cdot\,)$ is nonlinear. Furthermore, as we demonstrate in Section VI-A below, our approach allows recovery from SI samples with a kernel that is not supported by [18]. Moreover, in Section VI-B we apply our technique in a multichannel setting for which the algorithm of [19] is applicable, and show the advantage of our approach in the presence of noise.

To apply the Gauss-Newton or steepest decent methods, we note that, with the transformation $\boldsymbol{\theta} = p(\tilde{\boldsymbol{\theta}})$ of (34),

$$\frac{\partial \hat{\boldsymbol{c}}}{\partial \tilde{\boldsymbol{\theta}}} = \frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}} \frac{\partial p}{\partial \tilde{\boldsymbol{\theta}}}. \qquad (41)$$

Explicit computation shows that

$$\frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}} = \boldsymbol{C} \, (\, \boldsymbol{A} \quad \boldsymbol{B} \,) \qquad (42)$$

with

$$\boldsymbol{A} = \begin{pmatrix} -a_1 \langle g'(\cdot - t_1), s_1 \rangle & \cdots & -a_M \langle g'(\cdot - t_M), s_1 \rangle \\ \vdots & & \vdots \\ -a_1 \langle g'(\cdot - t_1), s_N \rangle & \cdots & -a_M \langle g'(\cdot - t_M), s_N \rangle \end{pmatrix} \qquad (43)$$

$$\boldsymbol{B} = \begin{pmatrix} \langle g(\cdot - t_1), s_1 \rangle & \cdots & \langle g(\cdot - t_M), s_1 \rangle \\ \vdots & & \vdots \\ \langle g(\cdot - t_1), s_N \rangle & \cdots & \langle g(\cdot - t_M), s_N \rangle \end{pmatrix} \qquad (44)$$

and

$$\boldsymbol{C} = \operatorname{diag} (\, f'(\langle x, s_1 \rangle) \quad \cdots \quad f'(\langle x, s_N \rangle) \,). \qquad (45)$$

Furthermore,

$$\frac{\partial p}{\partial \tilde{\boldsymbol{\theta}}} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{E} \end{pmatrix} \qquad (46)$$

with

$$\boldsymbol{D} = \operatorname{diag} (\, e^{\tilde{\theta}_1} \quad \cdots \quad e^{\tilde{\theta}_M} \,) \qquad (47)$$
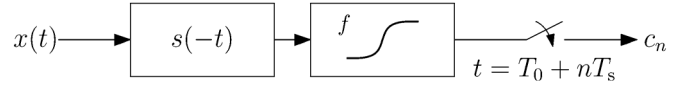

Fig. 3. Nonlinear and nonideal sampling.

and

$$\boldsymbol{E} = \frac{\Delta}{\pi} \begin{pmatrix} \frac{1}{1+\tilde{\theta}_{M+1}^2} & 0 & \cdots & 0 \\ \frac{1}{1+\tilde{\theta}_{M+1}^2} & \frac{1}{1+\tilde{\theta}_{M+2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1+\tilde{\theta}_{M+1}^2} & \frac{1}{1+\tilde{\theta}_{M+2}^2} & \cdots & \frac{1}{1+\tilde{\theta}_{2M}^2} \end{pmatrix}. \qquad (48)$$

We now demonstrate our method in several specific settings.

A. Gaussian Pulses and Gaussian Kernels With Nonlinear Distortion

Consider the sampling system of Fig. 3, in which $x(t)$ is sampled after being convolved with a filter $s(-t)$ and passing through an amplitude limiter $f(\,\cdot\,)$. The resulting samples can be described by (21), with $s_n(t) = s(t - T_0 - nT_s)$. Since the model (39) is clearly determined by $K = 2M$ parameters, we would like to recover any such $x(t)$ from $N = 2M$ samples. We choose the sampling period T_s to equal τ/N and the offset T_0 to be $T_s/2$, so that the sampling functions span the entire observation segment $[0, \tau]$.

Fig. 4 demonstrates the convergence of the Newton iterations for recovering $M = 2$ pulses over the period $[0, 1]$ from $N = 4$ samples. Here, the pulse shape and the sampling filter were taken to be Gaussian functions with variances $\sigma_g = 0.05$ and $\sigma_s = 0.1$, respectively. Note that, with this choice, all inner products in (43) and (44) can be computed analytically at every iteration. The nonlinear response curve was set to be $f(c) = 100 \arctan(0.01c)$. The constraints (18) we assumed on the parameters corresponded to $a_0 = 0.1$, $T_{\min} = 0.3$, $T_{\max} = 0.7$ and $t_0 = -0.3$.

The true parameters in this experiment were $t_1 = 0.2$, $t_2 = 0.8$, $a_1 = 1$ and $a_2 = 5$. As shown in Fig. 4(a), the iterations were initialized at $t_1 = 1/3$, $t_2 = 2/3$ and $a_1 = a_2 = 3$. The estimated samples at this point, shown in "x"-marks, deviate substantially from the true samples, marked with circles. As can be seen, though, this gap decreases quickly in the first 15 iterations [see Fig. 4(b)] and almost completely vanishes after 30 iterations [Fig. 4(c)]. Fig. 4(d) shows the rapid decrease in the LS objective (30) as a function of the iterations.

Fig. 5 demonstrates the behavior of the algorithm in the presence of noise. The setting here is the same as that of Fig. 4 with the distinction that white Gaussian noise is added to the samples prior to recovery. This figure depicts the mean-square error (MSE) in $x(t)$, defined as

$$\text{MSE} = \mathbb{E}\left[\int_0^\tau |x(t) - \hat{x}(t)|^2 \, dt \right] \qquad (49)$$

as a function of the signal-to-noise (SNR) ratio. The solid line corresponds to the Cramér–Rao bound (CRB), developed in [28], which is a lower bound on the MSE attainable by any unbiased estimation technique. As can be seen, the MSE of our
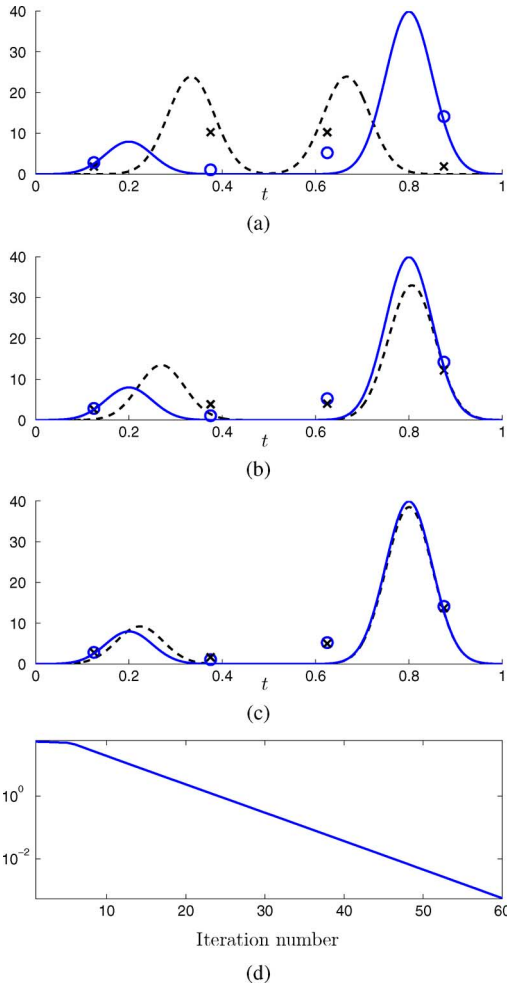
Fig. 4. Convergence of Newton iterations for pulse stream recovery. (a) Initialization. (b) Fifteenth iteration. (c) Thirtieth iteration. (d) LS objective value as a function of the iterations.
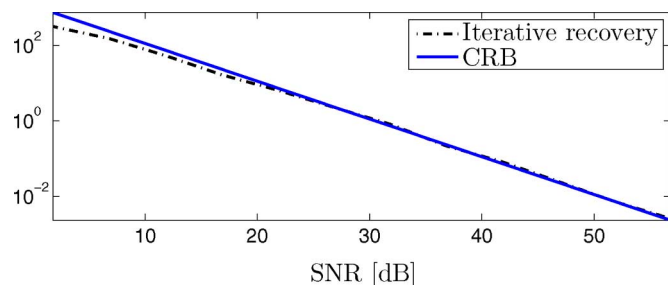

Fig. 5. MSE as a function of SNR for pulse stream recovery in the setting of Fig. 4.

method coincides with the CRB in high SNR scenarios and outperforms it at low SNR levels. This is a result of the fact that our technique is biased.

B. Periodic Pulses and Sinusoidal Kernels

Next, we turn to demonstrate our approach in a periodic pulse-stream scenario with the multichannel sampling system of [19]. Specifically, suppose that $g(t)$ in (39) is τ-periodic with Fourier coefficients $\tilde{g}_k = (1/\tau)\langle g, \phi_k \rangle$, where
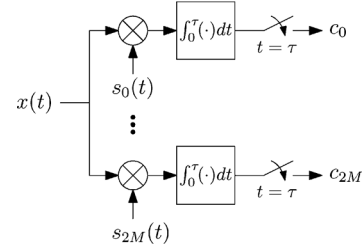
Fig. 6. Linear multichannel sampling.

$\phi_k(t) = e^{2\pi jkt/\tau}$. In [19], it was shown that the pulse parameters can be identified in this setting by using the multichannel sampling system depicted in Fig. 6, where the sampling kernels $s_n(t)$ correspond to combinations of the complex exponentials $\{\phi_k(t)\}_{k \in \mathcal{K}}$ with $\mathcal{K}$ being a set of consecutive indices. The algorithm of [19] was developed for linear sampling, so that $f(\cdot)$ of (21) is set to be the identity. This algorithm is based on applying techniques for identifying the frequencies of complex exponentials, such as the matrix pencil [32] or annihilating filter [16] methods.

If we restrict attention to real sampling functions, then the minimal number N of samples supported by the method of [19] is $2M + 1$. This is achieved by choosing[4]

$$s_n(t) = \begin{cases} 1 & n = 0, \\ \cos(2\pi nt/\tau) & 1 \leq n \leq M, \\ \sin(2\pi nt/\tau) & M+1 \leq n \leq 2M. \end{cases} \quad (50)$$

Due to the very small over-sampling factor, only the annihilating filter method is applicable within the approach of [19].

Our approach can operate with a budget of only $2M$ samples and with arbitrary sampling kernels, as long as (22) is satisfied. Nevertheless, we now wish to demonstrate that our method is advantageous over that of [19] even in settings in which the sampling kernels are chosen as (50).

We note that the convergence guarantees we provided in previous sections do not hold when sampling above the rate of innovation. However, in practice, the algorithm performs well also in mild over-sampling scenarios, such as the one treated here.

To compare between iterative recovery and the algorithm of [19], we concentrated on signals with period $\tau = 1$ comprising $M = 2$ pulses and thus used $N = 2M + 1 = 5$ samples to recover them. We chose a pulse with Fourier coefficients $\tilde{g}_k = 1/(5 + n^2)$, which, as shown in Fig. 7(a) is very wide in the time domain. This renders the determination of pulse positions a challenging task. The constraints (18) were the same as in Section VI-A. The true time delays were $t_1 = 1/\sqrt{15} \approx 0.2582$ and $t_2 = 1/\sqrt{2} \approx 0.7071$ and the true amplitudes were randomly generated to yield $a_1 \approx 0.5285$ and $a_2 \approx 0.14$. The initialization of the algorithm was the same as in Section VI-A. At each iteration of the algorithm, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ comprise the (weighted) Fourier coefficients of shifted versions of $g(t)$ and of $g'(t)$. These quantities can be obtained analytically from the Fourier coefficients of $g(t)$.

In Fig. 7(b), the performance of both approaches is compared against the CRB when the samples are contaminated by white Gaussian noise. As can be seen, the Gauss-Newton method outperforms the annihilating-filter-based algorithm at all SNR.

---

[4]For notational convenience the samples are indexed as $c_0, c_1, \ldots$ in this example rather than $c_1, c_2, \ldots$.
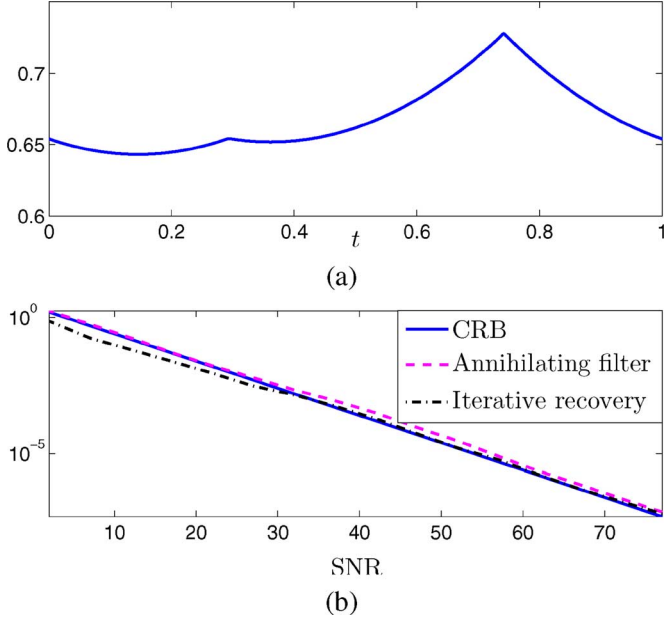
(a)

(b)

Fig. 7. (a) One period of the periodic signal $x(t)$ comprising two wide pulses. (b) MSE as a function of SNR for recovery with the $N = 5$ samples corresponding to (50). The dashed and dashed–dotted lines correspond, respectively, to the method of [19] and Gauss-Newton iterations. The solid line corresponds to the CRB for estimating $x(t)$ from the samples.

C. Stability

Although time-delay estimation is a long-studied problem, stability was not given much attention in past works. In [29] an example was presented in which the delay t_1 of a rectangular pulse $g(t - t_1)$ can be determined from uniformly-spaced samples taken at the output of a triangular impulse-response filter, but this cannot be achieved in a stable manner. For a general channel-sounding setting, it is not trivial to obtain simple-to-verify conditions on the pulse shape $g(t)$, sampling functions $\{s_n(t)\}$, nonlinearity $f(\cdot)$ and the parameters $T_{\min}$, $T_{\max}$, t_0, and a_0 such that stable recovery is guaranteed. However, as we now demonstrate, unstable settings can be identified numerically using the proposed approach.

Assume that $N = 2M$ samples are obtained with a monotonic nonlinearity $f(\cdot)$ and a set $\{s_n\}_{n=1}^{2M}$ of linearly independent sampling kernels. In this case, condition (29) is satisfied. Assume further that for a certain parameter vector $\boldsymbol{\theta}_0 = (\ t_1 \ \cdots \ t_M \ a_1 \ \cdots \ a_M\ )$ and certain initial guess $\boldsymbol{\theta}^0$, the algorithm terminates at a point $\boldsymbol{\theta}_1$ for which $\varepsilon(\boldsymbol{\theta}_1) \neq 0$. This means that $\boldsymbol{\theta}_0$ is not the unique stationary point of the LS objective so that, according to Theorem 1, stable recovery is not possible in this setting for all $\boldsymbol{\theta}$ in the constraint set $\mathcal{A}$. More specifically, either condition (17) or (22) (or both) are violated for some $\boldsymbol{\theta} \in \mathcal{A}$.

In fact, the point at which (17) or (22) are violated is no other than $\boldsymbol{\theta}_1$. Indeed, the fact that $\nabla \varepsilon(\boldsymbol{\theta}_1) = 0$ and $\varepsilon(\boldsymbol{\theta}_1) \neq 0$ implies that $(\partial \hat{c}/\partial \boldsymbol{\theta})|_{\boldsymbol{\theta}_1} = 0$ (see (31) and (30)). Therefore, by the definition of the Fréchet derivative,

$$0 = \lim_{\boldsymbol{\delta} \to 0} \frac{\|\hat{\boldsymbol{c}}(\boldsymbol{\theta}_1 + \boldsymbol{\delta}) - \hat{\boldsymbol{c}}(\boldsymbol{\theta}_1)\|_{\mathbb{R}^{2M}}}{\|\boldsymbol{\delta}\|_{\mathbb{R}^{2M}}}$$

$$= \lim_{\boldsymbol{\delta} \to 0} \frac{\|S(h(\boldsymbol{\theta}_1 + \boldsymbol{\delta})) - S(h(\boldsymbol{\theta}_1))\|_{\mathbb{R}^{2M}}}{\|\boldsymbol{\delta}\|_{\mathbb{R}^{2M}}} \quad (51)$$
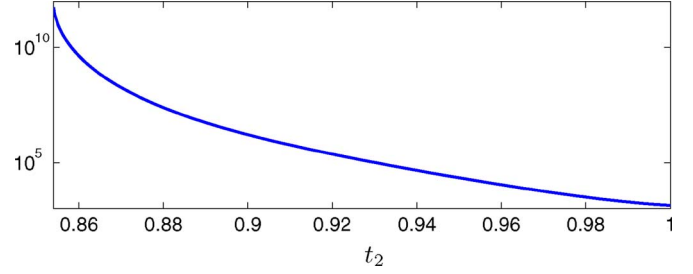

Fig. 8. CRB versus t_2 for fixed t_1 in a setting with sinusoidal sampling kernels with nonconsecutive frequencies.

contradicting the requirements (17) and (22) that

$$\|S(h(\boldsymbol{\theta}_1 + \boldsymbol{\delta})) - S(h(\boldsymbol{\theta}_1))\|_{\mathbb{R}^{2M}} \geq \alpha_s \alpha_h \|\boldsymbol{\delta}\|_{\mathbb{R}^{2M}}. \quad (52)$$

This can also be seen from an estimation viewpoint. Namely, suppose that the samples $\boldsymbol{c}$ are perturbed by white Gaussian noise with variance σ^2. Then the unbiased CRB for estimating $\boldsymbol{\theta}$ from these noisy measurements is given at $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ by [28]

$$\sigma^2 \left(\left(\left. \frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_1} \right)^* \left(\left. \frac{\partial \hat{\boldsymbol{c}}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}_1} \right) \right)^{-1}. \quad (53)$$

If $(\partial \hat{\boldsymbol{c}}/\partial \boldsymbol{\theta})|_{\boldsymbol{\theta}_1} = 0$ then there exists no unbiased technique that can recover the parameters with a finite MSE.

As a demonstration of the utilization of this approach, consider again the setting of Section VI-B. As mentioned above, existing techniques that do not involve discretization can only handle the case in which the frequencies of the sampling kernels are consecutive. An interesting question is whether there is a potential gain in using nonconsecutive indices. To study this setting, we used our algorithm to recover two time delays, where $g(t)$ was taken to be a pulse whose Fourier coefficients are equal 1 up to some large index and 0 otherwise. We used four sinusoidal sampling functions (two sines and two cosines) with frequencies 1 and 3. While the true parameters were $(\ t_1 \quad t_2 \quad a_1 \quad a_2\ ) = (\ 0.2 \quad 0.8 \quad 1 \quad 5\ )$, the algorithm converged to the point $(\ 0.34 \quad 0.85 \quad 0.41 \quad 3.1\ )$. This means that the CRB for estimating $\boldsymbol{\theta}$ explodes at this point. Fig. 8 depicts the CRB as function of $t_2 \in [0.85, 1]$ for $t_1 = 0.34$, verifying that this is indeed the case. We therefore conclude that in this setting there exist parameter values that *cannot be recovered stably by any technique*.

A word of caution is in place, though. For our approach to be able to recover a parameter vector $\boldsymbol{\theta}_0$, we need that *every* $\boldsymbol{\theta} \in \mathcal{A}$ can be stably recovered and not only $\boldsymbol{\theta}_0$ itself. Therefore, the fact that in some settings with nonconsecutive sampling frequencies there exist unstable points in $\mathcal{A}$ limits the applicability of our method in those scenarios. It may thus be of interest in certain applications to pursue methods that can recover any stably-reconstructible $\boldsymbol{\theta}_0$, regardless if there exist other points in $\mathcal{A}$ at which the CRB is infinite.

VII. APPLICATION TO CPM COMMUNICATION

As mentioned in Section II, an important application area not treated in the FRI literature is CPM communication (see (7)). For a general rational modulation index h and pulse width L, optimum coherent detection can be performed by means of the Viterbi algorithm. A major limitation with this approach,
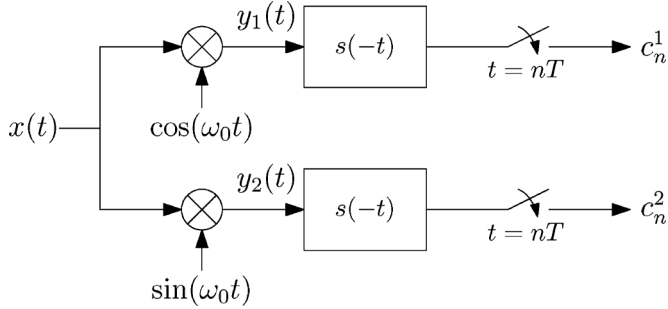
Fig. 9. Proposed CPM receiver.

though, is that it requires sampling at a rate of $1/T$ at the output of $4Q^L$ linear filters [33]. This corresponds to an over-sampling factor of $4Q^L$ beyond the rate of innovation. Furthermore, for $h = 2k/p$, where k and p have no common factors, the number of states in the Viterbi decoder is pQ^{L-1}. Here, we propose a suboptimal alternative, which employs an average sampling rate of only $2/T$, as depicted in Fig. 9. Our approach consists in treating the data symbols $\{a_m\}$ in (7) as continuous-valued and quantizing the resulting recoveries to the nearest element in the set $\{\pm 1, \ldots, \pm(Q-1)\}$. We emphasize that our proposed approach does not perform well in noise and serves here merely as a demonstration of treatment of non union-of-subspace models.

In principle, cleverly designed measurements at a rate of $1/T$ should suffice (in the noiseless setting) for perfect recovery. However, as we will see, neither of the branches of Fig. 9 suffices by itself for recovery of all symbols with our iterative approach. Instead, we will alternately use bunches of samples from each of the branches. The signals $y_1(t)$ and $y_2(t)$ contain one replica of the frequency content of $x(t)$ around $\omega = 0$ and one around $2\omega_0$. Suppose for the moment that the filter $s(-t)$ suppresses the replica around $2\omega_0$ so that, to high precision, for $i = 1, 2$,

$$y_i(t) = f_i\left(\sum_m \tilde{a}_m \tilde{g}(t - mT)\right) \tag{54}$$

where we adopted the representation (8) and denoted $f_1(\alpha) = 0.5\cos(\alpha)$ and $f_2(\alpha) = -0.5\sin(\alpha)$. Thus, for $i = 1, 2$,

$$c_n^i = \int_{-\infty}^{\infty} s(t - nT) f_i\left(\sum_{m \in \mathbb{Z}} \tilde{a}_m \tilde{g}(t - nT)\right) dt. \tag{55}$$

Linear sampling of a SI signal passing through memoryless nonlinearity, as in (55), was studied in [10], [15]. In particular, it was shown that if the nonlinearity is a monotone function that does not vary too rapidly, then a stationary point of the LS objective is necessarily a global minimum. In our setting, neither $f_1(\alpha)$ nor $f_2(\alpha)$ are monotone functions. However, since $\tilde{a}_m$ can only vary by ± 1 from one symbol to the next, the phase

$$\varphi(t) = \sum_m \tilde{a}_m \tilde{g}(t - mT) \tag{56}$$

is guaranteed to vary by less than $\pm \pi/2$ over short enough time segments. Specifically, $f_i(\varphi(t))$ is a monotone function of $\varphi(t)$ over a certain time interval if

$$(i-1)\pi/2 + 2\pi p < \varphi(t) < (i+1)\pi/2 + 2\pi p \tag{57}$$

or

$$(i+1)\pi/2 + 2\pi p < \varphi(t) < (i+3)\pi/2 + 2\pi p \tag{58}$$

for some $p \in \mathbb{Z}$ throughout this period. For such a segment $[t_1, t_2]$ and assuming that the support of $s(t)$ is contained in $[t_a, t_b]$, all samples c_n^i with indices $(t_1 - t_a)/T < n < (t_2 - t_b)/T$ conform to the model in [10] and [15]. These samples can be used to recover a corresponding set of symbols.

To summarize, our approach for the simple setting in which $s(t)$ is supported on $[0, T]$ is as follows. Suppose that all symbols up to index n_1 were recovered. These allow to determine $\varphi(n_1 T)$, which is used to decide, according to (57) and (58), weather the next batch of samples is to be taken from the first branch or from the second one. Next, the maximal index $n_{\max}$ such that the phase remains within the corresponding interval for every $t \in [(n_1 + 1)T, n_{\max}T]$ is determined.[5] The samples with indices $n_1 + 1, \ldots, n_{\max} - 1$ are then used to recover the symbols with the corresponding indices. This process is repeated sequentially.

The nth sample in the ith channel is given by $c_n^i = \langle y_i, s_n \rangle$, where $s_n(t) = s(t - nT)$. Assume, without loss of generality, that $\boldsymbol{\theta} = (a_1, \ldots, a_M)$. Direct computation shows that

$$\frac{\partial \hat{\boldsymbol{c}}^i}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \langle z_1^i, s_1 \rangle & \cdots & \langle z_M^i, s_1 \rangle \\ \vdots & & \vdots \\ \langle z_1^i, s_N \rangle & \cdots & \langle z_M^i, s_N \rangle \end{pmatrix} \tag{59}$$

where

$$z_m^1(t) = \frac{1}{2} q(t - mT)\left(\cos(2\omega_0 t + \varphi(t)) + \cos(\varphi(t))\right)$$

$$z_m^2(t) = \frac{1}{2} q(t - mT)\left(\sin(2\omega_0 t + \varphi(t)) - \sin(\varphi(t))\right) \tag{60}$$

and we denoted $q(t) = \int_{-\infty}^{t} g(\tau) d\tau$. To account for the fact that $|a_m| \leq Q - 1$, we chose to enforce the constraint $|a_m| < Q$ by using the parametrization $\tilde{\theta}_m = \tan(\pi a_m/(2Q))$. The derivative of the corresponding transformation $\boldsymbol{\theta} = p(\tilde{\boldsymbol{\theta}})$ is $\partial p/\partial \tilde{\boldsymbol{\theta}} = (2Q/\pi)\mathrm{diag}(1/(1 + \tilde{\theta}_1^2) \quad \cdots \quad 1/(1 + \tilde{\theta}_M^2))$.

Fig. 10 shows the phase of a typical binary CPM signal (namely, with $Q = 1$) with modulation index $h = 1/7$ and with the 5 REC pulse $g(t) = \mathrm{rect}_{[0,5T]}(t)$. Fig. 10(c) shows the recovery of the symbols with only 2 iterations per batch of samples. Here the sampling kernels were taken as $s(t) = \mathrm{rect}_{[0,T]}(t)$. The batches of samples on which the algorithm operated are marked with dashed vertical lines. As can be seen, even with two iterations, the original symbols can be recovered by quantization of the recovered symbols.

VIII. CONCLUSION

In this paper, we studied recovery of the parameters defining an FRI signal from samples taken at the rate of innovation. We showed that in any situation in which the parameters can be stably recovered, this can be achieved by a general-purpose unconstrained optimization method. Our approach thus provides a simple means for treating a wide range of FRI signal classes and sampling methods. We demonstrated the usefulness of our strategy in reconstructing finite and periodic pulse streams from nonlinear and nonideal samples as well as in decoding CPM

---

[5]This can be done by noting that the change in phase for $t \geq (n_1 + 1)T$ is due both to the contribution of the known symbols $\{a_m\}_{m \leq n_1}$ and to the symbols $\{a_m\}_{m > n_1}$, which are yet to be recovered. The largest change occurs if the latter are all $+1$ or -1.
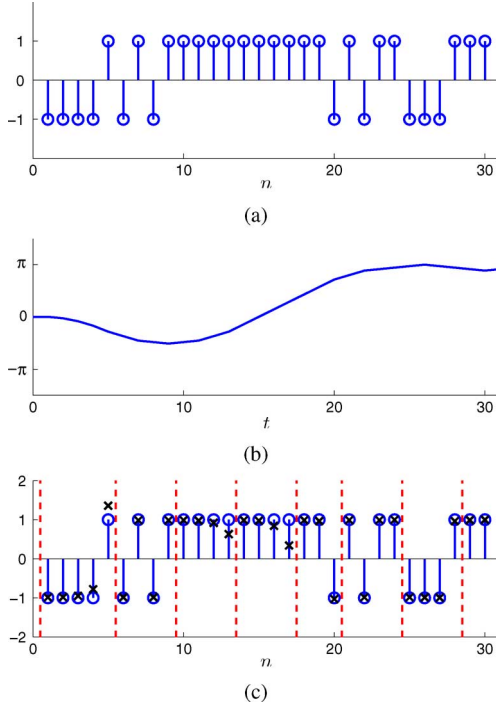
Fig. 10. Binary 5 REC CPM modulation with index $h = 1/7$. (a) Symbols a_m. (b) Corresponding phase $\varphi(t)$. (c) The "x"-marks denote the recovered coefficients using 2 Gauss-Newton iterations (before quantization).

modulated messages. We also showed that our method is often advantageous in noisy settings.

APPENDIX A
CONVERGENCE OF GAUSS-NEWTON ITERATIONS

Letting $Q = \partial \hat{c}/\partial \theta|_{\theta^\ell}$ and substituting $B^\ell = (Q^*Q)^{-1}$ and (31), the left-hand side of (36) becomes

$$
\frac{\left(\hat{c}(\theta^\ell) - c\right)^* Q(Q^*Q)^{-1}Q^* \left(\hat{c}(\theta^\ell) - c\right)}{\left\|(Q^*Q)^{-1}Q^* \left(\hat{c}(\theta^\ell) - c\right)\right\| \left\|Q^* \left(\hat{c}(\theta^\ell) - c\right)\right\|}
$$

$$
= \frac{\left\|\hat{c}(\theta^\ell) - c\right\|^2}{\left\|Q^{-1} \left(\hat{c}(\theta^\ell) - c\right)\right\| \left\|Q^* \left(\hat{c}(\theta^\ell) - c\right)\right\|}
$$

$$
\geq \frac{1}{\|Q^{-1}\|\|Q\|}. \tag{61}
$$

Here, we used the fact that $\mathcal{R}(Q) = \mathbb{R}^K$, which was established in the proof of Theorem 1, so that $Q(Q^*Q)^{-1}Q^* = I$. Now, the right-hand side of (22), together with (37), imply that $\|Q\| \leq \beta_s\beta_h$. Similarly, the left-hand side of (22), together with (17), imply that $\|Q^{-1}\| \leq 1/(\alpha_s\alpha_h)$. Therefore,

$$
\frac{1}{\|Q^{-1}\|\|Q\|} \geq \frac{\beta_s\beta_h}{\alpha_h\alpha_s} \tag{62}
$$

so that (36) is satisfied with any $\delta < (\beta_s\beta_h)/(\alpha_h\alpha_s)$.

APPENDIX B
PROOF OF GRADIENT LIPSCHITZ CONTINUITY

Denoting $e(\theta) = \hat{c}(\theta) - c$, we have

$$
\|\nabla \varepsilon(\theta_1) - \nabla \varepsilon(\theta_2)\|
$$

$$
= \left\| \left(\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1}\right)^* e(\theta_1) - \left(\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right)^* e(\theta_2)\right\|
$$

$$
= \frac{1}{2}\left\| \left(\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} - \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right)^* (e(\theta_1) + e(\theta_2))\right.
$$

$$
\left. + \left(\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} + \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right)^* (e(\theta_1) - e(\theta_2))\right\|
$$

$$
\leq \frac{1}{2}\left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} - \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right\| \|e(\theta_1) + e(\theta_2)\|
$$

$$
+ \frac{1}{2}\left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} + \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right\| \|e(\theta_1) - e(\theta_2)\|. \tag{63}
$$

Assuming that $\theta_1, \theta_2 \in \mathcal{N}$, conditions (22) and (38) imply that

$$
\left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} - \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right\| \leq \beta_s \left\|\frac{\partial h}{\partial \theta}\Big|_{\theta_1} - \frac{\partial h}{\partial \theta}\Big|_{\theta_2}\right\|
$$

$$
\leq \beta_s\beta_{h'}\|\theta_1 - \theta_2\|. \tag{64}
$$

Furthermore, (22) implies that

$$
\left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1} + \frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right\| \leq \left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_1}\right\| + \left\|\frac{\partial \hat{c}}{\partial \theta}\Big|_{\theta_2}\right\| \leq 2\beta_s. \tag{65}
$$

Since $\theta_1, \theta_2 \in \mathcal{N}$, it also follows that

$$
\|e(\theta_1) + e(\theta_2)\| \leq \|e(\theta_1)\| + \|e(\theta_2)\| \leq 2\varepsilon(\theta^0). \tag{66}
$$

Finally,

$$
\|e(\theta_1) - e(\theta_2)\| = \|\hat{c}(\theta_1) - \hat{c}(\theta_2)\| \leq \beta_s\|\theta_1 - \theta_2\|. \tag{67}
$$

Substituting (64), (65), (66) and (67) into (63) yields

$$
\|\nabla \varepsilon(\theta_1) - \nabla \varepsilon(\theta_2)\| \leq \left(\beta_s\beta_{h'}\varepsilon(\theta^0) + \beta_s^2\right)\|\theta_1 - \theta_2\|, \tag{68}
$$

which proves that $\nabla \varepsilon(\theta_1)$ is Lipschitz continuous over $\mathcal{N}$ with Lipschitz bound $\beta_s(\beta_{h'}\varepsilon(\theta^0) + \beta_s)$.

REFERENCES

[1] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.
[2] M. Unser, A. Aldroubi, and M. Eden, "B-Spline Signal Processing: Part I Theory," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 821–833, 1993.
[3] Y. C. Eldar and T. Michaeli, "Beyond bandlimited sampling," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 48–68, 2009.
[4] A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory," *Numer. Funct. Anal. Optim.*, vol. 15, no. 1, pp. 1–21, 1994.
[5] A. Aldroubi, "Oblique projections in atomic spaces," in *Proc. Amer. Math. Soc.*, 1996, vol. 124, no. 7, pp. 2051–2060.

[6] A. Aldroubi and K. Gröchenig, "Nonuniform sampling and reconstruction in shift-invariant spaces," *SIAM Rev.*, vol. 43, no. 4, pp. 585–620, 2001.

[7] Y. C. Eldar, "Sampling with arbitrary sampling and reconstruction spaces and oblique dual frame vectors," *J. Fourier Anal. Appl.*, vol. 9, no. 1, pp. 77–96, 2003.

[8] Y. C. Eldar and T. G. Dvorkind, "A minimum squared-error framework for generalized sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2155–2167, 2006.

[9] T. Michaeli and Y. C. Eldar, "Optimization techniques in modern sampling theory," in *Convex Optimization in Signal Processing and Communications*, Y. C. Eldar and D. Palomar, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010, pp. 266–314.

[10] T. G. Dvorkind, Y. C. Eldar, and E. Matusiak, "Nonlinear and non-ideal sampling: Theory and methods," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5874–5890, 2008.

[11] H. J. Landau and W. L. Miranker, "The recovery of distorted band-limited signals," *J. Math. Anal. Appl.*, vol. 2, pp. 97–104, 1961.

[12] I. W. Sandberg, "On the properties of some systems that distort signals I," *Bell Syst. Tech. J.*, vol. 42, pp. 2033–2047, 1963.

[13] E. Masry, "The recovery of distorted band-limited stochastic processes," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 398–403, 1973.

[14] I. W. Sandberg, "Notes on PQ theorems," *IEEE Trans. Circuits Syst. I*, vol. 41, no. 4, pp. 303–307, 1994.

[15] T. Faktor, T. Michaeli, and Y. C. Eldar, "Nonlinear and nonideal sampling revisited," *IEEE Trans. Signal Process.*, vol. 17, no. 2, pp. 205–208, 2010.

[16] 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.

[17] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 1741–1757, 2007.

[18] R. Tur, Y. C. Eldar, and Z. Friedman, "Low rate sampling of pulse streams with application to ultrasound imaging," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1827–1842, 2011.

[19] K. Gedalyahu, R. Tur, and Y. C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1491–1504, 2011.

[20] K. Gedalyahu and Y. C. Eldar, "Time delay estimation from low rate samples: A union of subspaces approach," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3017–3031, 2010.

[21] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334–2345, 2008.

[22] M. Mishali and Y. C. Eldar, "Blind multi-band signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 993–1009, 2009.

[23] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 375–391, 2010.

[24] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Dev., Syst.*, vol. 5, no. 1, pp. 8–20, 2011.

[25] M. Mishali and Y. C. Eldar, "Xampling: Compressed sensing of analog signals," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, to be published.

[26] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4719–4734, 2011.

[27] W. A. Frank, "Sampling requirements for Volterra system identification," *IEEE Signal Process. Lett.*, vol. 3, no. 9, pp. 266–268, 1996.

[28] Z. Ben-Haim, T. Michaeli, and Y. C. Eldar, "Performance bounds and design criteria for estimating finite rate of innovation signals," *IEEE Trans. Inf. Theory* [Online]. Available: http://arxiv.org/abs/1009.2221, submitted for publication

[29] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[30] Q. Sun, "Frames in spaces with finite rate of innovation," *Adv. Comput. Math.*, vol. 28, no. 4, pp. 301–329, 2008.

[31] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[32] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 5, pp. 814–824, 1990.

[33] C. E. Sundberg, "Continuous phase modulation," *IEEE Commun. Mag.*, vol. 24, no. 4, pp. 25–38, 1986.

Tomer Michaeli (S'12) received the B.Sc. degree in electrical engineering in 2004 (*summa cum laude*) from the Technion–Israel Institute of Technology, Haifa, Israel. He is currently working towards the Ph.D. degree in electrical engineering at the Technion.

From 2000 to 2008, he was a Research Engineer at RAFAEL Research Laboratories, Israel Ministry of Defense, Haifa. His research interests include statistical signal processing and sampling theory.

Mr. Michaeli was awarded the Andrew and Erna Finci Viterbi Fellowship in 2008, and in 2010, he was a Jacobs-QUALCOMM fellow. He won the Jury Award in 2011.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering both from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2001.

From January 2002 to July 2002, she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently a Professor in the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa, Israel. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and a Visiting Professor at Stanford University, Stanford, CA.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in Electrical Engineering at MIT, and in 2000, she held an IBM Research Fellowship. From 2002 to 2005, she was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, in 2008 the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, in 2009 the Technion's Award for Excellence in Teaching, and in 2010 the Michael Bruno Memorial Award from the Rothschild Foundation, and in 2011 the Weizmann Prize for Exact Sciences. She is a Signal Processing Society Distinguished Lecturer, a member of the IEEE Bio Imaging Signal Processing technical committee, an Associate Editor for the *SIAM Journal on Imaging Sciences*, and on the Editorial Board of *Foundations and Trends in Signal Processing*. In the past, she was a member of the IEEE Signal Processing Theory and Methods Technical Committee and served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, and the *SIAM Journal on Matrix Analysis and Applications*.