High Rate Quantization Analysis for a Class of Finite Rate of Innovation Signals

Ajinkya Jayawant and Animesh Kumar
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India 400076
ajinkyajayawant,animesh@ee.iitb.ac.in

***Abstract*—Acquisition and perfect reconstruction of finite rate of innovation (FRI) signals was proposed first by Vetterli, Marziliano, and Blu [1]. To the best of our knowledge, the stability of their reconstruction procedure in the presence of scalar quantizers has not been addressed in the literature. For periodic stream of Dirac FRI signal, which is an important subclass of FRI signals, the stability of reconstruction when quantization is introduced on acquired samples is analyzed in this work. It is shown that the parameters of stream of Diracs can be obtained with error $O(\varepsilon)$, where ε is the per sample quantization error. This result holds in the high-rate quantization regime when ε is sufficiently small.**

***Index Terms*—quantization (signal), signal sampling, signal reconstruction, signal analysis**

I. Introduction

The Shannon sampling theorem states that bandlimited signals can be sampled at the Nyquist rate, which is twice the bandwidth of the signal being sampled [2]. Signals such as stream of Dirac signals, which are useful in neuroscience, are not bandlimited and their acquisition and reconstruction method is desirable. Acquisition and *perfect reconstruction* of a periodic stream of Dirac signals was addressed in a seminal paper on sampling of finite rate of innovation (FRI) signals [1]. In their work, Vetterli, Marziliano, and Blu proposed a method to acquire and reconstruct a stream of Dirac signals using $(2K+1)$ samples, where K is the number of Diracs (with different magnitude) present in each period. Their work extended the signal sampling results to other FRI signals such as periodic splines and piecewise polynomials.

Noise and quantization are two impairments that are ubiquitous in signal acquisition and digitization [3]. To the best of our knowledge, the effect of scalar quantization on the acquired samples in the case of FRI signals is *not quantified*. In this work, the effect of scalar quantization is examined for the reconstruction algorithm used by Vetterli Marziliano and Blu [1]. Due to space constraints and for brevity, only the periodic stream of Dirac signals is considered in this work.

A periodic stream of Dirac signals is of the form $x(t) = \sum_{k=1}^{K} c_k \delta(t - t_k)$, with $x(t) = x(t+\tau)$ and $0 < t_1 < t_2 < \ldots < t_K < \tau$. It is assumed that τ is known. The acquisition scheme of this signal requires lowpass filtering of $x(t)$ followed by taking $(2K+1)$ samples [1] of the resultant signal. With L bit uniform scalar quantizers used after sampling, the per sample error will be proportional to 2^{-L} for bounded signals [3]. When quantized samples are subjected to the reconstruction algorithm in [1], an approximate reconstruction of Dirac locations $\vec{t} := \{t_1, \ldots, t_K\}$ and amplitudes $\vec{c} := \{c_1, \ldots, c_K\}$ will be obtained. Let $\widehat{\vec{t}}$ and $\widehat{\vec{c}}$ be the respective approximations for these signal parameters. The main result of this work states that for large L (high-rate)

$$\max_k |t_k - \widehat{t}_k| = O(2^{-L}) \text{ and } \max_k |c_k - \widehat{c}_k| = O(2^{-L}). \quad (1)$$

To the best of our knowledge, the stability of FRI signal acquisition and reconstruction with quantization is not known.

Related work: The FRI sampling and *perfect reconstruction* of a Dirac-stream from its lowpass filtered version was first presented by Vetterli, Marziliano, and Blu [1].

The effect of statistical additive noise on samples and signal reconstruction has been explored extensively. The effect of noise in power-sum series based reconstruction methods has been analyzed for two Diracs by Kusuma and Goyal [4]. FRI signal reconstruction for ensuring robustness to noise have been suggested [5]. Maravić and Vetterli [6] have studied subspace based algorithms that are more robust for combating the effect of additive noise compared to the annihilating filter method. Cramer-Rao bound in the case of single Dirac filtered with B and E-spline kernels have been derived in [7]. With Gaussian noise, [8] uses Gibbs sampling to find the position of the Diracs.

The reconstruction error for stream of Diracs with respect to quantization has been less studied. Barbotin [9] studied the effect of Monte-Carlo and Multiple threshold quantization. The effect of oversampling in frequency and time on the mean squared reconstruction error has been analyzed by Jovanović and Beferull [10]. Tenneti, Kumar and Karandikar [11] study the maximum error in c_k, t_k with a uniform scalar quantizer while using oversampling and resistor-capacitor filters.

In this work, we analyze the maximum possible reconstruction error due to quantization. To the best of our knowledge, the effect of quantization error on the reconstruction scheme in the seminal paper of Vetterli, Marziliano, and Blu has not been presented in the literature.

Notation: We use the notation $\|A\|$ for the spectral norm of a matrix A, and $\|x\|$ is the Euclidean norm of a vector

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

x. The notation $\|A\|_F$ will be used for the Frobenius norm of a matrix A. For a matrix A, its column sum norm will be denoted by $\|A\|_1$ and its row sum norm will be denoted by $\|A\|_\infty$. For a vector x, its l_1 norm will be denoted by $\|x\|_1$ and its l_∞ norm will be denoted by $\|x\|_\infty$. The largest absolute eigenvalue (that is, the spectral radius) of a matrix A will be denoted by $\rho(A)$. The singular values of a matrix A of dimension $n \times n$ will be denoted by $\sigma_1(A), \ldots, \sigma_n(A)$ such that $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_n(A)$. All these norms are well understood from the literature [12]. The inner-product of two complex-valued vectors v_1 and v_2 will be represented by $\langle v_1, v_2 \rangle := \bar{v}_1^T v_2$. The symbol j will be used for $\sqrt{-1}$.

II. Signal and system model

This section presents the modeling assumptions on stream of Dirac FRI signals, their sampling, and quantization.

A. Signal model

It will be assumed that the signal of interest is given by

$$x(t) = \sum_{k=1}^{K} c_k \delta(t - t_k) \tag{2}$$

where the parameters $t_1, \ldots, t_K \in [0, 1]$ and $c_1, \ldots, c_K \in \mathbb{R}$. It is further assumed that $0 < t_1 < t_2 < \ldots < t_K < 1$ and $0 < c_{\min} \leq c_k \leq c_{\max} < \infty$ for every $1 \leq k \leq K$. This model represents periodic stream of delta signals with period $\tau = 1$ [1]. The condition $0 < c_{\min} \leq c_k$ is needed since very small c_k will be not resolved due to quantization. For resolvability after quantization, it will also be assumed that

$$t_{i+1} - t_i \geq \Delta_1 > 0 \tag{3}$$

where $1 \leq i \leq K$ and $t_{K+1} = 1 + t_1$. Loosely speaking, these assumptions mean that the Diracs are not too close (modulo the period), not too far, not too tall, and not too short! All these assumptions are required to ensure that there is stability in error after quantization. Note that $t_1, \ldots, t_K$ and $c_1, \ldots, c_K$ are $2K$ unknown real numbers, and these have to be estimated so as to reconstruct the original signal $x(t)$.

B. The sampling model

The signal $y(t) := x(t) \star h(t)$ will be sampled near its rate of innovation [1] with the sampling period T given by

$$\frac{1}{T} = (2K + 1) \tag{4}$$

with sampling locations

$$s_i = \frac{i}{2K + 1}, i = 0, 1, \ldots, (2K) \tag{5}$$

where, K is the number of Diracs present in the signal $x(t)$. The filter $h(t)$ is ideal-lowpass (see [1]) with bandwidth $B = 2K\pi$.

C. The quantization model

The filtered signal $y(t) = x(t) \star h(t)$ is sampled to obtain $y(s_0), y(s_1), \ldots, y(s_{2K})$. These samples are quantized through an L-bit uniform scalar quantizer to obtain the quantized bits $\widehat{y}(s_0), \widehat{y}(s_1), \ldots, \widehat{y}(s_{2K})$ [3]. It is assumed that for a given Δ_1, K, the parameter $c_{\max}$ is such that $|y(t)| \leq 1$ in the sampling interval $[0, 1]$. A trivial bound on $c_{\max}$ is $1/(K(2K + 1))$ which follows by the triangle inequality on Fourier series coefficients of $x(t)$. If $|y(t)| \leq 1$, then

$$|\widehat{y}(s_k) - y(s_k)| \leq 2^{-L}, \quad k = 0, 1, \ldots, 2K. \tag{6}$$

The main point is that each sample is available with exponential accuracy in the quantizer precision. Finally, the reconstruction is decided to be satisfactory based on the error in the positions of the Diracs if

$$|t_k - \widehat{t}_k| \leq \frac{\Delta_1}{2}. \tag{7}$$

If this condition is not satisfied then the positions of Diracs lose their meaning as they could get confused with each other.

III. Quantization error analysis

The steps in the reconstruction of stream of Diracs are illustrated in Fig. 1 while using the algorithm of Vetterli, Marziliano, and Blu [1]. Accordingly, block A of Fig. 1 is analyzed first. Later subsections (of this section) analyze blocks B, C, and D of Fig. 1.

A. Estimation of Fourier series coefficients of $x(t)$

The signal $x(t)$ with a period $\tau = 1$ has a Fourier series

$$x(t) = \sum_{m=-\infty}^{\infty} X[m] \exp(j2\pi mt). \tag{8}$$

The lowpass filtered signal $y(t)$ will retain $2K + 1$ Fourier series coefficients and the samples $y(s_n)$ are given by

$$y(s_n) = \sum_{m=-K}^{K} X[m] \exp\left(\frac{j2\pi mn}{2K + 1}\right). \tag{9}$$

Let $p = \exp\left(\frac{j2\pi}{2K+1}\right)$, where p has magnitude one. Then,

$$\begin{bmatrix} y(0) \\ y(s_1) \\ \vdots \\ y(s_{2K}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ p^{-K} & p^{-(K-1)} & \cdots & p^{K} \\ \vdots & \vdots & \ddots & \vdots \\ p^{-2K^2} & p^{-2K(K-1)} & \cdots & p^{2K^2} \end{bmatrix} \begin{bmatrix} X[-K] \\ X[-K+1] \\ \vdots \\ X[K] \end{bmatrix},$$

or simply $\vec{y} = P\vec{X}$ where the notation is obvious. This equation is invertible since the matrix P has a Vandermonde structure. Let $\widehat{\vec{y}} := [\widehat{y}(0), \ldots, \widehat{y}(s_{2K})]^T$ be the quantized observations. Then the quantization error vector is,

$$\vec{\varepsilon}_y := \widehat{\vec{y}} - \vec{y} \text{ with } \|\vec{\varepsilon}_y\|_\infty \leq 2^{-L}. \tag{10}$$

The invertible relation between $\vec{X}$ and $\vec{y}$ is used to find $\widehat{\vec{X}}$ as $\widehat{\vec{X}} = P^{-1}\widehat{\vec{y}}$. The l_2 error of $\vec{\varepsilon}_X := \widehat{\vec{X}} - \vec{X}$ is bounded as

$$\|\vec{\varepsilon}_X\| = \|P^{-1}\vec{\varepsilon}_Y\| \leq \|P^{-1}\|\|\vec{\varepsilon}_Y\| \tag{11}$$

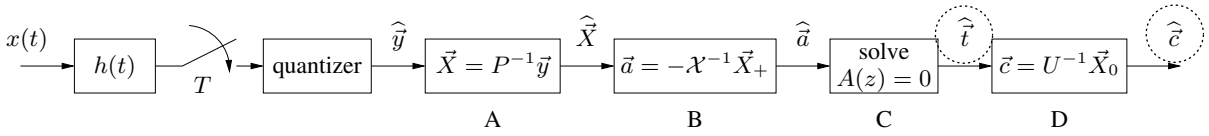

Fig. 1. *The reconstruction algorithm for periodic stream of Dirac signals from acquired samples [1] will be subjected to quantization in this work. From* (2)*, the location vector* $\vec{t}$ *and their amplitudes* $\vec{c}$ *determine the signal. Due to addition of the quantizer, their estimates* $\widehat{\vec{t}}$ *and* $\widehat{\vec{c}}$ *will be obtained.*

where the last inequality follows from [13]. The term $\|P^{-1}\|$ is bounded in terms of its smallest singular value

$$\|P^{-1}\| = \sigma_1(P^{-1}) = \frac{1}{\sigma_{2K+1}(P)}. \tag{12}$$

From [14], the smallest singular value of P is bounded as

$$\sigma_{2K+1}(P) \geq \left(\frac{2K}{\|P\|_F^2}\right)^K |\det(P)|. \tag{13}$$

Since all the entries of P have a magnitude 1, so its Frobenius norm squared is $\|P\|_F^2 = (2K+1)^2$. A lower bound on $|\det(P)|$ will be evaluated next. The matrix P is a Vandermonde matrix with second-row entries as $p^{-K}, p^{-(K-1)}, \cdots, p^K$, which are equidistant on the unit circle. After the computation of $|\det(P)|$, the algebraic details of which are omitted,

$$\|\vec{\varepsilon}_X\| \leq \left(\frac{2K+1}{2K}\right)^K \frac{1}{\sqrt{2K+1}} \|\vec{\varepsilon}_Y\| \tag{14}$$

In summary, the error-norm of the Fourier series coefficients is at most linear in the norm of the quantization errors.

B. Estimation of the annihilating filter

With $(2K+1)$ samples, the solution for the annihilating filter is found by solving the equations

$$\begin{bmatrix} X[0] & X[-1] & \cdots & X[-K+1] \\ X[1] & X[0] & \cdots & X[-K+2] \\ \vdots & \vdots & \ddots & \vdots \\ X[K-1] & X[K-2] & \cdots & X[0] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[K] \end{bmatrix} = - \begin{bmatrix} X[1] \\ X[2] \\ \vdots \\ X[K] \end{bmatrix}$$

or simply $\mathcal{X}\vec{a} = -\vec{X}_+$ where the notation is obvious. The $+$ mark in $\vec{X}_+$ signifies that only the positive index Fourier series coefficients are present. From Section III-A, $\widehat{\mathcal{X}}$ and $\widehat{\vec{X}}$ are available. This results in an estimate for $\vec{a}$ as follows:

$$\widehat{\mathcal{X}}\widehat{\vec{a}} = -\widehat{\vec{X}}_+. \tag{15}$$

Subsequently, $\widehat{\vec{a}}$ can be obtained if $\widehat{\mathcal{X}}$ is invertible. Let $\vec{\varepsilon}_a := \widehat{\vec{a}} - \vec{a}$. By (52), Appendix A we get,

$$\|\vec{\varepsilon}_a\| < 2\left(\|\mathcal{X}^{-1}\|\|\vec{\varepsilon}_{X_+}\| + \|\mathcal{X}^{-1}\|\|\mathcal{X} - \widehat{\mathcal{X}}\|\|\vec{a}\|\right) \tag{16}$$

It is also noted that

$$\|\widehat{\mathcal{X}} - \mathcal{X}\| \leq \sqrt{K}\|\vec{\varepsilon}_X\| \tag{17}$$

by solving the optimization problem

$$\max_{\|\vec{\varepsilon}_X\| = \varepsilon} K|\varepsilon_X[1]|^2 + \sum_{i=2}^{K} (K-i+1)[|\varepsilon_X[-i]|^2 + |\varepsilon_X[i]|^2]$$

where ε is an arbitrary constant. The above-mentioned cost function is essentially the Frobenius norm of $\widehat{\mathcal{X}} - \mathcal{X}$. Using $\|\vec{\varepsilon}_{X_+}\| \leq \|\vec{\varepsilon}_X\|$ and (17) in (16) results in

$$\|\vec{\varepsilon}_a\| \leq 2\Big(1 + \sqrt{K}\|\vec{a}\|\Big)\|\mathcal{X}^{-1}\|\|\vec{\varepsilon}_X\| \tag{18}$$

The term $\|\vec{a}\|$ is bounded since each $|a_k| \leq \binom{K}{k}$. In deriving this bound, $|u_k| = 1$ is utilized and the fact that a_k is the coefficient of z^{K-k} in $(z-u_1)(z-u_2)\ldots(z-u_K)$ (see (26)).

C. Invertibility of $\widehat{\mathcal{X}}$

In this section, a bound on the quantization precision L will be established such that $\widehat{\mathcal{X}}$ is invertible. From (2) and (8), the (p,q)-element of $\mathcal{X}$ is given by

$$\mathcal{X}_{p,q} = \sum_{k=1}^{K} c_k \exp(-j2\pi(p-q)t_k) \tag{19}$$

for $1 \leq p, q \leq K$. This matrix can be factorized as

$$\mathcal{X} = VD\bar{V}^T \tag{20}$$

where $\bar{V}^T$ is Hermitian of V, $D = \text{diag}(c_1, c_2, \ldots, c_K)$ and

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{-j2\pi t_1} & e^{-j2\pi t_2} & \cdots & e^{-j2\pi t_K} \\ \vdots & \vdots & \vdots & \vdots \\ e^{-j2\pi(K-1)t_1} & e^{-j2\pi(K-1)t_2} & \cdots & e^{-j2\pi(K-1)t_K} \end{bmatrix}.$$

Since $|c_i| \neq 0$ and $t_1, \ldots, t_K$ are distinct by assumption, V and D, and $\mathcal{X}$ are invertible. From the properties of matrices [13],

$$\sigma_K(\mathcal{X}) \geq [\sigma_K(V)]^2 \min_{1 \leq k \leq K} |c_k| \tag{21}$$

since the singular values of V and $\bar{V}^T$ are identical. From [15] and its results on the inverses of Vandermonde matrices,

$$\|V^{-1}\|_\infty < \max_k \prod_{v \neq k} \frac{\left(1 + |e^{-j2\pi t_v}|\right)}{|e^{-j2\pi t_v} - e^{-j2\pi t_k}|} \tag{22}$$

Note that $\prod_{v \neq k} |e^{-j2\pi t_v} - e^{-j2\pi t_k}|$ is minimum when for each i, $t_i - t_{i-1} = \Delta_1$ and $k = \lfloor K/2 \rfloor$. As a result,

$$\min_k \prod_{v \neq k} |e^{-j2\pi t_v} - e^{-j2\pi t_k}| \geq \prod_{v=1}^{\lfloor \frac{K}{2} \rfloor} 4\sin^2(\pi v \Delta_1). \tag{23}$$

Since $\|V^{-1}\| = [\sigma_K(V)]^{-1} \leq \sqrt{K}\|V^{-1}\|_\infty$, therefore,

$$\sigma_K(V) \geq \left[\prod_{v=1}^{\lfloor \frac{K}{2} \rfloor} 4\sin^2(\pi v \Delta_1)\right] \frac{1}{2^{K-1}\sqrt{K}} \tag{24}$$

Upon substitution of (24) in (21), we get

$$\sigma_K(\mathcal{X}) \geq \left[\left(\prod_{v=1}^{\lfloor \frac{K}{2} \rfloor} 4\sin^2(\pi v \Delta_1) \right) \frac{1}{2^{K-1}\sqrt{K}} \right]^2 c_{\min}. \quad (25)$$

The invertibility of $\widehat{\mathcal{X}}$ will be argued formally next. Let A and B be $n \times n$ matrices, and A be invertible. If $\sigma_1(B) < \sigma_n(A)$, then $A+B$ will be invertible. By using the result for $\mathcal{X}$ and $\widehat{\mathcal{X}} - \mathcal{X}$, the matrix $\widehat{\mathcal{X}}$ is invertible if $\|\widehat{\mathcal{X}} - \mathcal{X}\| \leq \sigma_K(\mathcal{X})$. For deriving (16), it was assumed that $\|\widehat{\mathcal{X}} - \mathcal{X}\| \leq \sigma_K(\mathcal{X})/2$. *This conditions results in a lower bound on L, which we omit for brevity*. This is the reason for 'high-rate' in this paper's title.

D. Error in Dirac locations and amplitudes

In this section, the error analysis for the positions and the amplitudes of the Diracs in $x(t)$ will be presented. From $\widehat{\vec{a}}$ obtained in (15) using the quantized samples, the approximate locations $\widehat{\vec{t}}$ will be obtained. Note that $a[0] = \widehat{a}[0] = 1$. Let

$$z^K + \widehat{a}[1]z^{K-1} + \ldots + \widehat{a}[K] = \prod_{k=1}^{K} (z - \widehat{u}_k) \quad (26)$$

be the annihilation filter's factors. The technique of Galántai and Hegedűs will be used to establish a bound on $\|\vec{u} - \widehat{\vec{u}}\|$ [16]. The companion matrix used in [16] for the annihilator polynomial is

$$C = \begin{bmatrix} -\widehat{a}[1] & \cdots & -\widehat{a}[K-1] & -\widehat{a}[K] \\ 1 & \ldots & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & \ldots & 1 & 0 \end{bmatrix} \quad (27)$$

and can be factorized as $\widehat{C} = \Pi \widehat{V} \widehat{U} \widehat{V}^{-1} \Pi^T$, where $\Pi = [e_K | e_{K-1} | \cdots | e_1]$, $\widehat{U} = \text{diag}(\widehat{u}_1, \widehat{u}_2, \ldots, \widehat{u}_K)$, and V is as in (20) with $\widehat{u}_1, \ldots, \widehat{u}_K$ generating the Vandermonde matrix. Similar factorization is there for exact roots $\vec{u}$ and the companion matrix made from $\vec{a}$. From [16, Theorem 4],

$$|u_k - \widehat{u}_k| \leq \|V^{-1} e_K\| \|V\| \|\vec{\varepsilon}_a\| \quad (28)$$

where $1 \leq k \leq K$. From [16], $V^{-1}e_K$ is given by

$$V^{-1}e_K = \begin{bmatrix} \frac{1}{(u_1-u_2)(u_1-u_3)\cdots(u_1-u_K)} \\ \cdots \\ \frac{1}{(u_K-u_1)(u_K-u_2)\cdots(u_K-u_{K-1})} \end{bmatrix}. \quad (29)$$

By similar calculations as in (23), it follows that

$$\|V^{-1}e_K\| \leq \frac{\sqrt{K}}{\prod_{v=1}^{\lfloor \frac{K}{2} \rfloor} 4\sin^2(\pi v \Delta_1)} \quad (30)$$

The norm $\|V\|$ is bounded using row and column norms [17]:

$$\|V\| \leq \sqrt{\|V\|_1 \|V\|_\infty} = \sqrt{K \cdot K} = K \quad (31)$$

since each element of V has magnitude 1. From (28),

$$\max_k |u_k - \hat{u}_k| \leq \frac{\sqrt{K}K}{\prod_{v=1}^{\lfloor \frac{K}{2} \rfloor} 4\sin^2(\pi v \Delta_1)} \|\vec{\varepsilon}_a\| \quad (32)$$

1) Bound on location error: From $\widehat{\vec{a}}$, the Dirac locations $\widehat{\vec{t}}$ will be obtained. The roots of the polynomial in (26) may not lie on the unit circle. Each approximate root will be mapped on the unit circle along a ray originating at the center of the unit circle and passing through the approximate root. Let the mapped approximate root be $\widehat{w}_k := e^{-j2\pi \widehat{t}_k}$. Since $|\widehat{w}_k - u_k| \leq |\widehat{u}_k - u_k|$ and $|\widehat{w}_k - u_k| = |e^{-j2\pi \widehat{t}_k} - e^{-j2\pi t_k}| = 2|\sin(\pi(\widehat{t}_k - t_k))|$, therefore,

$$|\widehat{t}_k - t_k| \leq \frac{1}{\pi} \sin^{-1}(\max_k |u_k - \widehat{u}_k|/2) \quad (33)$$

Maximizing the left hand side over k, and with $\sin^{-1}(x) \leq 2x$ for $0 \leq x < \frac{1}{2}$,

$$\max_k |\widehat{t}_k - t_k| \leq \frac{1}{\pi} \max_k |\widehat{u}_k - u_k|. \quad (34)$$

2) Error in amplitudes: Approximate values in $\widehat{t}$ result in approximate amplitudes $\widehat{c}$ (of Diracs in $x(t)$) by a Vandermonde matrix [1]

$$\begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[K-1] \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ u_1 & \cdots & u_{K-1} \\ \vdots & \ddots & \vdots \\ u_1^{K-1} & \cdots & u_{K-1}^{K-1} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix} \quad (35)$$

or simply $\vec{X}_0 = V\vec{c}$, where the notation is obvious. The matrix V is same as in (20) and recall that $u_k = e^{-j2\pi t_k}$. The approximate version of this relation is

$$\widehat{\vec{X}}_0 = \widehat{V}\widehat{\vec{c}} \quad (36)$$

with

$$\widehat{V} = \begin{bmatrix} 1 & \cdots & 1 \\ e^{-j2\pi \widehat{t}_1} & \cdots & e^{-j2\pi \widehat{t}_K} \\ \vdots & \ddots & \vdots \\ e^{-j2\pi(K-1)\widehat{t}_1} & \cdots & e^{-j2\pi(K-1)\widehat{t}_K} \end{bmatrix} \quad (37)$$

Using (52) in (36), we get

$$\|\widehat{\vec{c}} - \vec{c}\| \leq 2\|V^{-1}\| \|\widehat{\vec{X}}_0 - \vec{X}_0\| + 2\|V^{-1}\| \|\widehat{V} - V\| \|V^{-1}\vec{X}_0\| \quad (38)$$

Using the row and column norms [17],

$$\sigma_1(\widehat{V} - V) = \|\widehat{V} - V\| \leq \sqrt{\|\widehat{V} - V\|_1 \|\widehat{V} - V\|_\infty} \quad (39)$$

Note that $|u_k^l - \widehat{u}_k^l| = 2|\sin(l\pi(t_k - \widehat{t}_k)|$ for $l = 1, 2, \ldots, K$. Since $|\sin(\theta)| \leq |\theta|$, therefore,

$$\|\widehat{V} - V\|_1 \leq 2 \max_k \sum_{l=1}^{K-1} |\sin(l\pi(\widehat{t}_k - t_k)| \quad (40)$$

$$\leq K(K-1)\pi \max_k |\widehat{t}_k - t_k|. \quad (41)$$

For $\|\widehat{V} - V\|_\infty$, it is observed that sum of any row l in the matrix is bounded by

$$\sum_{k=1}^{K} |\sin(l\pi(t_k - \widehat{t}_k))| \leq \sum_{k=1}^{K} \pi l |t_k - \widehat{t}_k| \quad (42)$$

$$\leq \pi K(K-1) \max_k |t_k - \widehat{t}_k|. \quad (43)$$

Therefore, we get an upper bound as follows:

$$\|\widehat{V} - V\| \leq \pi K(K-1) \max_k |t_k - \widehat{t}_k|. \quad (44)$$

Finally, the derivation of $\max_k|t_k - \widehat{t}_k| = O(2^{-L})$ and $\max_k|c_k - \widehat{c}_k| = O(2^{-L})$ will be argued. From (10), $\|\vec{\varepsilon}_y\| \leq (\sqrt{2K+1})2^{-L}$. From this observation, (14), (18), and (32), it follows that

$$\max_k |\widehat{u}_k - u_k| = O(2^{-L}). \quad (45)$$

The first main result follows from the above and (34). Since $\vec{X}_0$ is a subset of $\vec{X}$, therefore $\|\vec{\varepsilon}_{X_0}\| \leq \|\vec{\varepsilon}_X\|$. From Section III-C, $\|V^{-1}\|$ is also bounded. From these observations, (44) and (38)

$$\|\widehat{\vec{c}} - \vec{c}\| = O(2^{-L}) \quad (46)$$

which yields the second part of the main result as $\|\vec{c}\|_\infty \leq \|c\|$.

IV. Conclusions

For periodic stream of Dirac FRI signals, the stability of reconstruction when quantization is introduced on acquired samples was analyzed in this work. It was shown that the parameters of stream of Diracs can be obtained with error $O(2^{-L})$, where L was the number of bits used in quantizing each sample. Analysis in Section III reveals that this result holds when L is sufficiently large, that is, in high-rate quantizer regime. Bounds on L and extension of this result to periodic splines and piecewise polynomials are interesting topics for future work.

Appendix A
Matrix perturbation error bound

Consider the system $\vec{y} = A\vec{x}$, where $\vec{y}$ and $\vec{x}$ are $n \times 1$, and A is $n \times n$. Let $\widehat{A}$ and $\widehat{\vec{y}}$ be the approximate versions. Then, $\widehat{\vec{x}} = (\widehat{A})^{-1}\widehat{\vec{y}}$. The approximation error in $\vec{x}$ is given by

$$\vec{\varepsilon}_x = (\widehat{A})^{-1}\widehat{\vec{y}} - A^{-1}\vec{y} \quad (47)$$

$$= \Big(I + A^{-1}(\widehat{A} - A)\Big)^{-1} A^{-1}\widehat{\vec{y}} - A^{-1}\vec{y} \quad (48)$$

Let $\vec{\varepsilon}_y := \widehat{\vec{y}} - \vec{y}$ and $\varepsilon_A = \widehat{A} - A$. If $\sigma_1(\varepsilon_A) < \sigma_n(A)/2$, then $\widehat{A}$ has an inverse and $\Big(I + A^{-1}(\widehat{A} - A)\Big)^{-1} = I - A^{-1}(A - \widehat{A}) + R$ where

$$\|R\| \leq \|A^{-1}\|\|\widehat{A} - A\| \quad (49)$$

Therefore,

$$\vec{\varepsilon}_x = \Big(I - A^{-1}(A - \widehat{A}) + R\Big)A^{-1}\widehat{\vec{y}} - A^{-1}\vec{y} \quad (50)$$

$$= A^{-1}\vec{\varepsilon}_y - A^{-1}\varepsilon_A A^{-1}\vec{y} - A^{-1}\varepsilon_A A^{-1}\vec{\varepsilon}_y + RA^{-1}\vec{y} + RA^{-1}\vec{\varepsilon}_y \quad (51)$$

Now we use triangle inequality and sub-multiplicativity properties for matrix norms [12], on $\|\vec{\varepsilon}_x\|$ and use (49) to get

$$\|\vec{\varepsilon}_x\| < 2\|A^{-1}\|\|\vec{\varepsilon}_y\| + 2\|A^{-1}\|\|\varepsilon_A\|\|A^{-1}\vec{y}\|. \quad (52)$$

Appendix B
Matrix power series

If a matrix B has spectral radius $\rho(B) < 1$, then the power series of $(I+B)^{-1}$ converges absolutely [18]. As a result,

$$(I+B)^{-1} - (I-B) = B^2 - B^3 + \ldots. \quad (53)$$

Therefore,

$$\|(I+B)^{-1} - (I-B)\| \leq \|B\|^2 + \|B\|^3 + \ldots. \quad (54)$$

If $\|B\| < 1$, then $\|B\|^2 + \|B\|^3 + \ldots$ converges, while if $\|B\| < 1/2$ then $\|B\|^2 + \|B\|^3 + \ldots < \|B\|$. As a result, if $\|B\| < 1/2$ (which also implies that $\rho(B) \leq \|B\| < 1/2$) then

$$\|(I+B)^{-1} - (I-B)\| \leq \|B\|. \quad (55)$$

References

[1] M. Vetterli, P. Marziliano, and T. Blu, "Sampling Signals with Finite Rate of Innovation," *IEEE Trans. Signal Proc.*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[2] R. J. Marks, II, *Introduction to Shannon Sampling and Interpolation Theory*. New York, USA: Springer-Verlag, 1990.

[3] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic, 1992.

[4] J. Kusuma and V. Goyal, "On the accuracy and resolution of powersum-based sampling methods," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 182 –193, Jan. 2009.

[5] A. Ridolfi, I. Maravic, J. Kusuma, and M. Vetterli, "Sampling signals with finite rate of innovation: the noisy case," Infoscience, EPFL, Tech. Rep., 2002.

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

[7] P. L. Dragotti and F. Homann, "Sampling signals with finite rate of innovation in the presence of noise," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 2941–2944.

[8] V. Tan and V. Goyal, "Estimating signals with finite rate of innovation from noisy samples: A stochastic algorithm," *IEEE Trans. on Signal Processing*, vol. 56, no. 10, pp. 5135–5146, 2008.

[9] Y. Barbotin, "Finite rate of innovation sampling technics for embedded uwb devices," Ph.D. dissertation, Swiss Federal Institute of Technology (EPFL), Mar. 2009. [Online]. Available: http://infoscience.epfl.ch/record/141997

[10] I. Jovanović and B. Beferull-Lozano, "Oversampled a/d conversion and error-rate dependence of nonbandlimited signals with finite rate of innovation," *IEEE Trans. Signal Proc.*, vol. 54, no. 6, Jun. 2006.

[11] S. Tenneti, A. Kumar, and A. Karandikar, "Finite rate of innovation signals: Quantization analysis with resistor-capacitor acquisition filter," in *Proc. of the 10th International Conference on Sampling Theory and Applications (SampTA)*, Jul. 2013.

[12] G. Golub and C. V. Loan, *Matrix Computations*. JHU Press, 2012, vol. 3.

[13] J. Merikoski and R. Kumar, "Inequalities for spreads of matrix sums and products," *Applied Mathematics E-Notes*, vol. 4, pp. 150–159, 2004.

[14] G. Piazza and T. Politi, "An upper bound for the condition number of a matrix in spectral norm," *Journal of Computational and Applied Mathematics*, vol. 143, pp. 141–144, Jun. 2002.

[15] W. Gautschi, "On inverses of vandermonde and confluent vandermonde matrices," *Numerische Mathematik*, vol. 4, no. 1, pp. 117–123, 1962.

[16] A. Galántai and C. J. Hegedűs, "Perturbation bounds for polynomials," *Numerische Mathematik*, vol. 109, no. 1, pp. 77–100, 2008.

[17] R. Horn and C. Johnson, "Topics in matrix analysis," *Cambridge University Press, Cambridge*, vol. 37, p. 39, 1991.

[18] N. J. Young, "The rate of convergence of a matrix power series," *Linear Algebra and its Applications*, vol. 35, pp. 261–278, 1981.