Performance Bounds and Design Criteria for Estimating Finite Rate of Innovation Signals

Zvika Ben-Haim, Tomer Michaeli, and Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*—In this paper, we consider the problem of estimating finite rate of innovation (FRI) signals from noisy measurements, and specifically analyze the interaction between FRI techniques and the underlying sampling methods. We first obtain a fundamental limit on the estimation accuracy attainable regardless of the sampling method. Next, we provide a bound on the performance achievable using any specific sampling approach. Essential differences between the noisy and noise-free cases arise from this analysis. In particular, we identify settings in which noise-free recovery techniques deteriorate substantially under slight noise levels, thus quantifying the numerical instability inherent in such methods. This instability, which is only present in some families of FRI signals, is shown to be related to a specific type of structure, which can be characterized by viewing the signal model as a union of subspaces. Finally, we develop a methodology for choosing the optimal sampling kernels for linear reconstruction, based on a generalization of the Karhunen–Loève transform. The results are illustrated for several types of time-delay estimation problems.**

***Index Terms*—Finite rate of innovation, Sampling, Cramér–Rao bound, Union of subspaces, Time-delay estimation**

I. Introduction

The field of digital signal processing hinges on the availability of techniques for sampling analog signals, thus converting them to discrete measurements. The sampling mechanism aims to preserve the information present in the analog domain, ideally permitting flawless recovery of the original signal. For example, one may wish to recover a continuous-time signal $x(t)$ from a discrete set of samples. The archetypical manifestation of this concept is the Shannon sampling theorem, which states that a B-bandlimited function can be reconstructed from samples taken at the Nyquist rate $2B$ [1].

Recently, considerable attention has been devoted to the extension of sampling theory to functions having a finite rate of innovation (FRI). These are signals determined by a finite number ρ of parameters per time unit [2]. Such a definition encompasses a rich variety of signals, including splines, shift-invariant signals, multiband signals, and pulse streams. In many FRI settings, several existing algorithms are guaranteed to recover the signal $x(t)$ from samples taken at rate ρ [2]–[9]. In other words, signals which correspond to the FRI model can be reconstructed from samples taken at the rate of innovation, which is potentially much lower than their Nyquist rate.

Real-world signals are often contaminated by continuous-time noise and thus do not conform precisely to the FRI model. Furthermore, like any mathematical model, the FRI framework suffers from mismodeling errors, i.e., the model holds only approximately in practical scenarios [10]. It is therefore of interest to quantify the effect of noise and mismodeling errors on FRI techniques. In the noisy case, it is no longer possible to perfectly recover the original signal from its samples. Nevertheless, one might hope for an appropriate finite-rate technique which achieves the best possible estimation accuracy, in the sense that increasing the sampling rate conveys no further performance benefits. For example, to recover a B-bandlimited signal contaminated by continuous-time white noise, one can use an ideal low-pass filter with cutoff B prior to sampling at a rate of $2B$. This strategy removes all noise components with frequencies larger than B, while leaving all signal components intact. Consequently, any alternative method which does not zero out frequencies above B can be improved upon, whereas methods which zero out some of the signal frequencies can suffer from an arbitrarily large reconstruction error. Thus, sampling at a rate of $2B$ is indeed optimal in the case of a B-bandlimited signal, if the signal is corrupted by continuous-time noise prior to sampling. Sampling at a rate higher than $2B$ can be beneficial only when the sampling process itself introduces additional noise into the system, e.g., as a result of quantization.

By contrast, empirical observations indicate that, for some noisy FRI signals, substantial performance improvements are achievable when the sampling rate is increased beyond the rate of innovation [4], [7], [8]. This phenomenon has also been studied in the specific setting of time delay estimation of rectangular pulses [11]. Thus, in some cases, there appears to be a fundamental difference between the noiseless and noise-corrupted settings, in terms of the required sampling rate. Our first goal in this paper will be to provide an analytical justification and quantification of these empirical findings. As we will see, the fact that oversampling improves performance is not merely indicative of flaws in existing algorithms; rather, it is a consequence of the inherent difficulty of reconstructing FRI signals under noise. Indeed, we will demonstrate that for some FRI signals, unless considerable oversampling is employed, performance will necessarily deteriorate by several orders of magnitude relative to the optimal achievable reconstruction capability. Such effects occur even when the noise level is exceedingly low. Our analysis will also enable us to identify and characterize the types of signals for which oversampling

Z. Ben-Haim was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. He is now with Google Israel R&D Center, PO Box 15096, Haifa 31905, Israel (e-mail: zvika@google.com). T. Michaeli and Y. C. Eldar are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: {tomermic@tx, yonina@ee}.technion.ac.il). This work was supported in part by a Magneton grant from the Israel Ministry of Industry and Trade, by the Ollendorf foundation, and by the Israel Science Foundation by Grant 170/10.

is necessary.

To demonstrate these results, we first derive the Cramér–Rao bound (CRB) for estimating a finite-duration segment of an FRI signal $x(t)$ directly from continuous-time measurements $y(t) = x(t) + w(t)$, where $w(t)$ is a Gaussian white noise process. This yields a lower bound on the accuracy whereby $x(t)$ can be recovered by any unbiased technique, regardless of its sampling rate.

The continuous-time measurement setting we study is to be distinguished from previous bounds in the FRI literature [12], [13] in four respects. First and most importantly, the measurements are a continuous-time process $y(t)$ and the bound therefore applies regardless of the sampling method. Second, in our model, the noise is added prior to sampling. Thus, as will be shown below, even sampling at an arbitrarily high rate will not completely compensate for the noise. Third, we bound the mean-squared error (MSE) in estimating $x(t)$ and not the parameters defining it, since we seek to determine the accuracy with which $x(t)$ itself can be recovered. Such a bound does not depend on the specific parametrization of the signal, and consequently, possesses a simpler analytical expression. Fourth, contrary to previous work, our analysis does not assume a specific FRI family. In particular, we do not limit ourselves to periodic time-delay estimation settings.

In practice, rather than processing the continuous-time signal $y(t)$, it is typically desired to estimate $x(t)$ from a discrete set of samples $\{c_n\}$ of $y(t)$. In this scenario, in addition to the continuous-time noise $w(t)$, digital noise may arise from the sampling process itself, for example due to quantization. To quantify the extent to which sampling degrades the ability to recover the signal, we next derive the CRB for estimating $x(t)$ from the measurements $\{c_n\}$. This analysis depends on the relative power of the two noise factors. When only digital noise is present, oversampling can be used to completely overcome its effect. On the other hand, when there exists only continuous-time noise, the bound converges to the continuous-time CRB as the sampling rate increases. In some cases, these bounds coincide at a finite sampling rate, which implies that the sampling scheme has captured all of the information present in the continuous-time signal, and any further increase in the sampling rate is useless. Conversely, when the continuous-time and sampled CRBs differ, the gap between these bounds is indicative of the degree to which information is lost in the sampling process. Our technique can then be used to plot the best possible performance as a function of the sampling rate, and thus provide the practitioner with a tool for evaluating the benefits of oversampling.

When a certain sampling technique achieves the performance of continuous-time measurements, it can be identified using the method described above. However, in some cases no such technique exists, or the sampling rate it requires may be prohibitive. In these cases, it is desirable to determine the optimal sampling scheme having an allowed rate. Since different signals are likely to be recovered successfully with different sampling kernels, a Bayesian or average-case analysis is well-suited for this problem. Specifically, we assume that the signal $x(t)$ has a known prior distribution over the class of signals, and determine the linear sampling and reconstruction technique which minimizes the MSE for recovering $x(t)$ from its measurements. While nonlinear reconstruction techniques are commonly used and typically outperform the best linear estimator, this approach provides a simple means for identifying an appropriate sampling method. The resulting method can then be used in conjunction with standard nonlinear FRI recovery algorithms.

We demonstrate our results via the problem of estimating a finite-duration sequence of pulses having unknown positions and amplitudes [2], [4], [5], [8]. In this case, a simple sufficient condition is obtained for the existence of a sampling scheme whose performance bound coincides with the continuous-time CRB. This scheme is based on sampling the Fourier coefficients of the pulse shape, and is reminiscent of recent time-delay estimation algorithms [7], [8]. However, while the sampling scheme is theoretically sufficient for optimal recovery of $x(t)$, we show that in some cases there is room for substantial improvement in the reconstruction stage of these algorithms. Finally, we demonstrate that the Fourier domain is also optimal (in the sense of minimizing the reconstruction MSE) when the sampling budget is limited. Specifically, given an allowed number of samples N, the reconstruction MSE is minimized by sampling the N highest-variance Fourier coefficients of the signal $x(t)$.

The rest of this paper is organized as follows. The problem setting is defined in Section II, and some examples of signals conforming to this model are presented in Section III. We then briefly summarize our main results in Section IV. In Section V, we provide a technical generalization of the CRB to general spaces. This result is used to obtain bounds on the achievable reconstruction error from continuous-time measurements (Section VI) and using a sampling mechanism (Section VII). Next, in Section VIII a Bayesian viewpoint is introduced and utilized to determine the optimal sampling kernels having a given rate budget. The results are demonstrated for the specific signal model of time-delay estimation in Section IX.

II. Definitions

A. Notation

The following notation is used throughout the paper. A boldface lowercase letter $\mathbf{v}$ denotes a vector, while a boldface uppercase letter $\mathbf{M}$ denotes a matrix. $\mathbf{I}_N$ is the $N \times N$ identity matrix. For a vector $\mathbf{v}$, the notation $\|\mathbf{v}\|$ indicates the Euclidean norm, and $\mathbf{v}_k$ is the kth element. Given a complex number $z \in \mathbb{C}$, the symbols z^* and $\Re\{z\}$ denote the complex conjugate and the real part of z, respectively. For an operator P, the range space and null space are $\mathcal{R}(P)$ and $\mathcal{N}(P)$, respectively, while the trace and adjoint are denoted, respectively, by $\mathrm{Tr}(P)$ and P^*. The Kronecker delta, denoted by $\delta_{m,n}$, equals 1 when $m = n$ and 0 otherwise. The expectation of a random variable v is written as $\mathbb{E}\{v\}$.

The Hilbert space of square-integrable complex-valued functions over $[0, T_0]$ is denoted $L_2[0, T_0]$ or simply L_2. The corresponding inner product is

$$\langle f, g \rangle \triangleq \int_0^{T_0} f(t) g^*(t) dt \tag{1}$$

and the induced norm is $\|f\|_{L_2}^2 \triangleq \langle f, f\rangle$. For an ordered set of K functions $g_1, \ldots, g_K$ in L_2, we define the associated *set transformation* $G : \mathbb{C}^K \to L_2$ as

$$(G\mathbf{v})(t) = \sum_{k=1}^{K} \mathbf{v}_k g_k(t). \tag{2}$$

By the definition of the adjoint, it follows that

$$G^* f = \left(\langle f, g_1\rangle, \ldots, \langle f, g_K\rangle\right)^T. \tag{3}$$

B. Setting

In this work, we are interested in the problem of estimating FRI signals from noisy measurements. To define FRI signals formally, let the T_0-local number of degrees of freedom $N_{T_0}(t)$ of a signal $x(t)$ at time t be the number of parameters defining the segment $\{x(t) : t \in [t - T_0/2, t + T_0/2]\}$. The T_0-local rate of innovation of $x(t)$ is then defined as [2]

$$\rho_{T_0} = \max_{t\in\mathbb{R}} \frac{N_{T_0}(t)}{T_0}. \tag{4}$$

We say that $x(t)$ is an FRI signal if ρ_{T_0} is finite for all sufficiently large values of T_0. In Section III, we will give several examples of FRI signals and compute their rates of innovation.

For concreteness, let us focus on the problem of estimating the finite-duration segment $\{x(t) : t \in [0, T_0]\}$, for some constant T_0, and let $K \triangleq N_{T_0}(T_0/2)$ denote the number of parameters defining this segment. We then have

$$x \in \mathfrak{X} \triangleq \{h_{\boldsymbol{\theta}} \in L_2[0, T_0] : \boldsymbol{\theta} \in \Theta\} \tag{5}$$

where $h_{\boldsymbol{\theta}}$ is a set of functions parameterized by the vector $\boldsymbol{\theta}$, and Θ is an open subset of $\mathbb{R}^K$.

We wish to examine the random process

$$y(t) = x(t) + w(t), \quad t \in [0, T_0] \tag{6}$$

where $w(t)$ is continuous-time white Gaussian noise. Recall that formally, it is not possible to define Gaussian white noise over a continuous-time probability space [14]. Instead, we interpret (6) as a simplified notation for the equivalent set of measurements

$$z(t) = \int_0^t x(\tau)d\tau + \sigma_c b(t), \quad t \in [0, T_0] \tag{7}$$

where $b(t)$ is a standard Wiener process (also called Brownian motion) [15]. It follows that $w(t)$ can be considered as a random process such that, for any $f, g \in L_2$, the inner products $a = \langle f, w\rangle$ and $b = \langle g, w\rangle$ are zero-mean jointly Gaussian random variables satisfying $\mathbb{E}\{ab^*\} = \sigma_c^2 \langle f, g\rangle$ [14]. The subscript c in σ_c is meant as a reminder of the fact that $w(t)$ is continuous-time noise. By contrast, when examining samples of the random process $y(t)$, we will also consider digital noise which is added during the sampling process.

In this paper, we consider estimators which are functions either of the entire continuous-time process (6) or of some subset of the information present in (6), such as a discrete set of samples of $y(t)$. To treat these two cases in a unified way, let $(\Omega, \mathscr{F})$ be a measurable space and let $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a family of probability measures over $(\Omega, \mathscr{F})$. Let $(\mathcal{Y}, \mathscr{U})$ be a measurable space, and let the random variable $y : \Omega \to \mathcal{Y}$ denote the measurements. This random variable can represent either $y(t)$ itself or samples thereof.

An estimator can be defined in this general setting as a measurable function $\hat{x} : \mathcal{Y} \to L_2$. The MSE of an estimator $\hat{x}$ at x is defined as

$$\mathrm{MSE}(\hat{x}, x) \triangleq \mathbb{E}\left\{\|\hat{x} - x\|_{L_2}^2\right\} = \mathbb{E}\left\{\int_0^{T_0} |\hat{x}(t) - x(t)|^2 dt\right\}. \tag{8}$$

An estimator $\hat{x}$ is said to be unbiased if

$$\mathbb{E}\{\hat{x}(t)\} = x(t) \text{ for all } x \in \mathfrak{X} \text{ and almost all } t \in [0, T_0]. \tag{9}$$

In the next section, we demonstrate the applicability of our model by reviewing several scenarios which can be formulated using the FRI framework. Some of these settings will also be used in the sequel to exemplify our theoretical results.

III. Types of FRI Signals

Numerous FRI signal structures have been proposed and analyzed in the sampling literature. Whereas most of these can be treated within our framework, some FRI structures do not conform exactly to our problem setting. Thus, before delving into the derivation of the CRB, we first provide examples for scenarios that can be analyzed via our model and discuss some of its limitations.

A. Shift-Invariant Spaces

Consider the class of signals that can be expressed as

$$x(t) = \sum_{m\in\mathbb{Z}} a[m] g(t - mT) \tag{10}$$

with some arbitrary square-integrable sequence $\{a[m]\}_{m\in\mathbb{Z}}$, where $g(t)$ is a given pulse in $L_2(\mathbb{R})$ and $T > 0$ is a given scalar. This set of signals is a linear subspace of $L_2(\mathbb{R})$, which is often termed a *shift-invariant* (SI) space [16], [17]. The class of functions that can be represented in the form (10) is quite large. For example, choosing $g(t) = \mathrm{sinc}(t/T)$ leads to the subspace of π/T-bandlimited signals. Other important examples include the space of spline functions (obtained by letting $g(t)$ be a B-spline function) and communication signals such as pulse-amplitude modulation (PAM) and quadrature amplitude modulation (QAM). Reconstruction in SI spaces from noiseless samples has been addressed in [18], [19] and extended to the noisy setting in [20]–[22].

Intuitively, every signal lying in a SI space with spacing T has one degree of freedom per T seconds (corresponding to one coefficient from the sequence $\{a[m]\}$). It is thus tempting to regard the rate of innovation of such signals as $1/T$. However, this is only true in an asymptotic sense and for compactly supported pulses $g(t)$. For any finite window size T_0, the T_0-local rate of innovation ρ_{T_0} is generally larger. Specifically, suppose that the support of $g(t)$ is contained in $[t_a, t_b]$ and consider intervals of the form $[t, t + MT]$, where M is an integer. Then, due to the overlaps of the pulses, for any such interval we can only assure that there are no more

than $M + \lceil (t_b - t_a)/T \rceil$ coefficients affecting the values of $x(t)$. Thus, the MT-local rate of innovation of signals of the form (10) is given by

$$\rho_{MT} = \frac{1}{T}\left(1 + \frac{\lceil \frac{t_b - t_a}{T} \rceil}{M}\right). \tag{11}$$

In particular, signals of the form (10) having a generator $g(t)$ which is not compactly supported have an infinite T_0-local rate of innovation, for any finite T_0. This is the case, for example, with bandlimited signals, which are therefore not FRI functions under our definition. As will be discussed in the sequel, this is not a flaw of the definition we use for the rate of innovation. Rather, it reflects the fact that it is impossible to recover any finite-duration segment $[T_1, T_2]$ of such signals from a finite number of measurements.

B. Nonlinearly-Distorted Shift-Invariant Spaces

In certain communication scenarios, nonlinearities are introduced in order to avoid amplitude clipping, an operation known as companding [23]. When the original signal lies in a SI space, the resulting transmission takes the form

$$x(t) = r\left(\sum_{m\in\mathbb{Z}} a[m]g(t - mT)\right), \tag{12}$$

where $r(\cdot)$ is a nonlinear, invertible function. Clearly, the MT-local rate of innovation ρ_{MT} of this type of signals is the same as that of the underlying SI function, and is thus given by (11). The recovery of nonlinearly distorted SI signals from noiseless samples was treated in [23]–[26]. We are not aware of research work treating the noisy case.

C. Union of Subspaces

Much of the FRI literature treats signal classes which are unions of subspaces [5], [7], [10], [27]. We now give examples of a few of these models.

1) Finite Union of Subspaces: There are various situations in which a continuous-time signal is known to belong to one of a finite set of spaces. One such signal model is described by

$$x(t) = \sum_{m\in\mathbb{Z}}\sum_{k=1}^{K} a_k[m]g_k(t - mT), \tag{13}$$

where $\{g_k(t)\}_{k=1}^K$ are a set of generators. In this model it is assumed that only $L < K$ out of the K sequences $\{a_1[m]\}_{m\in\mathbb{Z}}, \ldots, \{a_K[m]\}_{m\in\mathbb{Z}}$ are not identically zero [28]. Therefore, the signal $x(t)$ is known to reside in one of $\binom{K}{L}$ spaces, each of which is spanned by an L-element subset of the set of generators $\{g_k(t)\}_{k=1}^K$. This class of functions can be used to describe multiband signals [27], [29]. However, the discrete nature of these models precludes analysis using the differential tools employed in the remainder of this paper. Therefore, in this work we will focus on infinite unions of subspaces.

2) Single-Burst Channel Sounding: In certain medium identification and channel sounding scenarios, the echoes of a transmitted pulse $g(t)$ are analyzed to identify the positions and reflectance coefficients of scatterers in the medium [8], [30]. In these cases, the received signal has the form

$$x(t) = \sum_{\ell=1}^{L} a_\ell g(t - t_\ell), \tag{14}$$

where L is the number of scatterers and the amplitudes $\{a_\ell\}_{\ell=1}^L$ and time-delays $\{t_\ell\}_{\ell=1}^L$ correspond to the reflectance and location of the scatterers. Such signals can be thought of as belonging to a union of subspaces, where the parameters $\{t_\ell\}_{\ell=1}^L$ determine an L-dimensional subspace, and the coefficients $\{a_\ell\}_{\ell=1}^L$ describe the position within the subspace. In contrast with the previous example, however, in this setting we have a union of an infinite number of subspaces, since there are infinitely many possible values for the parameters $t_1, \ldots, t_L$.

In this case, for any window of size $T_0 > \max_\ell\{t_\ell\} - \min_\ell\{t_\ell\}$, the T_0-local rate of innovation is given by

$$\rho_{T_0} = \frac{2L}{T_0}. \tag{15}$$

3) Periodic Channel Sounding: Occasionally, channel sounding techniques consist of repeatedly probing the medium [31]. Assuming the medium does not change throughout the experiment, the result is a periodic signal

$$x(t) = \sum_{m\in\mathbb{Z}}\sum_{\ell=1}^{L} a_\ell g(t - t_\ell - mT). \tag{16}$$

As before, the set $\mathfrak{X}$ of feasible signals is an infinite union of finite-dimensional subspaces in which $\{t_\ell\}_{\ell=1}^L$ determine the subspace and $\{a_\ell\}_{\ell=1}^L$ define the position within the subspace. The T_0-local rate of innovation in this case coincides with (15).

4) Semi-Periodic Channel Sounding: There are situations in which a channel consists of L paths whose amplitudes change rapidly, but the time delays can be assumed constant throughout the duration of the experiment [5], [31], [32]. In these cases, the output of a channel sounding experiment will have the form

$$x(t) = \sum_{m\in\mathbb{Z}}\sum_{\ell=1}^{L} a_\ell[m]g(t - t_\ell - mT), \tag{17}$$

where $a_\ell[m]$ is the amplitude of the ℓth path at the mth probing experiment. This is, once again, a union of subspaces, but here each subspace is infinite-dimensional, as it is determined by the infinite set of parameters $\{a_\ell[m]\}$. In this case, the MT-local rate of innovation can be shown to be

$$\rho_{MT} = \frac{L}{T}\left(1 + \frac{1 + \lceil \frac{t_b - t_a}{T} \rceil}{M}\right). \tag{18}$$

5) Multiband Signals: Multiuser communication channels are often characterized by a small number of utilized subbands interspersed by large unused frequency bands [29]. The resulting signal can be described as

$$x(t) = \sum_{n\in\mathbb{Z}}\sum_{\ell=1}^{L} a_\ell[n]g(t - nT)e^{j\omega_\ell t}, \tag{19}$$

where $\{a_\ell[n]\}_{n\in\mathbb{Z}}$ is the data transmitted by the ℓth user, and ω_ℓ is the corresponding carrier frequency. In some cases the transmission frequencies are unknown [27], [29], resulting in an infinite union of infinite-dimensional subspaces. This setting is analogous in many respects to the semi-periodic channel sounding case; in particular, the MT-local rate of innovation can be shown to be the same as that given by (18).

IV. Summary of Main Results

In this section, we provide a high-level description of our main contributions and summarize the resulting conclusions. An overview of the dependencies between the remaining sections in this paper is provided in Fig. 1.

The overarching objective of this paper is to design and analyze sampling schemes for reconstructing FRI signals from noisy measurements. This goal is accomplished in four stages. We first derive a general form of the CRB suitable for the estimation of continuous-time signals (Section V). We then apply this result to identify the best achievable MSE for estimating an FRI signal $x(t)$ from its continuous-time measurements $y(t) = x(t) + w(t)$, providing a fundamental lower bound which is independent of the sampling method (Section VI). Next, we compare this continuous-time bound with the lowest possible MSE for a given sampling scheme, thus measuring the loss entailed in any particular technique (Section VII). Finally, we provide a mechanism for choosing the optimal sampling kernels (in a specific Bayesian sense), under the assumption of a linear reconstruction scheme (Section VIII). Our results can be applied to specific families of FRI signals, but they also yield some general conclusions as to the relative difficulty of various classes of estimation problems. These general observations are summarized below.

It should be noted that our CRB analysis is focused on unbiased estimators. This is a standard assumption which is required in order to obtain meaningful bounds [33]. For high values of the signal-to-noise ratio (SNR), unbiased estimators are optimal, and the CRB is then of practical relevance. However, under low SNR, it is quite possible for techniques to outperform the CRB. These effects are further discussed in Section IX, where we show that the CRB typically predicts the performance of practical techniques very accurately, but does occasionally fail to do so in specific cases.

A. Continuous-Time Bound

Our first goal in this paper is to derive the continuous-time CRB, which defines a fundamental limit on the accuracy with which an FRI signal can be estimated, regardless of the sampling technique. This bound turns out to have a particularly simple closed form expression which depends on the rate of innovation, but not on the class $\mathfrak{X}$ of FRI signals being estimated. Specifically, as we show in Theorem 2 in Section VI, the MSE of any unbiased estimator $\hat{x}$ satisfies

$$\frac{1}{T_0}\mathrm{MSE}(\hat{x},x) \geq \rho_{T_0}\sigma_c^2. \tag{20}$$

Thus, the rate of innovation can be given a new interpretation as the ratio between the best achievable MSE and the noise variance σ_c^2. This is to be contrasted with the characterization of the rate of innovation in the noise-free case as the lowest sampling rate allowing for perfect recovery of the signal; indeed, when noise is present, perfect recovery is no longer possible.

B. Bound for Sampled Measurements

In Section VII, we consider lower bounds for estimating $x(t)$ from *samples* of the signal $y(t)$. In this setting, the samples inherit the noise $w(t)$ embedded in the signal $y(t)$, and may suffer from additional discrete-time noise, for example, due to quantization. We derive the CRB for estimating $x(t)$ from sampled measurements in the presence of both types of noise. However, since the discrete-time noise model has been previously analyzed [12], [13], we only discuss the fundamental distinctions between both models and mostly focus in this paper on the assumption that the sampling noise is negligible.

In this setting, the sampled CRB can be designed so as to converge to the continuous-time bound as the sampling rate increases. Moreover, if the family $\mathfrak{X}$ of FRI signals is contained in a finite-dimensional subspace $\mathcal{M}$ of L_2, then a sampling scheme achieving the continuous-time CRB can be constructed. Such a sampling scheme is obtained by choosing kernels which span the subspace $\mathcal{M}$, and yields samples which fully capture the information present in the signal $y(t)$. Contrariwise, if $\mathfrak{X}$ is not contained in a finite-dimensional subspace, then no finite-rate sampling method achieves the continuous-time CRB. In this case, any increase in the sampling rate can improve performance, and the continuous-time bound is obtained only asymptotically.

It is interesting to examine this distinction from a union of subspaces viewpoint. Suppose that, as in the examples of Section III-C, the family $\mathfrak{X}$ can be described as a union of an infinite number of subspaces $\{\mathcal{U}_\alpha\}$ indexed by the continuous parameter α, so that

$$\mathfrak{X} = \bigcup_\alpha \mathcal{U}_\alpha. \tag{21}$$

In this case, a finite sampling rate captures all of the information present in the signal if and only if

$$\dim\left(\sum_\alpha \mathcal{U}_\alpha\right) < \infty \tag{22}$$

where $\dim(\mathcal{M})$ is the dimension of the subspace $\mathcal{M}$. By contrast, in the noise-free case, it has been previously shown [34] that the number of samples required to recover $x(t)$ is given by

$$\max_{\alpha_1,\alpha_2} \dim(\mathcal{U}_{\alpha_1} + \mathcal{U}_{\alpha_2}), \tag{23}$$

i.e., the largest dimension among sums of *two* subspaces belonging to the union. In general, the dimension of (22) will be much higher than (23), illustrating the qualitative difference between the noisy and noise-free settings. For example, if the subspaces $\mathcal{U}_\alpha$ are finite-dimensional, then (23) is also necessarily finite, whereas (22) need not be.

Nevertheless, one may hope that the structure embodied in $\mathfrak{X}$ will allow *nearly* optimal recovery using a sampling rate

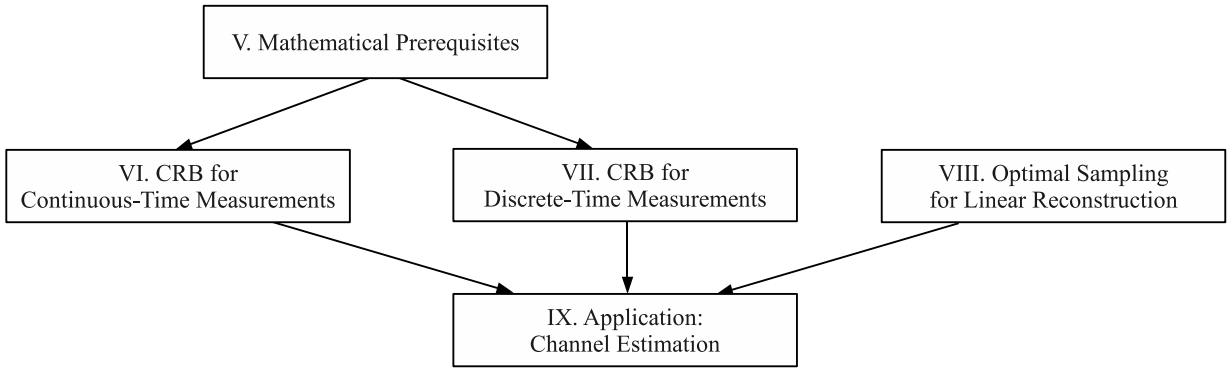


Fig. 1. Overview of the results presented in this paper.

close to the rate of innovation. This is certainly the case in many noise-free FRI settings. For example, there exist techniques which recover the pulse stream (14) from samples taken at the rate of innovation, despite the fact that in this case $\mathcal{X}$ is typically not contained in a finite-dimensional subspace. However, this situation often changes when noise is added, in which case standard techniques improve considerably under oversampling. This empirical observation can be quantified using the CRB: as we show, the CRB for samples taken at the rate of innovation is substantially higher in this case than the optimal, continuous-time bound. This demonstrates that the sensitivity to noise is a fundamental aspect of estimating signals of the form (14), rather than a limitation of existing algorithms. On the other hand, other FRI models, such as the semi-periodic pulse stream (17), exhibit considerable noise resilience, and indeed in these cases the CRB converges to the continuous-time value much more quickly.

As we discuss in Section IX-E, the different levels of robustness to noise can be explained when the signal models are examined in a union of subspaces context. In this case, the parameters $\boldsymbol{\theta}$ defining $x(t)$ can be partitioned into parameters defining the subspace $\mathcal{U}_\alpha$ and parameters pinpointing the position within the subspace. Our analysis hints that estimation of the position within a subspace is often easier than estimation of the subspace itself. Thus, when most parameters are used to select an intra-subspace position, estimation at the rate of innovation is successful, as occurs in the semi-periodic case (17). By contrast, when a large portion of the parameters define the subspace in use, a sampling rate higher than the rate of innovation is necessary; this is the case in the non-periodic pulse stream (14), wherein $\boldsymbol{\theta}$ is evenly divided among subspace-selecting parameters $\{t_\ell\}$ and intra-subspace parameters $\{a_\ell\}$. Thus we see that the CRB, together with the union of subspaces viewpoint, provide valuable insights into the relative degrees of success of various FRI estimation techniques.

C. Choosing the Sampling Kernels

In some cases, one may choose the sampling system according to design specifications, leading to the question: What sampling kernels should be chosen given an allotted number of samples? We tackle this problem in Section VIII by adopting a Bayesian framework, wherein the signal $x(t)$ is a complex random process distributed according to a known prior distribution. We further assume that both the sampling and reconstruction techniques are linear. While nonlinear reconstruction methods are typically used for estimating FRI signals, this assumption is required for analytical tractability, and is used only for the purpose of identifying sampling kernels. Once these kernels are chosen, they can be used in conjunction with nonlinear reconstruction algorithms (though in this case no analytic optimality conditions can be provided).

Under these assumptions, we identify the sampling kernels yielding the minimal MSE. An additional advantage of our assumption of linearity is that in this case, the optimal kernels depend only on the autocorrelation

$$R_X(t,\tau) = \mathbb{E}\{x(t)x^*(\tau)\} \tag{24}$$

of the signal $x(t)$, rather than on higher-order statistics. Indeed, given a budget of N samples, the optimal sampling kernels are given by the N eigenfunctions of R_X corresponding to the N largest eigenvalues. This is reminiscent of the Karhunen–Loève transform (KLT), which can be used to identify the optimal sampling kernels in the noiseless setting. However, in our case, shrinkage is applied to the measurements prior to reconstruction, as is typically the case with Bayesian estimation of signals in additive noise.

A setting of particular interest occurs when the autocorrelation R_X is cyclic, in the sense that

$$R_X(t,\tau) = R_X((t-\tau) \bmod T) \tag{25}$$

for some T. This scenario occurs, for example, in the periodic pulse stream (16) and the semi-periodic pulse stream (17), assuming a reasonable prior distribution on the parameters $\boldsymbol{\theta}$. It is not difficult to show that the eigenfunctions of R_X are given, in this case, by the complex exponentials

$$\psi_n(t) = \frac{1}{\sqrt{T}} e^{j\frac{2\pi}{T}nt}, \quad n \in \mathbb{Z}. \tag{26}$$

Furthermore, in the case of the periodic and semi-periodic pulse streams, the magnitudes of the eigenvalues of R_X are directly proportional to the magnitudes of the respective Fourier coefficients of the pulse shape $g(t)$. It follows that the optimal sampling kernels are the exponentials (26) corresponding to the largest Fourier coefficients of $g(t)$. This result is encouraging in light of recently proposed FRI reconstruction

TABLE I
CRB SETTINGS

		Measurements	
		Discrete	Continuous
Estimand	Discrete	Standard [35]	[15]
	Continuous, Unconstrained	No unbiased estimator	
	Continuous, Constrained	Section VII	Section VI

techniques which utilize exponential sampling kernels [7], and demonstrates the suitability of the Bayesian approach for designing practical estimation kernels.

V. Mathematical Prerequisites: CRB for General Parameter Spaces

In statistics and signal processing textbooks, the CRB is typically derived for parameters belonging to a finite-dimensional Euclidean space [33], [35], [36]. However, this result is insufficient when it is required to estimate a signal x belonging to other Hilbert spaces, such as the L_2 space defined above. When no knowledge about the structure of x is available, a bound for estimation from measurements contaminated by *colored* noise was derived in [37]. However, this bound does not hold when the noise $w(t)$ is white. Indeed, in the white noise case, it can be shown that no finite-MSE unbiased estimators exist, unless further information about $x(t)$ is available. For example, the naive estimator $\hat{x}(t) = y(t)$ has an error $\hat{x}(t) - x(t)$ equal to $w(t)$, whose variance is infinite.

In our setting, we are given the additional information that x belongs to the constraint set $\mathfrak{X}$ of (5). To the best of our knowledge, the CRB has not been previously defined for any type of constraint set $\mathfrak{X} \subset L_2$, a task which will be accomplished in the present section. As we show below, a finite-valued CRB can be constructed by requiring unbiasedness only within the constraint set $\mathfrak{X}$, as per (9). As we will see, the CRB increases linearly with the dimension of the manifold $\mathfrak{X}$. Thus, in particular, the CRB is infinite when $\mathfrak{X} = L_2$. However, for FRI signals, the dimension of $\mathfrak{X}$ is finite by definition, implying that a finite-valued CRB can be constructed. Although the development of this bound invokes some deep concepts from measure theory, it is a direct analog of the CRB for finite-dimensional parameters [33, Th. 2.5.15].

Table I summarizes the various CRB settings treated in the literature and in this work.

To derive the bound in the broadest setting possible, in this section we temporarily generalize the scenario of Section II, and consider estimation of a parameter x belonging to an arbitrary measurable and separable Hilbert space $\mathcal{H}$. The MSE of an estimator $\hat{x}$ in this setting is defined as $\mathrm{MSE}(\hat{x}, x) = \mathbb{E}\{\|\hat{x} - x\|_{\mathcal{H}}^2\}$. The concept of bias can similarly be extended if one defines expectation in the Pettis sense [38]. Specifically, the Pettis expectation $\mathbb{E}\{v\}$ of a random variable $v : \Omega \to \mathcal{H}$ is defined as an element $k \in \mathcal{H}$ such that $\langle k, \varphi \rangle = \mathbb{E}\{\langle v, \varphi \rangle\}$ for any $\varphi \in \mathcal{H}$. If no such element exist, then the expectation is said to be undefined.

The derivation of the CRB requires the existence of a "probability density" $p_{\boldsymbol{\theta}}(y)$ (more precisely, a Radon–Nikodym derivative of the probability measure) which is differentiable with respect to $\boldsymbol{\theta}$, and such that its differentiation with respect to $\boldsymbol{\theta}$ can be interchanged with integration with respect to y. The CRB also requires the mapping $h_{\boldsymbol{\theta}}$ from $\boldsymbol{\theta}$ to x to be non-redundant and differentiable. The formal statement of these regularity conditions is provided below. For the measurement setting (6), with reasonable mappings $h_{\boldsymbol{\theta}}$, these conditions are guaranteed to hold, as we will demonstrate in the sequel; in this section, however, we list these conditions in full so that a more general statement of the CRB will be possible.

P1) There exists a value $\boldsymbol{\theta}_0 \in \Theta$ such that the measure $P_{\boldsymbol{\theta}_0}$ dominates $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. In other words, there exists a Radon–Nikodym derivative $p_{\boldsymbol{\theta}}(y) \triangleq dP_{\boldsymbol{\theta}}/dP_{\boldsymbol{\theta}_0}$ such that, for any event $A \in \mathscr{U}$,

$$P_{\boldsymbol{\theta}}(A) = \int_A p_{\boldsymbol{\theta}}(y) P_{\boldsymbol{\theta}_0}(dy). \tag{27}$$

P2) For all y such that $p_{\boldsymbol{\theta}}(y) > 0$, the functions $p_{\boldsymbol{\theta}}(y)$ and $\log p_{\boldsymbol{\theta}}(y)$ are continuously differentiable with respect to $\boldsymbol{\theta}$. We denote by $\partial p_{\boldsymbol{\theta}}(y)/\partial \boldsymbol{\theta}$ and $\partial \log p_{\boldsymbol{\theta}}(y)/\partial \boldsymbol{\theta}$ the column vectors of the gradients of these two functions.

P3) The support $\{y \in \mathcal{Y} : p_{\boldsymbol{\theta}}(y) > 0\}$ of $p_{\boldsymbol{\theta}}(y)$ is independent of $\boldsymbol{\theta}$.

P4) There exists a measurable function $q : \mathcal{Y} \times \Theta \to \mathbb{R}$ such that for all sufficiently small $\Delta > 0$, for all $i = 1, \ldots, K$, for all y, and for all $\boldsymbol{\theta}$,

$$\frac{1}{\Delta} \left| p_{\boldsymbol{\theta} + \Delta \mathbf{e}_i}(y) - p_{\boldsymbol{\theta}}(y) \right| \leq q(y, \boldsymbol{\theta}) \tag{28}$$

and such that for all $\boldsymbol{\theta}$,

$$\int q^2(y, \boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy) < \infty. \tag{29}$$

In (28), $\mathbf{e}_i$ represents the ith column of the $K \times K$ identity matrix.

P5) For each $\boldsymbol{\theta}$, the $K \times K$ Fisher information matrix (FIM)

$$\mathbf{J}_{\boldsymbol{\theta}} \triangleq \mathbb{E}\left\{ \left(\frac{\partial \log p_{\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \log p_{\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}} \right)^* \right\} \tag{30}$$

is finite and invertible.

P6) $h_{\boldsymbol{\theta}}$ is Fréchet differentiable with respect to $\boldsymbol{\theta}$, in the sense that for each $\boldsymbol{\theta}$, there exists a continuous linear operator $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta} : \mathbb{R}^K \to L_2$ such that

$$\lim_{\boldsymbol{\delta} \to 0} \frac{\| h_{\boldsymbol{\theta} + \boldsymbol{\delta}} - h_{\boldsymbol{\theta}} - \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \boldsymbol{\delta} \|_{\mathcal{H}}}{\|\boldsymbol{\delta}\|} = 0 \tag{31}$$

P7) The null space of the mapping $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ contains only the zero vector. This assumption is required to ensure that the mapping from $\boldsymbol{\theta}$ to x is non-redundant, in the sense that there does not exist a parametrization of $\mathfrak{X}$ in which the number of degrees of freedom is smaller than K.

We are now ready to state the CRB for the estimation of a parameter $x \in L_2[0, T]$ parameterized by a finite-dimensional vector $\boldsymbol{\theta}$. The proof of this theorem is given in Appendix A.

Theorem 1. *Let $\boldsymbol{\theta} \in \Theta$ be a deterministic parameter, where Θ is an open set in $\mathbb{R}^K$. Let $\mathcal{H}$ be a measurable, separable Hilbert space and let $x \in \mathcal{H}$ be a signal defined by the mapping $h_{\boldsymbol{\theta}} : \Theta \to \mathcal{H}$, as in* (5)*. Let $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be*

a family of probability measures over a measurable space $(\Omega, \mathscr{F})$, and let $y : \Omega \to \mathcal{Y}$ be a random variable, where $\mathcal{Y}$ is a measurable Hilbert space. Assume regularity conditions P1–P6. Let $\hat{x} : \mathcal{Y} \to \mathcal{H}$ be an unbiased estimator of x from the measurements y such that

$$\mathbb{E}\{\|\hat{x}(y)\|_{\mathcal{H}}^2\} < \infty. \tag{32}$$

Then, the MSE of $\hat{x}$ satisfies

$$\mathrm{MSE}(\hat{x}, x) \geq \mathrm{Tr}\left[\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \mathbf{J}_{\boldsymbol{\theta}}^{-1}\right] \tag{33}$$

where $\mathbf{J}_{\boldsymbol{\theta}}$ is the FIM (30).

Theorem 1 enables us to obtain a lower bound on the estimation error of x based on the FIM for estimating $\boldsymbol{\theta}$. The latter can often be computed relatively easily since $\boldsymbol{\theta}$ is a finite-dimensional vector. Even more conveniently, the trace on the right-hand side of (33) is taken over a $K \times K$ matrix, despite the involvement of continuous-time operators. Thus, the computation of (33) is often possible either analytically or numerically, a fact which will be used extensively in the sequel.

VI. CRB for Continuous-Time Measurements

We now apply Theorem 1 to the problem of estimating a deterministic signal x from continuous-time measurements y given by (6).

Theorem 2. *Let x be a deterministic function defined by* (5)*, where $\boldsymbol{\theta} \in \Theta$ is an unknown deterministic parameter and Θ is an open subset of $\mathbb{R}^K$. Let y be a measurement signal given by* (6)*. Suppose that Assumptions P6–P7 are satisfied. Then, the MSE of any unbiased, finite-variance estimator $\hat{x}$ of x from y is bounded by*

$$\mathrm{MSE}(\hat{x}, x) \geq K\sigma_c^2. \tag{34}$$

The bound of Theorem 2 can be translated to units of the rate of innovation ρ_{T_0} if we assume that the segment $[0, T_0]$ under analysis achieves the maximum (4), i.e., this is a segment containing the maximum allowed number of degrees of freedom. In this case $\rho_{T_0} = K/T_0$, and any unbiased estimator $\hat{x}(t)$ satisfies

$$\frac{\frac{1}{T_0}\mathbb{E}\left\{\int_0^{T_0} |x(t) - \hat{x}(t)|^2 dt\right\}}{\sigma_c^2} \geq \rho_{T_0}. \tag{35}$$

In the noisy setting, ρ_{T_0} loses its meaning as a lower bound on the sampling rate required for perfect recovery, since the latter is no longer possible at any sampling rate. On the other hand, it follows from (35) that the rate of innovation gains an alternative meaning; namely, ρ_{T_0} is a lower bound on the ratio between the average MSE achievable by any unbiased estimator and the noise variance σ_c^2, *regardless of the sampling method*.

Before formally proving Theorem 2, note that (34) has an intuitive geometric interpretation. Specifically, the constraint set $\mathfrak{X}$ is a K-dimensional differential manifold in $L_2[0, T]$. In other words, for any point $x \in \mathfrak{X}$, there exists a K-dimensional subspace $\mathcal{U}$ tangent to $\mathfrak{X}$ at x. We refer to $\mathcal{U}$ as the feasible direction subspace [39]: any perturbation of x which remains within the constraint set $\mathfrak{X}$ must be in one of the directions in $\mathcal{U}$. Formally, $\mathcal{U}$ can be defined as the range space of $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$.

If one wishes to use the measurements y to distinguish between x and its local neighborhood, then it suffices to observe the projection of y onto $\mathcal{U}$. Projecting the measurements onto $\mathcal{U}$ removes most of the noise, retaining only K independent Gaussian components, each having a variance of σ_c^2. Thus we have obtained an intuitive explanation for the bound of $K\sigma_c^2$ in Theorem 2. To formally prove this result, we apply Theorem 1 to the present setting, as follows.

Proof of Theorem 2: The problem of estimating the parameters $\boldsymbol{\theta}$ from a continuous-time signal $y(t)$ of the form (6) was examined in [15, Example I.7.3], where the validity of Assumptions P1–P4 was demonstrated. It was further shown that the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ for estimating $\boldsymbol{\theta}$ from $y(t)$ is given by [15, *ibid.*]

$$\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}} = \frac{1}{\sigma_c^2}\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right). \tag{36}$$

Our goal will be to use (36) and Theorem 1 to obtain a bound on estimators of the continuous-time function $x(t)$. To this end, observe that the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ is finite since, by Assumption P6, the operator $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ is a bounded operator into L_2. Furthermore, by Assumption P7, $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ has a trivial null space, and thus $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ is invertible. Therefore, Assumption P5 has been demonstrated. We may consequently apply Theorem 1, which yields

$$\begin{aligned}
&\mathrm{MSE}(\hat{x}, x) \\
&\quad \geq \sigma_c^2 \,\mathrm{Tr}\left[\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \left(\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)\right)^{-1}\right] \\
&\quad = \sigma_c^2 \,\mathrm{Tr}\left(\mathbf{I}_K\right) \\
&\quad = K\sigma_c^2
\end{aligned} \tag{37}$$

thus completing the proof. ∎

To illustrate the use of Theorem 2 in practice, let us consider as a simple example a signal $x(t)$ belonging to a finite-dimensional subspace $\mathcal{G}$. Specifically, assume that

$$x(t) = \sum_{k=1}^{K} a_k g_k(t) \tag{38}$$

for some coefficient vector $\boldsymbol{\theta} = (a_1, \ldots, a_K)^T$ and a given set of linearly independent functions $\{g_k\}$ spanning $\mathcal{G}$. This includes, for example, families of shift-invariant subspaces with a compactly supported generator (see Section III-A). From Theorem 2, the MSE of any unbiased estimator of x is bounded by $K\sigma_c^2$, where K is the dimension of the subspace $\mathcal{G}$. We now demonstrate that this bound is achieved by the unbiased estimator

$$\hat{x} = \mathbf{P}_{\mathcal{G}}\, y \tag{39}$$

where $\mathbf{P}_{\mathcal{G}}$ is the orthogonal projector onto the subspace $\mathcal{G}$.

To verify that (39) achieves the CRB, let G denote the set transformation (2) associated with the functions $\{g_k\}_{k=1}^K$. One may then write $x = G\boldsymbol{\theta}$ and $\mathbf{P}_{\mathcal{G}} = G(G^*G)^{-1}G^*$. Thus (39)

becomes

$$\begin{aligned}\hat{x} &= G(G^*G)^{-1}G^*G\boldsymbol{\theta} + G(G^*G)^{-1}G^*w \\ &= G\boldsymbol{\theta} + \mathbf{P}_{\mathcal{G}}\, w\end{aligned} \tag{40}$$

and therefore

$$\mathbb{E}\big\{\|\hat{x}-x\|_{L_2}^2\big\} = \mathbb{E}\big\{\|\mathbf{P}_{\mathcal{G}}\, w\|_{L_2}^2\big\}. \tag{41}$$

Since $\mathcal{G}$ is a K-dimensional subspace, it is spanned by a set of K orthonormal[1] functions $u_1,\ldots,u_K \in L_2$. Thus

$$\mathbb{E}\big\{\|\mathbf{P}_{\mathcal{G}}\, w\|_{L_2}^2\big\} = \sum_{k=1}^{K}\mathbb{E}\Big\{|\langle w,u_k\rangle|^2\Big\} = K\sigma_c^2 \tag{42}$$

which demonstrates that $\hat{x}$ indeed achieves the CRB in this case.

In practice, a signal is not usually estimated directly from its continuous-time measurements. Rather, the signal $y(t)$ is typically sampled and digitally manipulated. In the next section, we will compare the results of Theorem 2 with the performance achievable from sampled measurements, and demonstrate that in some cases, a finite-rate sampling scheme is sufficient to achieve the continuous-time bound of Theorem 2.

VII. CRB for Sampled Measurements

In this section, we consider the problem of estimating $x(t)$ of (5) from a finite number of samples of the process $y(t)$ given by (6). Specifically, suppose our measurements are given by

$$c_n = \langle y, s_n\rangle + v_n = \int_0^{T_0} y(t)s_n^*(t)dt + v_n, \quad n=1,\ldots,N \tag{43}$$

where $\{s_n\}_{n=1}^N \in L_2[0,T_0]$ are sampling kernels, and v_n is a discrete white Gaussian noise process, independent of $w(t)$, having mean zero and variance σ_d^2. Note that the model (43) includes both continuous-time noise, which is present in the signal $y(t)=x(t)+w(t)$ prior to sampling, and digital noise v_n, which arises from the sampling process, e.g., as a result of quantization. In this section, we will separately examine the effect of each of these noise components.

From (6) and (43), it can be seen that the measurements $c_1,\ldots,c_N$ are jointly Gaussian with mean

$$\mu_n \triangleq \mathbb{E}\{c_n\} = \langle x, s_n\rangle \tag{44}$$

and covariance

$$\Gamma_{ij} \triangleq \mathrm{Cov}(c_i,c_j) = \sigma_c^2\langle s_i,s_j\rangle + \sigma_d^2\delta_{ij}. \tag{45}$$

A somewhat unusual aspect of this estimation setting is that the choice of the sampling kernels $s_n(t)$ affects not only the measurements obtained, but also the statistics of the noise. One example of the impact of this fact is the following. Suppose first that no digital noise is present, i.e., $\sigma_d=0$, and consider a modified set of sampling kernels $\{\tilde{s}_n(t)\}_{n=1}^N$ which are an invertible linear transformation of $\{s_n(t)\}_{n=1}^N$, so that

$$\tilde{s}_n(t) = \sum_{i=1}^{N}\mathbf{B}_{ni}s_i(t) \tag{46}$$

where $\mathbf{B}\in\mathbb{R}^{N\times N}$ is an invertible matrix. Then, the resulting measurements $\tilde{\mathbf{c}}$ are given by $\tilde{\mathbf{c}}=\mathbf{B}\mathbf{c}$, and similarly the original measurements $\mathbf{c}$ can be recovered from $\tilde{\mathbf{c}}$. It follows that these settings are equivalent in terms of the accuracy with which x can be estimated. In particular, the FIM for estimating x in the two settings is identical [15, Th. I.7.2].

When digital noise is present in addition to continuous-time noise, the sampling schemes $\{s_n(t)\}$ and $\{\tilde{s}_n(t)\}$ are no longer necessarily equivalent, since the gain introduced by the transformation $\mathbf{B}$ will alter the ratio between the energy of the signal and the digital noise. The two estimation problems are then equivalent if and only if $\mathbf{B}$ is a unitary transformation.

How should one choose the space $\mathcal{S}=\mathrm{span}\{s_1,\ldots,s_N\}$ spanned by the sampling kernels? Suppose for a moment that there exist elements in the range space of $\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}$ which are orthogonal to $\mathcal{S}$. This implies that one can perturb x in such a way that the constraint set $\mathfrak{X}$ is not violated, without changing the distribution of the measurements $\mathbf{c}$. This situation occurs, for example, when the number of measurements N is smaller than the dimension K of the parametrization of $\mathfrak{X}$. While it may still be possible to reconstruct some of the information concerning x from these measurements, this is an undesirable situation from an estimation point of view. Thus we will assume henceforth that

$$\mathcal{R}\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\right)\cap\mathcal{S}^{\perp} = \{\mathbf{0}\}. \tag{47}$$

As an example of the necessity of the condition (47), consider again the signal (38), which belongs to a K-dimensional subspace $\mathcal{G}\subset L_2$ spanned by the functions $g_1,\ldots,g_K$. In this case it is readily seen that for any vector $\mathbf{v}$

$$\frac{\partial h_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\mathbf{v} = \sum_{k=1}^{K}\mathbf{v}_k g_k(t). \tag{48}$$

Since the functions $\{g_k\}$ span the space $\mathcal{G}$, this implies that $\mathcal{R}(\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}) = \mathcal{G}$, and therefore the condition (47) can be written as

$$\mathcal{G}\cap\mathcal{S}^{\perp} = \{\mathbf{0}\} \tag{49}$$

which is a standard requirement in the design of a sampling system for signals belonging to a subspace $\mathcal{G}$ [40].

By virtue of Theorem 1, a lower bound on unbiased estimation of x can be obtained by first computing the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ for estimating $\boldsymbol{\theta}$ from $\mathbf{c}$. This yields the following result. For simplicity of notation, in this theorem we assume that the function $h_{\boldsymbol{\theta}}$ and the sampling kernels s_n are real. If complex sampling kernels are desired (as will be required in the sequel), the result below can still be used by translating each measurement to an equivalent pair of real-valued samples.

Theorem 3. *Let x be a deterministic real function defined by* (5)*, where $\boldsymbol{\theta}\in\Theta$ is an unknown deterministic parameter and Θ is an open subset of $\mathbb{R}^K$. Assume regularity conditions*

[1] We require the new functions $u_1,\ldots,u_K$ since the functions $g_1,\ldots,g_K$ are not necessarily orthonormal. The choice of non-orthonormal functions $g_1,\ldots,g_K$ will prove useful in the sequel.

P6–P7, and let $\hat{x}$ be an unbiased estimator of x from the real measurements $\mathbf{c} = (c_1, \ldots, c_N)^T$ of (43)*. Then, the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ for estimating $\boldsymbol{\theta}$ from $\mathbf{c}$ is given by*

$$\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}} = \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* S\left(\sigma_c^2 S^* S + \sigma_d^2 \mathbf{I}_N\right)^{-1} S^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \quad (50)$$

where S is the set transformation corresponding to the functions $\{s_n\}_{n=1}^N$. If (47) *holds, then $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ is invertible. In this case, any finite-variance, unbiased estimator $\hat{x}$ for estimating x from $\mathbf{c}$ satisfies*

$$\mathrm{MSE}(\hat{x}, x) \geq \mathrm{Tr}\left[\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) (\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}})^{-1}\right]. \quad (51)$$

Proof: In the present setting, the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ is given by [35]

$$\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)^* \boldsymbol{\Gamma}^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right) \quad (52)$$

where the matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{N \times N}$ is defined by (45) and the matrix $\partial \boldsymbol{\mu}/\partial \boldsymbol{\theta} \in \mathbb{R}^{N \times K}$ is given by

$$\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right)_{nk} = \frac{\partial \mu_n}{\partial \theta_k} \quad (53)$$

with μ_n defined in (44).

By the definition of the set transformation, the ijth element of the $N \times N$ matrix S^*S is given by

$$(S^*S)_{ij} = \langle S\mathbf{e}_j, S\mathbf{e}_i \rangle = \langle s_j, s_i \rangle \quad (54)$$

where $\mathbf{e}_i$ is the ith column of the $N \times N$ identity matrix. Therefore, we have

$$\boldsymbol{\Gamma} = \sigma_c^2 S^* S + \sigma_d^2 \mathbf{I}_N. \quad (55)$$

Similarly, observe that

$$\left(S^* \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)_{nk} = \left\langle \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \tilde{\mathbf{e}}_k, S\mathbf{e}_n \right\rangle = \left\langle \frac{\partial h_{\boldsymbol{\theta}}}{\partial \theta_k}, s_n \right\rangle = \frac{\partial \mu_n}{\partial \theta_k} \quad (56)$$

where $\tilde{\mathbf{e}}_k$ is the kth column of the $K \times K$ identity matrix. Thus

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = S^* \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}. \quad (57)$$

Substituting (55) and (57) into (52) yields the required expression (50).

We next demonstrate that if (47) holds, then $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ is invertible. To see this, note that from (50) we have

$$\mathcal{N}(\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}) = \mathcal{N}\left(S^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)\right). \quad (58)$$

Now, consider an arbitrary function $f \in \mathcal{R}(\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta})$. If (47) holds, then f is not orthogonal to the subspace $\mathcal{S}$. Therefore, $\langle f, s_n \rangle \neq 0$ for at least one value of n, and thus by (3), $S^* f \neq 0$. This implies that

$$\mathcal{N}\left(S^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)\right) = \mathcal{N}\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) = \{\mathbf{0}\}. \quad (59)$$

Combined with (58), we conclude that $\mathcal{N}(\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}) = \{\mathbf{0}\}$. This demonstrates that $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ is invertible, proving Assumption P5. Moreover, in the present setting, Assumptions P1–P4 are fulfilled for any value of $\boldsymbol{\theta}_0$ [33]. Applying Theorem 1 yields (51) and completes the proof. ■

In the following subsections we draw several conclusions from Theorem 3.

A. Discrete-Time Noise

Suppose first that $\sigma_c^2 = 0$, so that only digital noise is present. This setting has been analyzed previously [12], [13], [41], and we therefore only briefly examine the contrast with continuous-time noise. When only digital noise is present, its effects can be surmounted either by increasing the gain of the sampling kernels, or by increasing the number of measurements. These intuitive conclusions can be verified from Theorem 3 as follows. Assume that condition (47) holds, and consider the modified kernels $\tilde{s}_n(t) = 2s_n(t)$. The set transformation $\tilde{S}$ corresponding to the modified kernels is $\tilde{S} = 2S$, and since $\sigma_c^2 = 0$, this implies that the FIM obtained from the modified kernels is given by $\tilde{\mathbf{J}}_{\boldsymbol{\theta}}^{\mathrm{samp}} = 4\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$. Thus, a sufficient increase in the sampling gain can arbitrarily increase $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ and consequently reduce the bound (51) arbitrarily close to zero. Similarly, it is possible to increase the number of samples, for example, by repeating each measurement twice. Let S and $\tilde{S}$ denote the transformations corresponding to the original and doubled sets of measurements. It can then readily be seen from the definition of the set transformation (2) and its adjoint (3) that $\tilde{S}\tilde{S}^* = 4SS^*$. Consequently, by the same argument, in the absence of continuous-time noise one can achieve arbitrarily low error by repeated measurements.

In practice, rather than repeating each measurement, an increase in sampling rate is often obtained by sampling on a denser grid. In this case, the analysis is more complex and depends on the specific signal family in question. For example, in the case of SI spaces, it has been shown that the error converges to zero as the sampling rate increases [41], [42].

B. Continuous-Time Noise

As we have seen, sampling noise can be mitigated by increasing the sampling rate. Furthermore, digital noise is inherently dependent on the sampling scheme being used. Since our goal is to determine the fundamental performance limits regardless of the sampling technique, we will focus here and in subsequent sections on continuous-time noise. Thus, suppose that $\sigma_d^2 = 0$, so that only continuous-time noise is present. In this case, as we now show, it is generally impossible to achieve arbitrarily low reconstruction error, regardless of the sampling kernels used; indeed, it is never possible to outperform the continuous-time CRB of Section VI, which is typically nonzero. To see this formally, observe first that in the absence of digital noise, the FIM for estimating $\boldsymbol{\theta}$ can be written as

$$\begin{aligned}\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}} &= \frac{1}{\sigma_c^2}\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* S\left(S^*S\right)^{-1} S^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \\ &= \frac{1}{\sigma_c^2}\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \mathbf{P}_{\mathcal{S}} \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \end{aligned} \quad (60)$$

where $\mathbf{P}_{\mathcal{S}}$ is the orthogonal projection onto the subspace $\mathcal{S}$. It is insightful to compare this expression with the FIM

$\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ obtained from continuous-time measurements in (36). In both cases, a lower bound on the MSE for unbiased estimation of x was obtained from $\mathbf{J}_{\boldsymbol{\theta}}$ by applying Theorem 1. Consequently, if it happens that $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}} = \mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$, then the continuous-time bound of Theorem 2 and the sampled bound of Theorem 3 coincide. Thus, if no digital noise is added, then it is possible (at least in terms of the performance bounds) that estimators based on the samples $\mathbf{c}$ will suffer no degradation compared with the "ideal" estimator based on the entire set of continuous-time measurements. This occurs if and only if $\mathcal{R}(\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}) \subseteq \mathcal{S}$; in this case, the projection $\mathbf{P}_{\mathcal{S}}$ will have no effect on the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$, which will then coincide with $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ of (36). In the remainder of this section, we will discuss several cases in which this fortunate circumstance arises.

C. Example: Sampling in a Subspace

The simplest situation in which samples provide all of the information present in the continuous-time signal is the case in which $x(t)$ belongs to a K-dimensional subspace $\mathcal{G}$ of L_2. This is the case, for example, when the signal lies in a shift-invariant subspace having a compactly supported generator (see Section III-A). As we have seen above (cf. (48)), in this scenario $\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}$ is a mapping onto the subspace $\mathcal{G}$. Assuming that there is no discrete-time noise, it follows from (60) that the optimal choice of a sampling space $\mathcal{S}$ is $\mathcal{G}$ itself. Such a choice requires $N = K$ samples and yields $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}} = \mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$. Of course, such an occurrence is not possible if the sampling process contributes additional noise to the measurements.

In some cases, it may be difficult to implement a set of sampling kernels spanning the subspace $\mathcal{G}$. It may then be desirable to choose a K-dimensional subspace $\mathcal{S}$ which is close to $\mathcal{G}$ but does not equal it. We will now compute the CRB for this setting and demonstrate that it can be achieved by a practical estimation technique. This will also demonstrate achievability of the CRB in the special case $\mathcal{S} = \mathcal{G}$. We first note from (2) and (48) that $\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta} = G$, where G is the set transformation corresponding to the generators $\{g_k\}_{k=1}^K$. Furthermore, it follows from (49) that S^*G and G^*S are invertible $K \times K$ matrices [40]. Using Theorem 3, we thus find that the CRB is given by

$$\begin{aligned}\mathrm{MSE}(\hat{x},x) &\geq \sigma_c^2 \operatorname{Tr}\Big(G\left(G^*S(S^*S)^{-1}S^*G\right)^{-1}G^*\Big)\\ &= \sigma_c^2 \operatorname{Tr}\big(G(S^*G)^{-1}S^*S(G^*S)^{-1}G^*\big). \end{aligned} \tag{61}$$

It is readily seen that when $\mathcal{S} = \mathcal{G}$, the bound (61) reduces to $K\sigma_c^2$, which is (as expected) the continuous-time bound of Theorem 2. When $\mathcal{S} \neq \mathcal{G}$, the bound (61) will generally be higher than $K\sigma_c^2$, since $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ of (60) will exceed $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}}$ of (36). In this case, it is common to use the consistent, unbiased estimator [18], [40]

$$\hat{x} = G(S^*G)^{-1}\mathbf{c}. \tag{62}$$

As we now show, the bound (61) is achieved by this estimator. Indeed, observe that $\mathbf{c} = S^*y = S^*G\boldsymbol{\theta} + S^*w$, and thus

$$\begin{aligned}\mathbb{E}\big\{\|\hat{x}-x\|_{L_2}^2\big\} &= \mathbb{E}\big\{\|G(S^*G)^{-1}S^*w\|_{L_2}^2\big\}\\ &= \mathbb{E}\big\{\operatorname{Tr}\big(G(S^*G)^{-1}S^*ww^*S(G^*S)^{-1}G^*\big)\big\}\\ &= \operatorname{Tr}\big(G(S^*G)^{-1}\operatorname{Cov}(S^*w)(G^*S)^{-1}G^*\big).\end{aligned} \tag{63}$$

Note that $\operatorname{Cov}(S^*w) = \operatorname{Cov}(\mathbf{c})$, which by (55) is equal to $\sigma_c^2 S^*S$. Substituting this result into (63) and comparing with (61) verifies that $\hat{x}$ achieves the CRB.

D. Nyquist-Equivalent Sampling

We refer to situations in which the dimension of the sampling space equals the dimension of the signal space as "Nyquist-equivalent" sampling schemes. In the previous section, we saw that Nyquist-equivalent sampling is possible using K samples when the signal lies in a K-dimensional subspace $\mathfrak{X}$, and that the resulting system achieves the continuous-time CRB. A similar situation occurs when the set of possible signals $\mathfrak{X}$ is a subset of an M-dimensional subspace $\mathcal{M}$ of L_2 with $M > K$. In this case, it can be readily shown that $\mathcal{R}(\partial h_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}) \subseteq \mathcal{M}$. Thus, by choosing $N = M$ sampling kernels such that $\mathcal{S} = \mathcal{M}$, we again achieve $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{cont}} = \mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$, demonstrating that all of the information content in x has been captured by the samples. This is again a Nyquist-equivalent scheme, but the number of samples it requires is higher than the number of parameters K defining the signals. Therefore, in this case it is not possible to sample at the rate of innovation without losing some of the information content of the signal.

In general, the constraint set $\mathfrak{X}$ will not be contained in any finite-dimensional subspace of L_2. In such cases, it will generally not be possible to achieve the performance of the continuous-time bound using *any* finite number of samples, even in the absence of digital noise. This implies that in the most general setting, sampling above the rate of innovation can often improve the performance of estimation schemes. This conclusion will be verified by simulation in Section IX.

VIII. Optimal Sampling for Linear Reconstruction Procedures

In this section, we address the problem of designing a sampling method which minimizes the MSE. One route towards this goal could be to minimize the sampled CRB of Theorem 3 with respect to the sampling space $\mathcal{S}$. However, the CRB is a function of the unknown parameter vector $\boldsymbol{\theta}$. Consequently, for each value of $\boldsymbol{\theta}$, there may be a different sampling space $\mathcal{S}$ which minimizes the bound. To obtain a sampling method which is optimal on average over all possible choices of $\boldsymbol{\theta}$, we now make the additional assumption that the parameter vector $\boldsymbol{\theta}$ is random and has a known distribution. Our goal, then, is to determine the sampling space $\mathcal{S}$ that minimizes the MSE $\mathbb{E}\big\{\|\hat{x}-x\|_{L_2}^2\big\}$ within a class of allowed estimators. Note that the mean is now taken over realizations of both the noise $w(t)$ and the parameter $\boldsymbol{\theta}$.

Since $\boldsymbol{\theta}$ is random, the signal $x(t)$ is random as well. To make our discussion general, we will derive the optimal sampling functions for estimating a general complex random

process $x(t)$ (not necessarily having realizations in $\mathfrak{X}$ of (5)) from samples of the noisy process $y(t) = x(t) + w(t)$. In Sections VIII-C and IX, we will specialize the results to a specific type of real-valued FRI signal and obtain explicit expressions for the optimal sampling kernels in these scenarios.

Let $x(t)$ denote a zero-mean random process defined over $t \in [0, T_0]$, and suppose that its autocorrelation function

$$R_X(t, \eta) \triangleq \mathbb{E}\{x(t)x^*(\eta)\} \tag{64}$$

is continuous in t and η. Our goal is to estimate $x(t)$ based on a finite number N of samples of the signal $y(t) = x(t) + w(t)$, $t \in [0, T_0]$, where $w(t)$ is a white noise process (not necessarily Gaussian) with variance σ_c^2 which is uncorrelated with $x(t)$. We focus our attention on *linear sampling* schemes, i.e., we assume the samples are given by

$$c_n = \langle y, s_n \rangle . \tag{65}$$

Finally, we restrict the discussion to *linear estimation* methods, namely those techniques in which the estimate $\hat{x}(t)$ is constructed as

$$\hat{x}(t) = \sum_{n=1}^{N} c_n v_n(t), \tag{66}$$

for some set of reconstruction functions $\{v_n(t)\}_{n=1}^N$. It is important to note that for any given set of sampling functions $\{s_n(t)\}_{n=1}^N$, the minimum MSE (MMSE) estimator of $x(t)$ is often a nonlinear function of the measurements $\{c_n\}_{n=1}^N$. Indeed, typical FRI reconstruction techniques involve a nonlinear stage. Consequently, restricting the discussion to linear recovery schemes may seem inadequate. However, this choice has two advantages. First, as we will see, the optimal linear scheme is determined only by the second-order statistics of $x(t)$ and $w(t)$, whereas the analysis of nonlinear methods necessitates exact knowledge of their entire distribution functions. Second, it is not the final estimate $\hat{x}(t)$ that interests us in this discussion, but merely the set of optimal sampling functions. Once such a set is determined, albeit from a linear recovery perspective, it can be used in conjunction with existing nonlinear FRI techniques, though of course the optimality guarantees will no longer hold in this case. As we will see in Section IX, the conclusions obtained through our analysis appear to apply to FRI techniques in general. Under the above assumptions, our goal is to design the sampling kernels $\{s_n(t)\}_{n=1}^N$ and reconstruction functions $\{v_n(t)\}_{n=1}^N$ such that the MSE (8) is minimized.

As can be seen from (65), we assume henceforth that only continuous-time noise is present in the sampling system. The situation is considerably more complicated in the presence of digital noise. First, without digital noise, one must choose only the subspace spanned by the sampling kernels, as the kernels themselves do not affect the performance; this is no longer the case when digital noise is added. Second, digital noise may give rise to a requirement that a particular measurement be repeated in order to average out the noise. This is undesirable in the continuous noise regime, since the repeated measurement will contain the exact same noise realization.

A. Relation to the Karhunen–Loève Expansion and Finite-Dimensional Generalizations

The problem posed above is closely related to the Karhunen–Loève transform (KLT) [43], [44], which is concerned with the reconstruction of a random signal $x(t)$ from its noiseless samples. Specifically, one may express $x(t)$ in terms of a complete orthonormal basis $\{\psi_k(t)\}_{k=1}^\infty$ for L_2 as

$$x(t) = \sum_{k=1}^{\infty} \langle x, \psi_k \rangle \, \psi_k(t). \tag{67}$$

The goal of the KLT is to choose the functions $\{\psi_k(t)\}_{k=1}^\infty$ such that the MSE resulting from the truncation of this series after N terms is minimal. It is well known that the solution to this problem is given by the N-term truncation of the Karhunen–Loève expansion [43], [45].

Since $R_X(t, \eta)$ is assumed to be continuous in our setting, by Mercer's theorem [45] it possesses a discrete set of eigenfunctions $\{\psi_k(t)\}_{k=1}^\infty$, which constitute an orthonormal basis for L_2. These functions satisfy the equations

$$\lambda_k \psi_k(t) = \int_0^{T_0} R_X(t, \eta)\psi_k(\eta) d\eta, \tag{68}$$

in which the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are nonnegative and are assumed to be arranged in descending order. With these functions, (67) is known as the Karhunen–Loève expansion. It can be easily shown that the first N terms in this series constitute the best N-term approximation of $x(t)$ in an MSE sense [45]. In other words, in the noiseless case, the optimal sampling and reconstruction functions are $s_n(t) = v_n(t) = \psi_n(t)$.

In our setting, we do not have access to samples of $x(t)$ but rather only to samples of the noisy process $y(t)$. In this case, it is not clear *a priori* whether the optimal sampling and reconstruction filters coincide or whether they match the Karhunen–Loève expansion of $x(t)$.

The finite-dimensional analogue of our problem, in which x, y, and w are random vectors taking values in $\mathbb{C}^M$, was treated in [46], [47]. The derivation in these works, however, relied on the low-rank approximation property of the singular-value decomposition (SVD) of a matrix. The generalization of this concept to infinite-dimensional operators is subtle and will thus be avoided here. Instead, we provide a conceptually simple (if slightly cumbersome) derivation of the optimal linear sampling and reconstruction method for noisy signals. As we will see, it still holds that $s_n(t) = \psi_n(t)$, but $v_n(t) = \alpha_n \psi_n(t)$, where α_n is a shrinkage factor depending on the SNR of the nth sample.

B. Optimal Sampling in Noisy Settings

As explained in Section VII, in the absence of discrete-time noise, the MSE is not affected by modifications of the sampling kernels which leave the set $\mathcal{S} = \operatorname{span}\{s_1(t), \ldots, s_N(t)\}$ unchanged. Thus, without loss of generality, we constrain $\{s_n(t)\}_{n=1}^N$ to satisfy

$$\left\langle s_n \, , \, \sigma_c^2 s_m + \int_0^T R_X(\cdot, \tau) s_m(\tau) d\tau \right\rangle = \delta_{m,n} \tag{69}$$

for every $m,n=1,\ldots,N$. This can always be done since the operator $R_Y : L_2 \to L_2$ defined by $(R_Y f)(t) = \int_0^T R_X(t,\tau)f(\tau)d\tau + \sigma_c^2 f(t)$ is positive definite. This choice is particularly convenient as it results in a set of uncorrelated samples $\{c_n\}$. Indeed

$$\begin{aligned}\mathbb{E}\{c_m c_n^*\} &= \mathbb{E}\left\{\left(\int_0^T s_m^*(\tau)y(\tau)d\tau\right)\left(\int_0^T s_n^*(\eta)y(\eta)d\eta\right)^*\right\}\\ &= \iint_0^T s_m^*(\tau)\mathbb{E}\{y(\tau)y^*(\eta)\}\,s_n(\eta)d\tau d\eta\\ &= \iint_0^T s_m^*(\tau)R_X(\tau,\eta)s_n(\eta)d\tau d\eta\\ &\qquad + \mathbb{E}\left\{\langle s_n,w\rangle\langle s_m,w\rangle^*\right\}\\ &= \iint_0^T s_m^*(\tau)R_X^*(\eta,\tau)s_n(\eta)d\tau d\eta + \sigma_c^2\langle s_n,s_m\rangle\\ &= \delta_{m,n}. \end{aligned} \tag{70}$$

We are now ready to determine the optimal sampling method. We begin by expressing the MSE (8) as

$$\begin{aligned}\int_0^T \mathbb{E}\left\{|x(t)-\hat{x}(t)|^2\right\}dt &= \int_0^T \mathbb{E}\left\{|x(t)|^2\right\}dt\\ &\quad - 2\int_0^T \Re\left\{\mathbb{E}\{x^*(t)\hat{x}(t)\}\right\}dt + \int_0^T \mathbb{E}\left\{|\hat{x}(t)|^2\right\}dt.\end{aligned} \tag{71}$$

The first term in this expression does not depend on the choice of $\{s_n(t)\}_{n=1}^N$ and $\{v_n(t)\}_{n=1}^N$, and is therefore irrelevant for our purpose. Substituting (66) and (65), and using the fact that $w(t)$ is uncorrelated with $x(t)$, the second term can be written as

$$\begin{aligned}&\int_0^T 2\Re\left\{\mathbb{E}\left\{x^*(t)\sum_{n=1}^N c_n v_n(t)\right\}\right\}dt\\ &= \sum_{n=1}^N 2\int_0^T \Re\left\{\mathbb{E}\left\{x^*(t)\int_0^T y(\tau)s_n^*(\tau)d\tau\right\}v_n(t)\right\}dt\\ &= \sum_{n=1}^N 2\iint_0^T \Re\left\{\mathbb{E}\{x^*(t)(x(\tau)+w(\tau))\}\,s_n^*(\tau)v_n(t)\right\}d\tau dt\\ &= \sum_{n=1}^N 2\iint_0^T \Re\left\{s_n^*(\tau)R_X(\tau,t)v_n(t)\right\}d\tau dt\\ &= \sum_{n=1}^N 2\Re\left\{\left\langle v_n\,,\int_0^T R_X(\cdot,\tau)s_n(\tau)d\tau\right\rangle\right\}.\end{aligned} \tag{72}$$

Similarly, using the fact that $\{c_n\}_{n=1}^N$ are uncorrelated and have unit variance (see (70)), the last term in (71) becomes

$$\begin{aligned}\int_0^T \mathbb{E}\left\{\left|\sum_{n=1}^N c_n v_n(t)\right|^2\right\}dt &= \sum_{m=1}^N\sum_{n=1}^N \mathbb{E}\{c_m^* c_n\}\langle v_m,v_n\rangle\\ &= \sum_{n=1}^N \|v_n\|^2.\end{aligned} \tag{73}$$

Substituting (72) and (73) back into (71), we conclude that minimization of the MSE is equivalent to minimization of

$$\sum_{n=1}^N\left(\|v_n\|^2 - 2\Re\left\{\left\langle v_n,\int_0^T R_X(\cdot,\tau)s_n(\tau)d\tau\right\rangle\right\}\right) \tag{74}$$

with respect to $\{s_n(t)\}_{n=1}^N$ and $\{v_n(t)\}_{n=1}^N$, subject to the set of constraints (69).

As a first stage, we minimize (74) with respect to the reconstruction functions $\{v_n(t)\}_{n=1}^N$. To this end, we note that the nth summand in (74) is lower bounded by

$$\begin{aligned}\|v_n\|^2 - 2\|v_n\|\left\|\int_0^T R_X(\cdot,\tau)s_n(\tau)d\tau\right\|&\\ \geq -\left\|\int_0^T R_X(\cdot,\tau)s_n(\tau)d\tau\right\|^2&,\end{aligned} \tag{75}$$

where we used the Cauchy–Schwarz inequality and the fact that $\min_z\{z^2-2bz\} = -b^2$. This bound is achieved by choosing

$$v_n(t) = \int_0^T R_X(t,\tau)s_n(\tau)d\tau, \tag{76}$$

thus identifying the optimal reconstruction functions.

Substituting (76) into (74), our goal becomes the maximization of

$$\sum_{n=1}^N\left\|\int_0^T R_X(\cdot,\tau)s_n(\tau)d\tau\right\|^2 \tag{77}$$

with respect to the sampling functions $\{s_n(t)\}_{n=1}^N$. As we show in Appendix B, the maximum of this expression is achieved by any set of kernels of the form

$$s_n(t) = \sum_{k=1}^N \mathbf{A}_{k,n}\left(\lambda_k+\sigma_c^2\right)^{-\frac{1}{2}}\psi_k(t), \tag{78}$$

where $\mathbf{A}$ is a unitary $N\times N$ matrix and λ_k and $\psi_k(t)$ are the eigenvalues and eigenfunctions of $R_x(t,\eta)$ respectively (see (68)). In particular, we can choose $\mathbf{A}=\mathbf{I}_N$, leading to

$$s_n(t) = \frac{1}{\sqrt{\lambda_n+\sigma_c^2}}\psi_n(t), \quad n=1,\ldots,N. \tag{79}$$

From (76), the optimal reconstruction kernels are given by

$$v_n(t) = \frac{\lambda_n}{\sqrt{\lambda_n+\sigma_c^2}}\psi_n(t), \quad n=1,\ldots,N. \tag{80}$$

The following theorem summarizes the result.

Theorem 4. *Let $x(t)$, $t\in[0,T]$ be a random process whose autocorrelation function $R_X(t,\eta)$ is jointly continuous in t and η. Assume that $y(t)=x(t)+w(t)$, where $w(t)$ is a white noise process uncorrelated with $x(t)$. Then, among all estimates $\hat{x}(t)$ of $x(t)$ having the form*

$$\hat{x}(t) = \sum_{n=1}^N v_n(t)\int_0^T s_n^*(\tau)y(\tau)dt \tag{81}$$

the MSE (8) *is minimized with $\{s_n(t)\}_{n=1}^N$ and $\{v_n(t)\}_{n=1}^N$ of* (79) *and* (80) *respectively. In these expressions, λ_n and $\psi_n(t)$*

are the eigenvalues and eigenfunctions of $R_x(t,\eta)$ respectively (see (68)*).*

Interestingly, the optimal sampling and reconstruction functions in our noisy setting are similar to those dictated by the KLT. The only difference is that in the present scenario, the nth sample is shrunk by a factor of $\lambda_n/(\lambda_n+\sigma_c^2)$ prior to reconstruction. This ensures that the low-SNR measurements do not contribute to the recovery as much as their high-SNR counterparts. From the viewpoint of designing the sampling mechanism, however, this difference is of no importance.

As stated above, in practice one would generally favor nonlinear processing of the samples (namely, applying standard nonlinear FRI techniques) rather than a simple element-wise shrinkage. Thus, the importance of Theorem 4 for our purposes is in identifying that the eigenfunctions of $R_X(t,\tau)$ remain the optimal sampling kernels even in the noisy setting.

C. Example: Sampling in a Subspace

To demonstrate the utility of Theorem 4, we now revisit the situation in which $x(t)$ is given by (38) for some set of linearly independent functions $\{g_k(t)\}_{k=1}^K$ spanning a subspace $\mathcal{G}\in L_2$. We assume that the coefficients $\boldsymbol{\theta}=\{a_1,\ldots,a_K\}^T$ form a zero-mean random vector and denote its autocorrelation matrix by $\mathbf{R}_{\boldsymbol{\theta}}$. In this case, the signal's autocorrelation function is given by

$$\begin{aligned} R_X(t,\eta) &= \mathbb{E}\{x(t)x^*(\eta)\} \\ &= \mathbb{E}\left\{\sum_{k=1}^K a_k g_k(t)\sum_{\ell=1}^K a_\ell^* g_k^*(\eta)\right\} \\ &= \sum_{k=1}^K\sum_{\ell=1}^K g_k(t)g_\ell^*(\eta)(\mathbf{R}_{\boldsymbol{\theta}})_{k,\ell}. \end{aligned} \tag{82}$$

Consequently, the operator $R_X: L_2\to L_2$ defined by $(R_x h)(t)=\int_0^T R_x(t,\eta)h(\eta)d\eta$ can be expressed as

$$R_x = G\mathbf{R}_{\boldsymbol{\theta}}G^*, \tag{83}$$

where G is the set transformation (2) associated with $\{g_k\}_{k=1}^K$.

Now, let $\mathbf{U}$ be a unitary matrix and let $\mathbf{D}$ be a diagonal matrix, such that

$$\mathbf{U}\mathbf{D}\mathbf{U}^* = (G^*G)^{1/2}\mathbf{R}_{\boldsymbol{\theta}}(G^*G)^{1/2}. \tag{84}$$

Since the dimension of $\mathcal{R}(G)$ is K, the operator R_X has at most K nonzero eigenvalues $\{\lambda_k\}_{k=1}^K$. Let Ψ denote the set transformation associated with the N eigenfunctions $\{\psi_n\}_{n=1}^N$ corresponding to the N largest eigenvalues, for some $N\le K$. Then, it can be shown that

$$\Psi = G(G^*G)^{-1/2}\mathbf{U} \tag{85}$$

and the corresponding eigenvalues are

$$\lambda_n = \mathbf{D}_{n,n}. \tag{86}$$

To see this, note that according to (85), Ψ is an isometry, since

$$\Psi^*\Psi = \mathbf{U}^*(G^*G)^{-1/2}G^*G(G^*G)^{-1/2}\mathbf{U} = \mathbf{U}^*\mathbf{U} = \mathbf{I}_K. \tag{87}$$

Furthermore, (83) and (84) imply that $R_X=\Psi\mathbf{D}\Psi^*$. Consequently

$$R_X\Psi = \Psi\mathbf{D}\Psi^*\Psi = \Psi\mathbf{D}, \tag{88}$$

which proves the claim.

It is important to emphasize that the K functions $\{\psi_n(t)\}_{n=1}^K$ span $\mathcal{G}$. Therefore, if one is allowed to take $N=K$ samples, then the optimal choice is a set of kernels that span $\mathcal{G}$. This conclusion is compatible with the CRB analysis of the previous sections. However, the advantage of the Bayesian viewpoint is that it allows us to identify the optimal sampling space when less than K samples are allowed. For example, suppose that $\{g_n(t)\}$ are orthonormal, and the coefficients $\{a_n\}$ are uncorrelated. Then the optimal sampling space is the one spanned by the N functions $\{g_n(t)\}$ corresponding to the N largest-variance coefficients $\{a_n\}$.

A second example demonstrating the derivation of the optimal sampling kernels will be given in the next section.

IX. Application: Channel Estimation

In this section, we focus on a specific application of FRI signals, namely, that of estimating a signal consisting of a number of pulses having unknown positions and amplitudes [5], [7], [8]. More precisely, we consider real periodic signals $x(t)$ of the form (16), which were discussed in Section III-C. These are T-periodic pulse sequences, in which the (real-valued) pulse shape $g(t)$ is known, but the amplitudes $\{a_\ell\}$ and delays $\{t_\ell\}$ are unknown. After analyzing periodic signals of this type, we will also compare estimation performance in this case with the semi-periodic family (17), and attempt to explain the empirically observed differences in stability between these two cases.

By defining the T-periodic function $h(t)=\sum_{n\in\mathbb{Z}}g(t-nT)$, we can write $x(t)$ of (16) as

$$x(t)=\sum_{\ell=1}^L a_\ell h(t-t_\ell). \tag{89}$$

Our goal is now to estimate $x(t)$ from samples of the noisy process $y(t)$ of (6). As before, we will assume that only continuous-time noise is present in the system. Since $x(t)$ is T-periodic, it suffices to recover the signal in the region $[0,T]$. In particular, we would like to identify the optimal sampling kernels for this setting, and to compare existing algorithms with the resulting CRB in order to determine when the optimal estimation performance is achieved.

Let

$$\tilde{h}_k = \frac{1}{T}\langle h,\varphi_k\rangle, \quad k\in\mathbb{Z} \tag{90}$$

be the Fourier series of $h(t)$, where $\varphi_k(t)=e^{j2\pi kt/T}$. The Fourier series of $x(t)$ is then given by

$$\tilde{x}_k \triangleq \frac{1}{T}\langle x,\varphi_k\rangle = \tilde{h}_k\sum_{\ell=1}^L a_\ell e^{-j\frac{2\pi}{T}kt_\ell}, \quad k\in\mathbb{Z}. \tag{91}$$

Let $\mathcal{K}=\{k\in\mathbb{Z}:\tilde{h}_k\neq 0\}$ denote the indices of the nonzero Fourier coefficients of $h(t)$. Suppose for a moment that $\mathcal{K}$ is finite. It then follows from (91) that $x(t)$ also has

a finite number of nonzero Fourier coefficients. Consequently, the set $\mathfrak{X}$ of possible signals $x(t)$ is contained in the finite-dimensional subspace $\mathcal{M} = \mathrm{span}\{\varphi_k\}_{k\in\mathcal{K}}$. Therefore, as explained in Section VII-D, choosing the $N = |\mathcal{K}|$ sampling kernels $\{s_n(t) = e^{-j2\pi nt/T}\}_{n\in\mathcal{K}}$ results in a sampled CRB which is equivalent to the continuous-time bound. This result is compatible with recent work demonstrating successful performance of FRI recovery algorithms using exponentials as sampling kernels [7].

Note, however, that this is a "Nyquist-equivalent" sampling scheme, i.e., the number of samples required $N = |\mathcal{K}|$ is potentially much higher than the number of degrees of freedom $2L$ in the signal $x(t)$ (see Section VII-D). This provides a theoretical explanation of the empirically recognized fact that sampling above the rate of innovation improves the performance of FRI techniques in the presence of noise [7], [8], a fact which stands in contrast to the noise-free performance guarantees of many FRI algorithms.

Moreover, if there exists an infinite number of nonzero coefficients $\tilde{h}_k$, then in general the set $\mathfrak{X}$ will not belong to any finite-dimensional subspace. Consequently, it will not be possible in this case for an algorithm based on a finite number of samples to achieve the performance obtainable from the complete signal $y(t)$. This occurs, for example, whenever the pulse $g(t)$ of (16) is time-limited. In such cases, any increase in the sampling rate will potentially continue to reduce the CRB, although the sampled CRB will converge to the asymptotic value of $\rho_{T_0}\sigma_c^2$ in the limit as the sampling rate increases.

A. Choosing the Sampling Kernels

An important question in the current setting is how to choose the sampling kernels so as to achieve the best possible performance under a limited budget of samples. This can be done via the Bayesian analysis provided in Section VIII. Assume, for example, that the time delays $\{t_\ell\}_{\ell=1}^L$ are independently drawn from a uniform distribution over the interval $[0, T]$. Furthermore, suppose that the amplitudes $\{a_\ell\}_{\ell=1}^L$ are mutually uncorrelated real zero-mean random variables which are independent of the time delays and have variance σ_a^2. Then,

$$\begin{aligned} R_X(t,\tau) &= \mathbb{E}\{x(t)x^*(\tau)\} \\ &= \sum_{k=1}^{L}\sum_{\ell=1}^{L} \mathbb{E}\{a_k a_\ell^*\}\, \mathbb{E}\{h(t-t_k)h^*(\tau-t_\ell)\} \\ &= \sigma_a^2 \sum_{\ell=1}^{L} \mathbb{E}\{h(t-t_\ell)h^*(\tau-t_\ell)\} \\ &= \sigma_a^2 L \frac{1}{T}\int_0^T h(t-t_\ell)h^*(\tau-t_\ell)dt_\ell \\ &= \sigma_a^2 L \sum_{k\in\mathbb{Z}} |\tilde{h}_k|^2 e^{j\frac{2\pi}{T}k(t-\tau)}, \end{aligned} \tag{92}$$

where we used Parseval's theorem. It is easily verified from the above expansion that the eigenfunctions of $R_X(t,\tau)$ are given by

$$\psi_n(t) = \frac{1}{\sqrt{T}}\, e^{j\frac{2\pi}{T}nt}, \quad n \in \mathbb{Z} \tag{93}$$

and the corresponding eigenvalues are

$$\lambda_n = L\sigma_a^2 T|\tilde{h}_n|^2, \quad n \in \mathbb{Z}. \tag{94}$$

Since $h(t)$ is real, $\tilde{h}_{-n} = \tilde{h}_n^*$ so that $|\tilde{h}_n| = |\tilde{h}_{-n}|$. Consequently, $\lambda_n = \lambda_{-n}$, and thus an equivalent set of eigenfunctions is

$$\tilde{\psi}_0(t) = \sqrt{1/T}, \tag{95}$$

$$\tilde{\psi}_n(t) = \sqrt{2/T}\, \cos(2\pi nt/T), \quad n \in \mathbb{N}, \tag{96}$$

$$\tilde{\psi}_{-n}(t) = \sqrt{2/T}\, \sin(2\pi nt/T), \quad n \in \mathbb{N}. \tag{97}$$

Therefore, the optimal set of N sampling functions is

$$s_n(t) = e^{j\frac{2\pi}{T}p_n t}, \quad n = 1, \ldots, N, \tag{98}$$

or, equivalently,

$$\begin{aligned} \tilde{s}_0(t) &= 1, \\ \tilde{s}_n(t) &= \cos(2\pi p_n t/T), \quad n = 1, \ldots, \frac{N-1}{2}, \\ \tilde{s}_{n+\frac{N+1}{2}}(t) &= \sin(2\pi p_n t/T), \quad n = 1, \ldots, \frac{N-1}{2}. \end{aligned} \tag{99}$$

Here, we denoted by p_n the index of the nth largest Fourier coefficient $|\tilde{h}_{p_n}|$ and we assumed for simplicity that N is odd and that the DC component $|\tilde{h}_0|$ is one of the N largest coefficients.

The above analysis again lends credence to the recently proposed time-delay estimation technique of Gedalyahu et al. [7], which makes use of complex exponentials as sampling functions. A disadvantage of this algorithm is that it can only handle a set of exponents with successive frequencies, while for general pulses, the indices of the N largest Fourier coefficients may be sporadic. As we will see in Section IX-C, this limitation may result in deteriorated performance of the algorithm in some cases.

B. Computing the CRB

Having identified the optimal sampling kernels (99), we would now like to compute the CRB for estimating $x(t)$ from the resulting samples. In order to compare these results with the continuous-time CRB, we assume that no digital noise is added in the sampling process. However, the calculations described below can be adapted without difficulty to situations containing both continuous-time and digital noise.

We define the parameter vector

$$\boldsymbol{\theta} = (a_1, \ldots, a_L, t_1, \ldots, t_L)^T \tag{100}$$

whose length is $K = 2L$. Theorem 3 provides a two-step process for computing the CRB of the signal $x(t)$ from its samples. First, the FIM $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$ for estimating $\boldsymbol{\theta}$ is determined. Second, the formula (51) is applied to compute the CRB. While Theorem 3 also provides a means for calculating $\mathbf{J}_{\boldsymbol{\theta}}^{\mathrm{samp}}$, it is more convenient in the present setting to derive the FIM directly. This can be done by calculating the expectations μ_n

of (44) and applying (52). In our setting, $\mu_n = \langle x, \tilde{s}_n \rangle$ are given by

$$\begin{aligned} \mu_n &= \frac{\tilde{x}_{p_n} + \tilde{x}_{-p_n}}{2}, \qquad n = 0, \ldots, \frac{N-1}{2} \\ \mu_{n+\frac{N+1}{2}} &= \frac{\tilde{x}_{p_n} - \tilde{x}_{-p_n}}{2j}, \qquad n = 1, \ldots, \frac{N-1}{2} \end{aligned} \tag{101}$$

where $\{\tilde{x}_n\}$ are the Fourier coefficients of $x(t)$. These coefficients depend in turn on the parameter vector $\boldsymbol{\theta}$, as shown in (91). Substituting μ_n into (52) yields a closed-form expression for $\mathbf{J}_{\boldsymbol{\theta}}^{\text{samp}}$. Since the resulting formula is cumbersome and not very insightful, it is not explicitly written herein.

To obtain the sampled CRB, our next step is to compute the $2L \times 2L$ matrix

$$\mathbf{M} \triangleq \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right). \tag{102}$$

The function $h_{\boldsymbol{\theta}} : \mathbb{R}^{2L} \to L_2$ maps a given parameter vector $\boldsymbol{\theta}$ to the resulting signal $x(t)$ as defined by (89). Differentiating this function with respect to $\boldsymbol{\theta}$, we find that the operator $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta} : \mathbb{R}^{2L} \to L_2$ is defined by

$$\begin{aligned} \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \mathbf{v} = \mathbf{v}_1 h(t-t_1) + \cdots + \mathbf{v}_L h(t-t_L) \\ - \mathbf{v}_{L+1} a_1 h'(t-t_1) - \cdots - \mathbf{v}_{2L} a_L h'(t-t_L) \end{aligned} \tag{103}$$

for any vector $\mathbf{v} \in \mathbb{R}^{2L}$.

One may now compute the ikth element of $\mathbf{M}$ as

$$M_{ik} = \mathbf{e}_i^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right)^* \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right) \mathbf{e}_k = \left\langle \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \mathbf{e}_i, \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \mathbf{e}_k \right\rangle. \tag{104}$$

Thus, each element of $\mathbf{M}$ is an inner product between two of the terms in (103). To calculate this inner product numerically for a given function $h(t)$, it is more convenient to use Parseval's theorem in order to convert the (continuous-time) inner product to a sum over Fourier coefficients. For example, in the case $1 \le i, k \le L$, we have

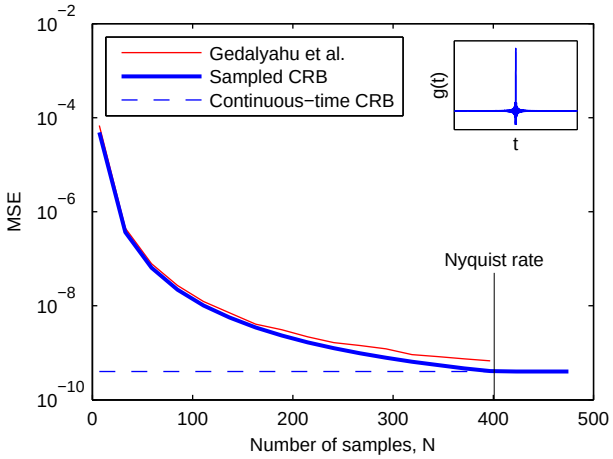
$$\begin{aligned} M_{ik} &= \int_0^T h(t-t_i) h^*(t-t_k) dt \\ &= T \sum_{n \in \mathbb{Z}} \tilde{h}_n e^{-j2\pi t_i n/T} \tilde{h}_n^* e^{j2\pi t_k n/T} \\ &= T \sum_{n \in \mathbb{Z}} |\tilde{h}_n|^2 e^{-j2\pi (t_k - t_i) n/T}, \quad 1 \le i, k \le L. \end{aligned} \tag{105}$$

An analogous derivation can be carried out when i or k are in the complementary range $L+1, \ldots, 2L$.

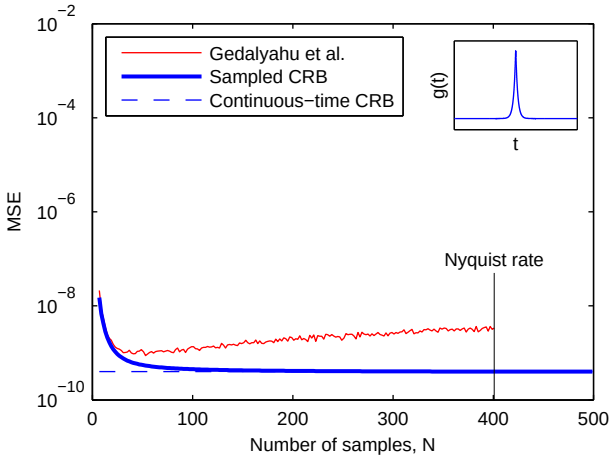
Finally, having calculated the matrices $\mathbf{J}_{\boldsymbol{\theta}}^{\text{samp}}$ and $\mathbf{M}$, the CRB for sampled measurements is obtained using (51). We are now ready to compare this bound to the performance of practical estimators in some specific scenarios.

C. Effect of the Pulse Shape

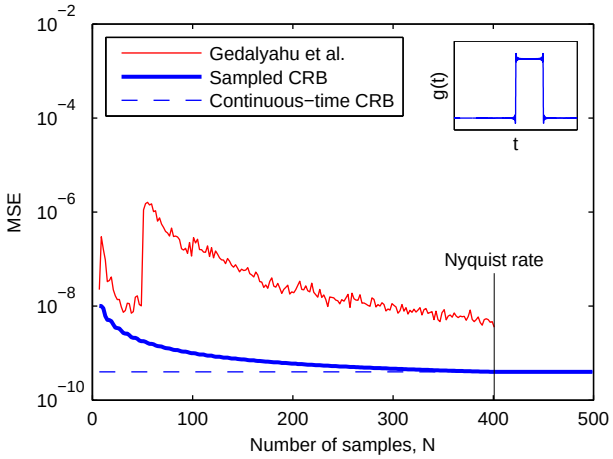
In Fig. 2, we document several experiments comparing the CRB with the time-delay estimation technique of Gedalyahu et al. [7]. Specifically, we sampled the signal $x(t)$ of (16) using a set of exponential kernels, and used the matrix pencil



(a) The pulse $g(t)$ is a filtered Dirac with 401 Fourier coefficients.



(b) The pulse $g(t)$ contains 401 nonzero Fourier coefficients which decrease monotonically with the frequency.



(c) The pulse $g(t)$ is a filtered $\text{rect}(\cdot)$ with 401 Fourier coefficients.

Fig. 2. Comparison of the CRB and the performance of a practical estimator, as a function of the number of samples.

method [48] to estimate $x(t)$ from the resulting measurements. Since we are considering only continuous-time noise, applying an invertible linear transformation to the sampling kernels has no effect on our performance bounds (see Section VII). The various kernels suggested in [7] amount to precisely such an invertible linear transformation, and the same performance bound applies to all of these approaches. Moreover, under the continuous-time noise model, it can be shown that these techniques also exhibit the same performance. For the same reason, the performance reported here is also identical to the method of Vetterli et al. [2].

In our experiments, a signal containing $L = 2$ pulses was constructed. The delays and amplitudes of the pulses were chosen randomly and are given by

$$\begin{aligned} a_1 &= 0.3204, & t_1 &= 0.6678, \\ a_2 &= 0.6063, & t_2 &= 0.9863. \end{aligned} \tag{106}$$

Modifications of these parameters does not appear to significantly affect the reported results, except when the time delays are close to one another, a situation which will be discussed in depth in Section IX-D. The pulse $h(t)$ consisted of $|\mathcal{K}| = 401$ nonzero Fourier coefficients at positions $\mathcal{K} = \{-200, \dots, 200\}$. The CRB is plotted as a function of the number of samples N, where the sampling kernels are given by $s_n(t) = e^{j2\pi nt/T}$ with $n \in \{-\lfloor N/2 \rfloor, \dots, \lfloor N/2 \rfloor\}$. This is done because the matrix pencil method requires the sampling kernels to have contiguous frequencies.

In Fig. 2(a), we chose $\tilde{h}_k = 1$ for $-200 \leq k \leq 200$ and $\tilde{h}_k = 0$ elsewhere; these are the low-frequency components of a Dirac delta function. The noise standard deviation was $\sigma_c = 10^{-5}$. In this case, for a fixed budget of N samples, any choice of N exponentials having frequencies in the range $-200 \leq k \leq 200$ is optimal according to the criterion of Section VIII. As expected, the sampled CRB achieves the continuous-time bound $K\sigma_c^2$ when $N \geq |\mathcal{K}|$. However, the CRB obtained at low sampling rates is higher by several orders of magnitude than the continuous-time limit. This indicates that the maxim of FRI theory, whereby sampling at the rate of innovation suffices for reconstruction, may not always hold in the presence of mild levels of noise. Indeed, if no noise is added in the present setting, then perfect recovery can be guaranteed using as few as $N = 4$ samples; yet even in the presence of mild noise, our bounds demonstrate that performance is quite poor unless the number of samples is increased substantially. This result may provide an explanation for the previously observed numerical instability of FRI techniques [2], [7], [8].

As a further observation, we note that in this scenario, existing algorithms come very close to the CRB. Thus, the previously observed improvements achieved by oversampling are a result of fundamental limitations of low-rate sampling, rather than drawbacks of the specific technique used.

The same experiment is repeated in Fig. 2(b) with a pulse having Fourier coefficients $\tilde{h}_k = 1/(1 + 0.01k^2)$. Since the Fourier coefficients decrease with $|k|$, in this case our choice of low-frequency sampling kernels aligns with the choice dictated by the linear recovery analysis of Section VIII. However, the SNR of the measurements c_n decreases with n. As can be seen, this has a negative effect on the performance of the algorithm, which is not designed for high noise levels. Indeed, including low-SNR measurements causes the MSE not only to depart from the CRB, but eventually even to increase as more noisy samples are provided. In other words, one would do better to ignore the high-frequency measurements than to feed them to the recovery algorithm. Yet information is clearly present in these high-frequency samples, as indicated by the continual decrease of the CRB with increasing N. Thus, our analysis indicates that improved estimation techniques should be achievable in this case, in particular by careful utilization of low-SNR measurements.

The adverse effect of low-SNR measurements is exacerbated if, for a given N, one does not choose the N largest Fourier coefficients. This is demonstrated in Fig. 2(c). Here, the results of a similar experiment are plotted, in which $\tilde{h}_k = P\,\mathrm{sinc}(nP/T)$, $-200 \leq k \leq 200$. These are the 401 lowest-frequency Fourier coefficients of a rectangular pulse having width P. In this case, the Fourier coefficients are no longer monotonically decreasing with $|k|$. Consequently, the sampling kernels $s_n(t) = e^{j2\pi nt/T}$ with $n \in \{-\lfloor N/2 \rfloor, \dots, \lfloor N/2 \rfloor\}$ do not correspond to the N largest Fourier coefficients, and thus are not optimal, at least from a linear recovery perspective. In particular, for the chosen parameters, $|\tilde{h}_{25}| = |\tilde{h}_{-25}|$ are considerably smaller than the rest of the coefficients. When $N \geq 50$, the corresponding measurements are included, causing the MSE to deteriorate significantly.

D. Closely-Spaces Pulses

It is well-known that the estimation of pulse positions becomes ill-conditioned when several of the pulses are located close to one another. Intuitively, this is a consequence of the overlap between the pulses, which makes it more difficult to identify the precise location of each pulse. However, our goal is to estimate the signal $x(t)$ itself, rather than the positions of its constituent pulses. As we will see, for this purpose the effect of closely-spaced pulses is less clear-cut.

To study the effect of pulse position on the estimation error, we used a setup similar to the one of Fig. 2(b), with the following differences. First, a higher noise level of $\sigma_c = 10^{-3}$ was chosen. Second, the signal consisted of $L = 2$ pulses, with the first pulse at position $t_1 = 0.5$. The position of the second pulse was varied in the range $[0.3, 0.7]$ to demonstrate the effect of pulse proximity on the performance. The setting was otherwise identical to that of Section IX-C. In particular, recall that $T = 1$.

The results of this experiment are plotted in Fig. 3, which documents both the values of the sampled CRB and the actual MSE obtained by the estimator of Gedalyahu et al. [7]. The continuous-time CRB is also plotted, although, as is evident from Theorem 2, this bound is a function only of the number of parameters determining the signal, and is therefore unaffected by the proximity of the pulses.

Several different effects are visible in Fig. 3. First, as the two pulses begin to come closer, both the CRB and the observed MSE increase by several orders of magnitude; this occurs

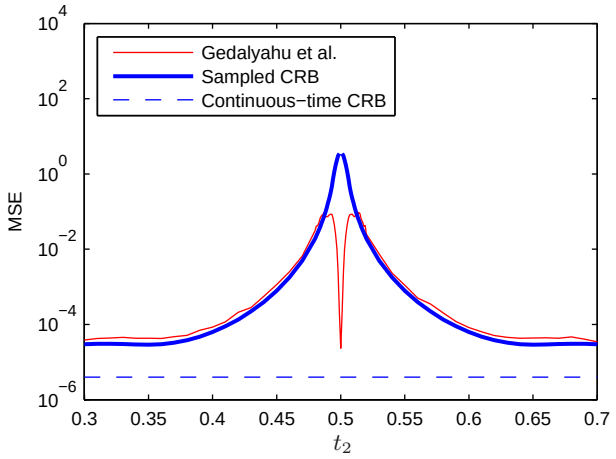
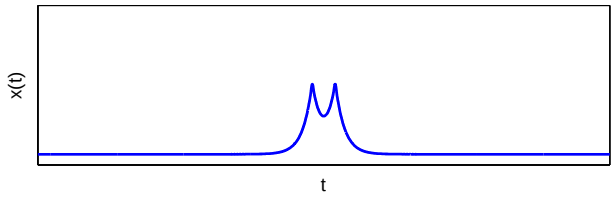
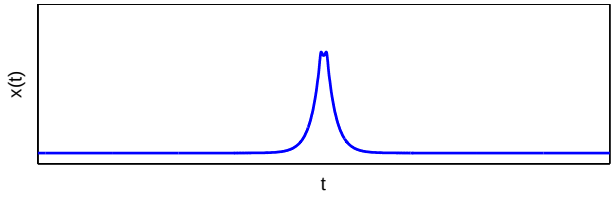


Fig. 3. Comparison between the CRB and the performance of a practical estimator as a function of the pulse positions. The signal contains $L = 2$ pulses, the first of which is located at $t_1 = 0.5$. The MSE is plotted as a function of the position of the second pulse.



(a) The spacing between the pulses is 0.04.



(b) The spacing between the pulses is 0.01.

Fig. 4. Demonstration of the different levels of overlap between pulses.

when $|t_1 - t_2|$ is between about 0.15 and 0.03. (Of course, the precise distances at which these effects occur depend on the pulse width and other parameters of the experiment.) This level of proximity is demonstrated in Fig. 4(a). At this stage, the overlap between the pulses is sufficient to make it more difficult to estimate their positions accurately, but the separation between the pulses is still large, so that they are not mistaken for a single pulse.

As the pulses draw nearer each other, they begin to resemble a single pulse located at $(t_1+t_2)/2$ (see Fig. 4(b)). Depending on the noise level, at some point the estimation algorithm will indeed identify the two pulses as one. Since our goal is to estimate $x(t)$ and not the pulse positions, such an "error" causes little deterioration in MSE. This is visible in Fig. 3 as the region in which the MSE of the practical algorithm ceases to deteriorate and ultimately decreases.

Interestingly, the CRB does not capture this improvement in performance. This failure is due to the fact that the CRB applies only to unbiased estimators, while the strategy utilized in [7] becomes biased for closely-spaced pulses. For an estimator to be unbiased, it is required that the mean estimate, averaged over noise realizations, will converge to the true value of $x(t)$, which has a form similar to that of Fig. 4(b). The expectation of an estimator reconstructing a single pulse will not have the form of two closely-spaced pulses; such an estimator is thus necessarily biased. In other words, the discrepancy observed here results from the fact that in this case, biased techniques outperform the best unbiased approach.

E. Non-Periodic and Semi-Periodic Signal Models

As we have seen above, the reconstruction of signals of the form (16) in the presence of noise is often severely hampered when sampled at or slightly above the rate of innovation. Rather than indicating a lack of appropriate algorithms, in many cases this phenomenon results from fundamental limits on the ability to recover such signals from noisy measurements. A similar effect was demonstrated [7], [8] in the non-periodic (or finite) pulse stream model (14). In fact, if one is allowed to sample a non-periodic pulse stream with arbitrary sampling kernels, then by designing kernels having sufficiently large time-domain support, one can capture all or most of the energy in the signal. This setting then essentially becomes equivalent to a periodic signal model (16) in which the period is larger than the effective support of the pulse stream: One can imagine that the signal repeats itself beyond the sampled region, as this would not affect the measurements. Consequently, it is not surprising that the non-periodic model demonstrates substantial improvement in the presence of oversampling [8].

On the other hand, some types of FRI and union of subspace signals exhibit remarkable noise resilience, and do not appear to require substantial oversampling in the presence of noise [5], [27]. As we now show, the CRB can be used to verify that such phenomena arise from a fundamental difference between families of FRI signals.

As an example, we compare the CRB for reconstructing the periodic signal (16) with the semi-periodic signal (17). Recall that in the former case, each period consists of pulses having unknown amplitudes and time shifts. By contrast, in the latter signal, the time delays are identical throughout all periods, but the amplitudes can change from one period to the next.

While these are clearly different types of signals, an effort was made to form a fair comparison between the reconstruction capabilities in the two cases. To this end, we chose an identical pulse $g(t)$ in both cases. We selected the signal segment $[0, T_0]$, where $T_0 = 1$, and chose the signal parameters so as to guarantee an identical T_0-local rate of innovation. We also used identical sampling kernels in both settings: specifically, we chose the kernels (99) which measure the N lowest frequency components of the signal.

To simplify the analysis and focus on the fundamental differences between these settings, we will assume in this section that the pulses $g(t)$ are compactly supported, and that the time delays are chosen such that pulses from one period do not overlap with other periods. In other words, if the support of $g(t)$ is given by $[t_a, t_b]$, then we require

$$\min_{\ell}\{t_\ell\} > t_a \quad \text{and} \quad \max_{\ell}\{t_\ell\} < T - t_b. \tag{107}$$

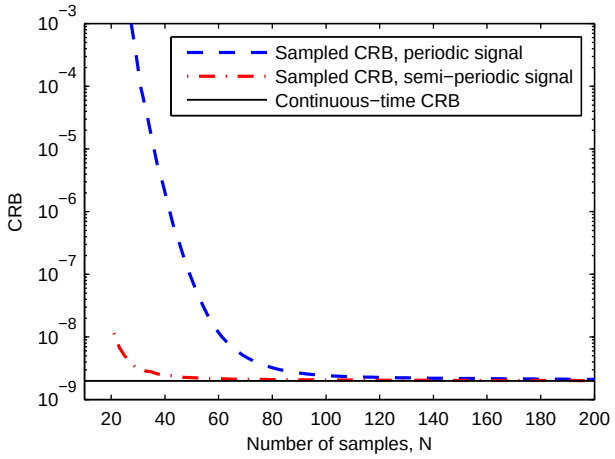


Fig. 5. Comparison between the CRB for a periodic signal (16) and a semi-periodic signal (17).

Specifically, we chose the pulse $g(t)$ used in Fig. 2(b), which is compactly supported to a high approximation.

For the periodic signal, we chose $L = 10$ pulses with random delays and amplitudes, picked so as to satisfy the condition (107). A period of $T = 1$ was selected. This implies that the signal of interest is determined by $2L = 20$ parameters (L amplitudes and L time delays).

To construct a semi-periodic signal with the same number of parameters, we chose a period of $T = 1/9$ containing $L = 2$ pulses. The segment $[0, T_0]$ then contains precisely $M = 9$ periods, for a total of 20 parameters. While it may seem plausible to require the same number of periods for both signals, this would actually disadvantage the periodic approach, as it would require the estimation of much more closely-spaced pulses.

The CRB for the periodic signal was computed as explained in Section IX-B, and the CRB for the semi-periodic signal can be calculated in a similar fashion. The results are compared with the continuous-time CRB in Fig. 5. Note that since the number of parameters to be estimated is identical in both signal models, the continuous-time CRB for the two settings coincides. Consequently, for a large number of measurements, the sampled bounds also converge to the same values. However, when the number of samples is closer to the rate of innovation, the bound on the reconstruction error for the semi-periodic signal is much lower than that of the periodic signal. As mentioned above, this is in agreement with previously reported findings for the two types of signals [2], [5], [7].

To find an explanation for this difference, it is helpful to recall that both signals can be described using the union of subspaces viewpoint (see Section III-C). Each of the signals in this experiment is defined by precisely 20 parameters, which determine the subspace to which the signal belongs and the position within this subspace. Specifically, the values of the time delays select the subspace, and the pulse amplitudes define a point within this subspace. Thus, in the above setting, the periodic signal contains 10 parameters for selecting the subspace and 10 additional parameters determining the position within it; whereas for the semi-periodic signal, only 2 parameters determine the subspace while the remaining 18 parameters set the location in the subspace. Evidently, identification of the subspace is challenging, especially in the presence of noise, but once the subspace is determined, the remaining parameters can be estimated using a simple linear operation (a projection onto the chosen subspace). Consequently, if many of the unknown parameters identify the position within a subspace, estimation can be performed more accurately. This may provide an explanation for the difference between the two examined signal models.

As further evidence in support of this explanation, we recall from Section III-C that the multiband signal model (19) can also be viewed as a union of subspaces. Here, again, the parameters $\{\omega_\ell\}_{\ell=1}^{L}$ determining the subspace (i.e., the utilized frequency bands) are far fewer than the parameters $\{a_\ell[n]\}$ selecting the point within the subspace (i.e., the content of each frequency band). In support of the proposed explanation, highly noise resistant algorithms can be constructed for the recovery of multiband signals [27], [29]. An even more extreme case is the single subspace setting, exemplified by shift-invariant signals (Section III-A). In this case, all of the signal parameters are used to determine the position within the one possible subspace. As we have seen in Section VII-C, in this case Nyquist-equivalent sampling at the rate of innovation achieves the continuous-time CRB.

X. Summary

In this paper, we studied the inherent limitations in recovering FRI signals from noisy measurements. We derived a continuous-time CRB which provides a lower bound on the achievable MSE of any unbiased estimation method, regardless of the sampling mechanism. We showed that the rate of innovation ρ_{T_0} is a lower bound on the ratio between the average MSE achievable by any unbiased estimator and the noise variance σ_c^2, *regardless of the sampling method*. This stands in contrast to the noise-free interpretation of ρ_{T_0} as the minimum sampling rate required for perfect recovery.

We next examined the CRB for estimating an FRI signal from a discrete set of noisy samples. We showed that the sampled bound is in general higher than the continuous-time CRB, and approaches it as the sampling rate increases. In general, the rate which is needed in order to achieve the continuous-time CRB is equal to the rate associated with the smallest subspace that encompasses all possible signal realizations. In particular, if a signal belongs to a union of subspaces, then the rate required to achieve the continuous-time bound is that associated with the sum of the subspaces. In some cases, this rate is finite, but in other cases the sum covers the entire space L_2 and no finite-rate technique achieves the CRB.

A consequence of these results is that oversampling can generally improve estimation performance. Indeed, our experiments demonstrate that sampling rates much higher than ρ_{T_0} are required in certain settings in order to approach the optimal performance. Furthermore, these gains can be substantial: In some cases, oversampling can improve the MSE by several orders of magnitude. We showed that the CRB can be used to

determine which estimation problems require substantial oversampling to achieve stable performance. As a rule of thumb, it appears that for union of subspace signals, performance is improved at low rates if most of the parameters identify the position within the subspace, rather than the subspace itself. Our analysis can also be used to identify cases in which no existing algorithm comes close to the CRB, implying that better approaches can be constructed. In particular, it seems that existing algorithms do not deal well with measurement sets having a wide dynamic range.

Lastly, we addressed the problem of choosing the sampling kernels. This was done by adopting a Bayesian framework, so that an optimality criterion can be rigorously defined. Using a generalization of the KLT, we showed that the optimal kernels are the eigenfunctions of the autocorrelation function of the signal. In the context of time-delay estimation, these kernels are exponentials with appropriately chosen frequencies. This choice coincides with recent FRI techniques [5].

Appendix A
Proof of Theorem 1

The following notation will be used within this appendix. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two measurable Hilbert spaces, and let $(\Omega, \mathscr{F}, P)$ be a probability space. Consider two random variables $u : \Omega \to \mathcal{H}_1$ and $v : \Omega \to \mathcal{H}_2$. Then, the notation $\mathbb{E}\{uv^*\}$ will be used to denote the linear operator $\mathcal{H}_2 \to \mathcal{H}_1$ such that, for any $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$,

$$\langle h_1, \mathbb{E}\{uv^*\}\, h_2\rangle_{\mathcal{H}_1} = \mathbb{E}\big\{\langle h_1, u\rangle_{\mathcal{H}_1}\, \langle v, h_2\rangle_{\mathcal{H}_2}\big\} \tag{108}$$

if the expectation exists for all h_1 and h_2.

We begin by stating two general lemmas which will be of use in the proof of Theorem 1.

Lemma 1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces, and consider the operators*

$$\begin{aligned} &A : \mathcal{H}_1 \to \mathcal{H}_1,\\ &B : \mathcal{H}_2 \to \mathcal{H}_1,\\ &C : \mathcal{H}_2 \to \mathcal{H}_2. \end{aligned} \tag{109}$$

Suppose C is self-adjoint and invertible. Define the product Hilbert space $\mathcal{H}_1 \times \mathcal{H}_2$ in the usual manner, and suppose the operator $M : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathcal{H}_1 \times \mathcal{H}_2$ defined by

$$M\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} Ah_1 + Bh_2 \\ B^*h_1 + Ch_2 \end{pmatrix} \tag{110}$$

is positive semidefinite (psd). Then,

$$A \succeq BC^{-1}B^* \tag{111}$$

in the sense that the $\mathcal{H}_1 \to \mathcal{H}_1$ operator $A - BC^{-1}B^*$ is psd.

Proof: Since M is psd, we have for any $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$

$$\left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, M\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right\rangle_{\mathcal{H}_1 \times \mathcal{H}_2} \geq 0 \tag{112}$$

which implies

$$\langle h_1, Ah_1\rangle_{\mathcal{H}_1} + 2\Re\left[\langle h_1, Bh_2\rangle_{\mathcal{H}_2}\right] + \langle h_2, Ch_2\rangle_{\mathcal{H}_2} \geq 0. \tag{113}$$

Choosing $h_2 = -C^{-1}B^*h_1$, we have that $\langle h_1, Bh_2\rangle_{\mathcal{H}_2} = -\left\langle B^*h_1, C^{-1}B^*h_1\right\rangle_{\mathcal{H}_1}$, which is real since C^{-1} is self-adjoint. It follows from (113) that

$$\langle h_1, Ah_1\rangle_{\mathcal{H}_1} - \left\langle h_1, BC^{-1}B^*h_1\right\rangle_{\mathcal{H}_1} \geq 0 \tag{114}$$

which leads to (111), as required. ∎

Lemma 2. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces and let $(\Omega, \mathscr{F}, P)$ be a probability space. Let $u : \Omega \to \mathcal{H}_1$ and $v : \Omega \to \mathcal{H}_2$ be random variables, and suppose the expectations $\mathbb{E}\{uu^*\}$, $\mathbb{E}\{uv^*\}$, and $\mathbb{E}\{vv^*\}$ exist as linear operators as defined in* (108)*. If $\mathbb{E}\{vv^*\}$ is invertible, then*

$$\mathbb{E}\{uu^*\} \succeq \mathbb{E}\{uv^*\}\left(\mathbb{E}\{vv^*\}\right)^{-1}\mathbb{E}\{vu^*\}. \tag{115}$$

Proof: Let us denote $A = \mathbb{E}\{uu^*\}$, $B = \mathbb{E}\{uv^*\}$, and $C = \mathbb{E}\{vv^*\}$ and define the linear operator $M : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathcal{H}_1 \times \mathcal{H}_2$ as in (110). From (108), for any $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ we have

$$\begin{aligned} &\left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, M\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right\rangle_{\mathcal{H}_1 \times \mathcal{H}_2}\\ &= \langle h_1, Ah_1\rangle_{\mathcal{H}_1} + 2\Re\left[\langle h_1, Bh_2\rangle_{\mathcal{H}_1}\right] + \langle h_2, Ch_2\rangle_{\mathcal{H}_2}\\ &= \mathbb{E}\Big\{\left|\langle h_1, u\rangle_{\mathcal{H}_1}\right|^2 + 2\Re\left[\langle h_1, u\rangle_{\mathcal{H}_1}\langle v, h_2\rangle_{\mathcal{H}_2}\right]\\ &\qquad + \left|\langle h_2, v\rangle_{\mathcal{H}_2}\right|^2\Big\}\\ &= \mathbb{E}\Big\{\left|\langle h_1, u\rangle_{\mathcal{H}_1} + \langle v, h_2\rangle_{\mathcal{H}_2}\right|^2\Big\}\\ &\geq 0. \end{aligned} \tag{116}$$

Thus M is a psd operator. Invoking Lemma 1 yields (115), as required. ∎

We are now ready to prove Theorem 1.

Proof of Theorem 1: Throughout the proof, let $\boldsymbol{\theta}$ be a fixed parameter and consider all functions as implicitly dependent on $\boldsymbol{\theta}$. Define the random variables

$$u : \Omega \to \mathcal{H} : \qquad u(\omega) = \hat{x}(y(\omega)) - h_{\boldsymbol{\theta}}, \tag{117}$$

$$v : \Omega \to \mathbb{R}^K : \qquad v(\omega) = \frac{\partial \log p_{\boldsymbol{\theta}}(y(\omega))}{\partial \boldsymbol{\theta}}. \tag{118}$$

We then have the linear operators $\mathbb{E}\{vv^*\} : \mathbb{R}^K \to \mathbb{R}^K$, $\mathbb{E}\{uu^*\} : \mathcal{H} \to \mathcal{H}$, and $\mathbb{E}\{uv^*\} : \mathbb{R}^K \to \mathcal{H}$, which satisfy

$$\mathbb{E}\{vv^*\} = \mathbf{J}_{\boldsymbol{\theta}}, \tag{119}$$

$$\langle \varphi_i, \mathbb{E}\{uu^*\}\, \varphi_j\rangle = \mathbb{E}\{\langle \varphi_i, u\rangle\, \langle u, \varphi_j\rangle\}, \tag{120}$$

$$\langle \varphi_i, \mathbb{E}\{uv^*\}\, \mathbf{e}_j\rangle = \mathbb{E}\left\{\langle \varphi_i, u\rangle\, \frac{\partial \log p_{\boldsymbol{\theta}}(y)}{\partial \theta_j}\right\}, \tag{121}$$

where $\{\varphi_n\}_{n\in\mathbb{Z}}$ denotes a complete orthonormal basis for $\mathcal{H}$. The operator $\mathbb{E}\{uu^*\}$ can be thought of as the covariance of $\hat{x}$, and is well-defined since, by (32), $\hat{x}$ has finite variance. Indeed, we have

$$\sum_{i\in\mathbb{Z}} \langle \varphi_i, \mathbb{E}\{uu^*\}\, \varphi_i\rangle = \mathbb{E}\big\{\|\hat{x} - h_{\boldsymbol{\theta}}\|_{L_2}^2\big\} < \infty \tag{122}$$

so that $\mathbb{E}\{uu^*\}$ is not only well-defined, but a trace class operator. Furthermore, $\mathbb{E}\{vv^*\} = \mathbf{J}_{\boldsymbol{\theta}}$ is well-defined and invertible by Assumption P5. The operator $\mathbb{E}\{uv^*\}$ is thus also well-defined by virtue of the Cauchy–Schwarz inequality.

To prove the theorem, we will show that

$$\mathbb{E}\{uv^*\} = \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \tag{123}$$

and then obtain the required result by applying Lemma 2. To demonstrate (123), observe that

$$\begin{aligned}\langle \varphi_i, \mathbb{E}\{uv^*\}\,\mathbf{e}_j\rangle &= \mathbb{E}\left\{ \langle \varphi_i, u\rangle \frac{\partial \log p_{\boldsymbol{\theta}}(y)}{\partial \theta_j} \right\} \\ &= \int \langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle \frac{1}{p(y;\boldsymbol{\theta})} \frac{\partial p(y;\boldsymbol{\theta})}{\partial \theta_j} p(y;\boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy) \\ &= \int \langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle \lim_{\Delta\to 0} \frac{p(y;\boldsymbol{\theta}+\Delta\mathbf{e}_j) - p(y;\boldsymbol{\theta})}{\Delta} P_{\boldsymbol{\theta}_0}(dy).\end{aligned} \tag{124}$$

By Assumption P4, for any sufficiently small $\Delta > 0$ we have

$$\begin{aligned}&\left| \langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle \frac{p(y;\boldsymbol{\theta}+\Delta\mathbf{e}_j) - p(y;\boldsymbol{\theta})}{\Delta} \right| \\ &\quad \le |\langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle|\, q(y,\boldsymbol{\theta}).\end{aligned} \tag{125}$$

Let us demonstrate that the right-hand side of (125) is absolutely integrable. By the Cauchy–Schwarz inequality,

$$\begin{aligned}&\left(\int |\langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle|\, q(y,\boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy) \right)^2 \\ &\quad \le \int |\langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle|^2 P_{\boldsymbol{\theta}_0}(dy) \cdot \int q^2(y,\boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy).\end{aligned} \tag{126}$$

The rightmost integral in (126) is finite by virtue of (29). As for the remaining integral, we have

$$\begin{aligned}&\int |\langle \varphi_i, \hat{x}(y) - h_{\boldsymbol{\theta}}\rangle|^2 P_{\boldsymbol{\theta}_0}(dy) \\ &\quad \le \int \|\hat{x}(y) - h_{\boldsymbol{\theta}}\|^2 P_{\boldsymbol{\theta}_0}(dy) \\ &\quad \overset{\text{(a)}}{\le} \int \left(\|\hat{x}(y)\| + \|h_{\boldsymbol{\theta}}\|\right)^2 P_{\boldsymbol{\theta}_0}(dy) \\ &\quad \overset{\text{(b)}}{\le} \int \|\hat{x}(y)\|^2 P_{\boldsymbol{\theta}_0}(dy) + \|h_{\boldsymbol{\theta}}\|^2 \int P_{\boldsymbol{\theta}_0}(dy) \\ &\qquad + 2\|h_{\boldsymbol{\theta}}\| \left(\int \|\hat{x}(y)\|^2 P_{\boldsymbol{\theta}_0}(dy) \right)^{1/2} \\ &\quad \overset{\text{(c)}}{<} \infty\end{aligned} \tag{127}$$

where we have used the triangle inequality in (a), the Cauchy–Schwarz inequality in (b), and the assumption (32) that $\hat{x}$ has finite energy in (c). We conclude that (125) is bounded by an absolutely integrable function, and we can thus apply the dominated convergence theorem to (124), obtaining

$$\begin{aligned}\langle \varphi_i, \mathbb{E}\{uv^*\}\,\mathbf{e}_j\rangle = \frac{\partial}{\partial \theta_j} \int \langle \varphi_i, \hat{x}(y)\rangle\, p(y;\boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy) \\ - \langle \varphi_i, h_{\boldsymbol{\theta}}\rangle \frac{\partial}{\partial \theta_j} \int p(y;\boldsymbol{\theta}) P_{\boldsymbol{\theta}_0}(dy).\end{aligned} \tag{128}$$

The second integral in (128) equals 1 and its derivative is therefore 0. Thus we have

$$\langle \varphi_i, \mathbb{E}\{uv^*\}\,\mathbf{e}_j\rangle = \frac{\partial \mathbb{E}\{\langle \varphi_i, \hat{x}(y)\rangle\}}{\partial \theta_j} = \frac{\partial \langle \varphi_i, h_{\boldsymbol{\theta}}\rangle}{\partial \theta_j}. \tag{129}$$

On the other hand, note that the Fréchet derivative $\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}$ of (31) coincides with the Gâteaux derivative of $h_{\boldsymbol{\theta}}$. In other words, for any vector $\mathbf{v} \in \mathbb{R}^K$, we have

$$\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \mathbf{v} = \lim_{\varepsilon \to 0} \frac{h_{\boldsymbol{\theta}+\varepsilon\mathbf{v}} - h_{\boldsymbol{\theta}}}{\varepsilon}. \tag{130}$$

It follows that

$$\left\langle \varphi_i, \frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \mathbf{e}_j \right\rangle = \frac{\partial \langle \varphi_i, h_{\boldsymbol{\theta}}\rangle}{\partial \theta_j}. \tag{131}$$

Since $\mathbb{E}\{uv^*\}$ and $(\partial h_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta})^*$ are both linear operators, (129) and (131) imply that the two operators are equal, demonstrating (123). Applying Lemma 2 and using the results (119) and (123), we have

$$\mathbb{E}\{(\hat{x} - h_{\boldsymbol{\theta}})(\hat{x} - h_{\boldsymbol{\theta}})^*\} \succeq \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right) \mathbf{J}_{\boldsymbol{\theta}}^{-1} \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right)^*. \tag{132}$$

As we have seen, the left-hand side of (132) is trace class, and thus so is the right-hand side. Taking the trace of both sides of the equation, we obtain

$$\mathbb{E}\{\|\hat{x} - h_{\boldsymbol{\theta}}\|^2\} \ge \mathrm{Tr}\left(\left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right) \mathbf{J}_{\boldsymbol{\theta}}^{-1} \left(\frac{\partial h_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right)^* \right) \tag{133}$$

which is equivalent to (33), as required. ∎

Appendix B
Maximization of (77)

The task of maximizing (77) is most easily accomplished by optimizing the coordinates of $s_n(t)$ in the orthonormal basis for $L_2[0,T]$ generated by the eigenfunctions of $R_X(t,\eta)$. Specifically, the function $s_n(t)$ can be written as

$$s_n(t) = \sum_{k=1}^{\infty} \alpha_k^n \left(\lambda_k + \sigma_c^2\right)^{-\frac{1}{2}} \psi_k(t), \tag{134}$$

with $\{\psi_k(t)\}_{k=1}^{\infty}$ and $\{\lambda_k\}_{k=1}^{\infty}$ of (68). (The coefficients $(\lambda_k + \sigma_c^2)^{-1/2}$ are inserted since they simplify the subsequent analysis.) Now, by Mercer's theorem, $R_X(t,\eta)$ can be expressed as

$$R_X(t,\eta) = \sum_{\ell=1}^{\infty} \lambda_\ell \psi_\ell(t) \psi_\ell^*(\eta), \tag{135}$$

where the convergence is absolute and uniform. Therefore

$$\begin{aligned}&\int_0^T R_X(t,\tau) s_n(\tau) d\tau \\ &\quad = \int_0^T \sum_{k=1}^{\infty} \alpha_k^n \left(\lambda_k + \sigma_c^2\right)^{-\frac{1}{2}} \psi_k(\tau) \sum_{\ell=1}^{\infty} \lambda_\ell \psi_\ell(t) \psi_\ell^*(\tau) d\tau \\ &\quad = \sum_{k=1}^{\infty} \alpha_k^n \frac{\lambda_k}{(\lambda_k + \sigma_c^2)^{\frac{1}{2}}} \psi_k(t),\end{aligned} \tag{136}$$

and consequently, by Parseval's theorem, (77) is given by

$$\begin{aligned}&\sum_{n=1}^{N} \int_0^T \left| \int_0^T R_X(t,\tau) s_n(\tau) d\tau \right|^2 dt \\ &\quad = \sum_{n=1}^{N} \int_0^T \left| \sum_{k=1}^{\infty} \alpha_k^n \frac{\lambda_k}{(\lambda_k + \sigma_c^2)^{\frac{1}{2}}} \psi_k(t) \right|^2 dt \\ &\quad = \sum_{n=1}^{N} \sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k + \sigma_c^2}.\end{aligned} \tag{137}$$

Similarly, using (134) and (136), we have

$$\iint_0^T s_m^*(t)R_X(t,\tau)s_n(\tau)d\tau dt = \sum_{k=1}^{\infty} \alpha_k^n(\alpha_k^m)^* \frac{\lambda_k}{\lambda_k+\sigma_c^2} \tag{138}$$

and by (134)

$$\sigma_c^2 \int_0^T s_m^*(t)s_n(\tau)dt = \sum_{k=1}^{\infty} \alpha_k^n(\alpha_k^m)^* \frac{\sigma_c^2}{\lambda_k+\sigma_c^2}. \tag{139}$$

Combining (138) and (139), the set of constraints (69) is translated to

$$\sum_{k=1}^{\infty} \alpha_k^m(\alpha_k^n)^* = \delta_{m,n} \tag{140}$$

for every $m,n = 1,\ldots,N$. Consequently, our problem has now been reduced to

$$\max_{\{\alpha_k^n\}} \sum_{n=1}^{N}\sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} \quad \text{s.t.} \quad \sum_{k=1}^{\infty} \alpha_k^m(\alpha_k^n)^* = \delta_{m,n}. \tag{141}$$

We now show that the sequences $\{\alpha_k^n\}$ which solve (141) must satisfy $\alpha_k^n = 0$ for every $k > N$ and $n = 1,\ldots,N$. To see this, assume to the contrary that the nth sequence satisfies $\alpha_\ell^n \neq 0$ for some $\ell > N$. We can then replace this sequence by a sequence $\{\tilde{\alpha}_k^n\}_{k\in\mathbb{Z}}$ satisfying

$$|\tilde{\alpha}_k^n|^2 = \begin{cases} |\alpha_k^n|^2 + a_k^2 & 1 \le k \le N \\ 0 & k = \ell \\ |\alpha_k^n|^2 & N < k \text{ and } k \neq \ell \end{cases} \tag{142}$$

where $\sum_{k=1}^N a_k^2 = |\alpha_\ell^n|^2$ (to ensure that $\sum_{k\in\mathbb{Z}} |\tilde{\alpha}_k^n|^2 = 1$). Such a set of coefficients $\{a_k\}_{k=1}^N$ can always be found since the N-term truncation of the remaining $N-1$ sequences cannot span $\mathbb{C}^N$. With this sequence, the nth summand in the objective of (141) becomes

$$\begin{aligned} \sum_{k=1}^{\infty} |\tilde{\alpha}_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} &= \sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} \\ &\quad + \left(\sum_{k=1}^{N} a_k^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} - |\alpha_\ell^n|^2 \frac{\lambda_\ell^2}{\lambda_\ell+\sigma_c^2}\right) \\ &\ge \sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} + \frac{\lambda_\ell^2}{\lambda_\ell+\sigma_c^2}\left(\sum_{k=1}^{N} a_k^2 - |\alpha_\ell^n|^2\right) \\ &= \sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2}, \end{aligned} \tag{143}$$

where we used the fact that $\lambda_k \ge \lambda_\ell$ for every $k < \ell$ and that $z^2/(a+z)$ is a monotone increasing function of z for all $z > 0$. This contradicts the optimality of $\{\alpha_k^n\}_{k\in\mathbb{Z}}$. Therefore, the set of sequences maximizing (141) satisfy $\alpha_k^n = 0$ for every $k > N$ and $n = 1,\ldots,N$.

It remains to determine the optimal values of the first N elements of each of the N sequences $\{\alpha_k^n\}_{k\in\mathbb{Z}}$, $n = 1,\ldots,N$. For this purpose, let $\mathbf{A}$ denote the $N\times N$ matrix whose entries are $\mathbf{A}_{k,n} = \alpha_k^n$ and let $\mathbf{\Lambda}$ be a diagonal matrix with $\mathbf{\Lambda}_{k,k} = \lambda_k^2/(\lambda_k+\sigma_c^2)$. Then, the constraint (140) can be written as $\mathbf{A}^*\mathbf{A} = \mathbf{I}_N$, which is equivalent to $\mathbf{A}\mathbf{A}^* = \mathbf{I}_N$. Now, the objective in (141) can be expressed as

$$\begin{aligned} \sum_{n=1}^{N}\sum_{k=1}^{\infty} |\alpha_k^n|^2 \frac{\lambda_k^2}{\lambda_k+\sigma_c^2} &= \mathrm{Tr}\{\mathbf{A}^*\mathbf{\Lambda}\mathbf{A}\} \\ &= \mathrm{Tr}\{\mathbf{A}\mathbf{A}^*\mathbf{\Lambda}\} \\ &= \mathrm{Tr}\{\mathbf{\Lambda}\}, \end{aligned} \tag{144}$$

which is independent of $\mathbf{A}$. Therefore, we conclude that any set of orthonormal sequences $\{\alpha_k^n\}_{k\in\mathbb{Z}}$, $n = 1,\ldots,N$, whose elements vanish for every $k > N$ is optimal.

REFERENCES

[1] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[2] 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.

[3] I. Maravic and M. Vetterli, "Sampling and reconstruction of signals with finite rate of innovation in the presence of noise," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2788–2805, Aug. 2005.

[4] 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.

[5] 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, Jun. 2010.

[6] D. Kandaswamy, T. Blu, and D. van de Ville, "Analytic sensing: Noniterative retrieval of point sources from boundary measurements," *SIAM J. Scientific Computing*, vol. 31, no. 4, pp. 3179–3194, 2009.

[7] 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, Apr. 2011.

[8] 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, Apr. 2011.

[9] T. Michaeli and Y. C. Eldar, "Xampling at the rate of innovation," *IEEE Trans. Signal Process.*, to appear.

[10] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[11] A. Bartov and H. Messer, "Lower bound on the achievable DSP performance for localizing step-like continuous signals in noise," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2195–2201, 1998.

[12] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood, and Cramer-Rao bound," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 720–741, 1989.

[13] J. Berent, P. L. Dragotti, and T. Blu, "Sampling piecewise sinusoidal signals with finite rate of innovation methods," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 613–625, Feb. 2010.

[14] H.-H. Kuo, "White noise theory," in *Handbook of Stochastic Analysis and Applications*, D. Kannan and V. Lakshmikantham, Eds. New York: Marcel Dekker, 2002, ch. 3.

[15] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer, 1981.

[16] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88, pp. 569–587, Apr. 2000.

[17] Y. C. Eldar and T. Michaeli, "Beyond bandlimited sampling," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 48–68, May 2009.

[18] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 2915–2925, Nov. 1994.

[19] 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, Jun. 2006.

[20] M. Unser and T. Blu, "Generalized smoothing splines and the optimal discretization of the Wiener filter," *IEEE Trans. Signal Process.*, vol. 53, no. 6, pp. 2146–2159, 2005.

[21] Y. C. Eldar and M. Unser, "Nonideal sampling and interpolation from noisy observations in shift-invariant spaces," *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2636–2651, Jul. 2006.

[22] S. Ramani, D. Van De Ville, T. Blu, and M. Unser, "Nonideal Sampling and Regularization Theory," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1055–1070, 2008.

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

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

[25] T. G. Dvorkind, Y. C. Eldar, and E. Matusiak, "Nonlinear and nonideal sampling: theory and methods," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5874–5890, Dec. 2008.

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

[27] 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, Apr. 2010.

[28] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 2986–2997, Aug. 2009.

[29] 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, Mar. 2009.

[30] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[31] A. Bruckstein, T. J. Shan, and T. Kailath, "The resolution of overlapping echos," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 6, pp. 1357–1367, 1985.

[32] W. U. Bajwa, K. Gedalyahu, and Y. C. Eldar, "Identification of parametric underspread linear systems and super-resolution radar," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2548–2561, Jun. 2011.

[33] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[34] 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, Jun. 2008.

[35] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[36] J. Shao, *Mathematical Statistics*, 2nd ed. New York: Springer, 2003.

[37] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. New York: Wiley, 1968, vol. 1.

[38] V. I. Sobolev, "Pettis integral," in *Encyclopaedia of Mathematics*, M. Hazewinkel, Ed. Springer, 2001. [Online]. Available: http://eom.springer.de/p/p072490.htm

[39] Z. Ben-Haim and Y. C. Eldar, "The Cramér–Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384–3389, Jun. 2010.

[40] Y. C. Eldar, "Sampling and reconstruction in arbitrary spaces and oblique dual frame vectors," *J. Fourier Analys.*, vol. 1, no. 9, pp. 77–96, Jan. 2003.

[41] A. Aldroubi, C. Leonetti, and Q. Sun, "Error analysis of frame reconstruction from noisy samples," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2311–2325, june 2008.

[42] N. Bi, M. Z. Nashed, and Q. Sun, "Reconstructing signals with finite rate of innovation from noisy samples," *Acta Applicandae Mathematicae*, vol. 107, no. 1, pp. 339–372, 2009.

[43] M. Loève, "Fonctions aléatoires du second ordre," in *Processes Stochastiques et Mouvement Brownien*, P. Lévy, Ed. Paris: Gauthier-Villars, 1948.

[44] ——, *Probability Theory*. Springer, 1978, vol. II.

[45] O. P. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification*. Springer, 2010.

[46] Y. Yamashita and H. Ogawa, "Relative Karhunen–Loève transform," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 371–378, 1996.

[47] Y. Hua and W. Liu, "Generalized Karhunen–Loève transform," *IEEE Signal Process. Lett.*, vol. 5, no. 6, pp. 141–142, 1998.

[48] Y. Hua and T. K. Sarpar, "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, May 1990.

PLACE PHOTO HERE

Zvika Ben-Haim received the B.Sc. degree in electrical engineering and the B.A. degree in physics in 2000, the M.Sc. in electrical engineering in 2005, and the Ph.D. degree in electrical engineering in 2010, all from the Technion—Israel Institute of Technology, Haifa, Israel. From 2000 to 2006, he served as a researcher in the Israel Defense Forces. He has been with the Google Israel R&D Center since 2011.

From 1996 to 2000, he was a member of the Technion Excellence Program. In 2008, he was awarded the Irwin and Joan Jacobs Excellence Scholarship. In 2009, he received the Andrew and Erna Finci Viterbi Fellowship for Ph.D. Students, and in 2010, he was awarded the Wolf Foundation Prize for Outstanding Ph.D. Students. His research is in the field of statistical signal processing, estimation performance bounds, and estimation under sparsity constraints.

PLACE PHOTO HERE

Tomer Michaeli 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 pursuing 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. In 2008 he was awarded the Andrew and Erna Finci Viterbi Fellowship and in 2010 he was a Jacobs–QUALCOMM fellow. He won the Jury Award in 2011 and the Hershel Rich Technion Innovation award in 2012. His research interests include statistical signal processing and sampling theory.

PLACE PHOTO HERE

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 2002.

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 TechnionIsrael Institute of Technology, Haifa. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and a Visiting Professor at Stanford University, Stanford, CA. Her research interests are in the broad areas of statistical signal processing, sampling theory and compressed sensing, optimization methods, and their applications to biology and optics.

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 Technions Award for Excellence in Teaching, 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.