Single-Bit Oversampled A/D Conversion with Exponential Accuracy in the Bit-Rate

Zoran Cvetković
AT&T Shannon Laboratory

Ingrid Daubechies
Princeton University

Abstract

We present a scheme for simple oversampled analog-to-digital conversion, with single bit quantization and exponential error decay in the bit-rate. The scheme is based on recording positions of zero-crossings of the input signal added to a deterministic dither function. This information can be represented in a manner which requires only logarithmic increase of the bit rate with the oversampling factor, r. The input band-limited signal can be reconstructed from this information locally, and with a mean squared error which is inversely proportional to the square of the oversampling factor, MSE $= O(1/r^2)$. Consequently, the mean squared error of this scheme exhibits exponential decay in the bit-rate.

1 Introduction

Analog-to-digital (A/D) conversion involves discretization of an analog continuous-time signal in both time and amplitude. In the simple analog-to-digital conversion considered here, the time discretization is performed as regular sampling with an interval τ, and uniform scalar quantization with a step q is used for amplitude discretization. The accuracy of this conversion scheme depends both on the resolution of the discretization in time and the resolution of the discretization in amplitude; it is commonly studied using statistical analysis, and assuming linear reconstruction, which amounts to low-pass filtering of the sequence of quantized samples with cut-off frequency equal to the signal bandwidth. For most practical purposes the signals involved can be assumed to be bandlimited; in the conversion these signals are sampled above their Nyquist rate, so the time discretization is reversible. The amplitude discretization, however, introduces an irreversible loss of information; the error made in analog-to-digital conversion is therefore referred to as quantization error. In the 1940's, Bennett showed that the quantization error can be well modeled as a white noise independent of the input, provided the quantization step is small enough, and a sufficiently large number of quantization levels is available to ensure that the quantizer does not overload [1]. According to the white noise model the mean squared error between a signal f and the corresponding signal f_r, obtained with the linear reconstruction, behaves as

$$E(|f(t) - f_r(t)|^2) = \frac{1}{12} q^2 \frac{\tau}{\tau_N} \tag{1}$$

where τ_N is the Nyquist sampling interval. This formula suggests that the conversion accuracy can be improved beyond the precision of the quantizer by introducing oversampling; because of the costs involved in building high precision quantizers, modern techniques for high accuracy analog-to-digital conversion are based on oversampling. However, apart from its technological and economical justification, using oversampling to improve the conversion accuracy is considered to be dramatically inferior, in rate-distortion sense, to refining quantization. For a given precision of the quantizer, the bit-rate of oversampled A/D conversion, unless some entropy coding is used, is inversely proportional to the sampling interval, $R = O(1/\tau)$; hence, the mean squared error decays only inversely to the bit-rate, $E(|f(t)-f_r(t)|^2) = O(1/R)$. On the other hand, if the sampling interval remains fixed but the quantization step is refined, the mean squared error decays exponentially in the bit-rate, $E(|f(t)-f_r(t)|^2) = O(2^{-2R})$.

Only recently has it been demonstrated that the accuracy of oversampled A/D conversion is better than suggested by formula (1), and that even without tedious entropy coding the quantized samples can be efficiently represented so that an exponentially decaying rate-distortion characteristic can be attained [2, 3, 4]. The deterministic analysis of oversampled A/D conversion in [2] and [3] reveals that increasing oversampling provides more precise information about the locations of quantization threshold crossings of the input analog signal and that this information describes the signal with higher accuracy than indicated by formula (1). In particular, the deterministic analysis shows that if the quantization threshold crossings of a signal f form a sequence of stable sampling for the space of square integrable bandlimited signals to which f belongs, and f_r is a square integrable function, in that same bandlimited space, for which the oversampled A/D conversion produces the *same* digital sequence as for f, then the squared error between f and f_r can be bounded as

$$|f(t) - f_r(t)|^2 < c_f \left(\frac{\tau}{\tau_N}\right)^2, \tag{2}$$

uniformly in time. Moreover, representing the information about the quantization threshold crossings requires only a logarithmic increase of the bit-rate as the sampling interval tends to zero [4]. Hence, using this quantization threshold crossings based representation the rate-distortion characteristic of oversample A/D conversion becomes

$$|f(t) - f_r(t)|^2 < a_f e^{-\alpha R}, \tag{3}$$

where α is a positive constant, and a_f is proportional to the factor c_f in (2) [4].

While these results provided a radically new perspective on oversampled A/D conversion, some important issues remained unanswered. The τ^2 conversion accuracy is established under the assumption that the quantization threshold crossings of the input signal constitute a sequence of stable sampling in an appropriate class of bandlimited signals. Although there are bandlimited signals which have this property, there also exist bandlimited signals for which the threshold crossings are too sparse to ensure this conversion accuracy; giving a precise characterization of these two classes of bandlimited signals is very intricate [3]. Another problem is that the stronger error bound in (2) is not uniform on sufficiently general compact sets for simple oversampled A/D conversion in its standard form. Note also that the result about $O(\tau)$ accuracy with the linear reconstruction is valid only in a small range of the discretization parameters q and τ, and that the error does not tend to zero along with the sampling interval but rather reaches a floor level for some finite τ [2]. So, there are basically no sufficiently general results about the accuracy of simple oversampled analog-to-digital conversion. Moreover, no explicit algorithms for reconstructing a bandlimited signal with τ^2 accuracy are known, and the feasibility of local reconstruction with this accuracy is not clear. This paper addresses and basically answers all of these issues for a single-bit oversampled A/D conversion scheme with a deterministic dither. The purpose of the dither functions is to enforce a sufficiently dense sequence of quantization threshold crossings so that the $O(\tau^2)$ accuracy and exponential rate-distortion characteristic is guaranteed for *all* bandlimited functions with amplitude bounded by a pre-set constant, and that uniformly on compact sets. Moreover, we prove the existence of local reconstruction algorithms which lead to good error and stability estimates. The details of the proposed 1-bit quantization scheme are given in Section 2 below. In Section 3 we show how the quantized information allows for stable reconstruction of a good approximation to the input signal, and we provide error estimates. Section 4 gives a short preliminary discussion of concrete reconstruction schemes.

2 A dithered A/D conversion scheme

In this section we study a simple single-bit analog-to-digital conversion with a deterministic dither. We apply this conversion scheme to signals that belong to the set $\mathcal{C}$, that is the set of π-bandlimited signals with finite energy and amplitude smaller than 1, $\mathcal{C} = \{f : f \in \mathcal{V}_\pi,\ \|f\|_\infty \leq 1\}$. Here we use $\mathcal{V}_\pi$ to denote the space of square integrable π-bandlimited signals, $\mathcal{V}_\pi = \{f : f \in L^2(I\!\!R), \hat{f}(\omega) = 0\ |\omega| > \pi\}$, and we use $\hat{f}$ to denote the Fourier transform of f.

The single-bit dithered A/D converter that we shall apply to f in $\mathcal{C}$ is defined by means of a dither function d and a parameter $\lambda > 1$. We shall assume that the dither function satisfies some special conditions, ensuring that the composite signal $f + d$ changes sign frequently. In particular, we shall require that d is a C^1-function that satisfies moreover, for all $n \in Z\!\!\!Z$,

$$\left|d\left(\frac{n}{\lambda} + \frac{1}{2\lambda}\right)\right| \geq \gamma > 1 \quad , \quad \operatorname{sgn}\left[d\left(\frac{n}{\lambda} - \frac{1}{2\lambda}\right)\right] = -\operatorname{sgn}\left[d\left(\frac{n}{\lambda} + \frac{1}{2\lambda}\right)\right] \quad . \tag{4}$$

An example of an appropriate dither is the sine function, $d(t) = \gamma \sin(\lambda\pi t)$. Since the sequence $|(f+d)(n/\lambda+1/(2\lambda))|$, for $n \in \mathbb{Z}$, is bounded below by $\gamma-1$, and alternates in sign, there must be at least one zero-crossing in every interval $(n/\lambda-1/(2\lambda), n/\lambda+1/(2\lambda))$. We can therefore select one zero-crossing t_n in every interval; the resulting sequence $(t_n)_{n\in\mathbb{Z}}$ is sufficiently dense to form a sequence of stable sampling in $\mathcal{V}_\pi$. This motivates the following definition:

Definition 2.1 *Let d be a bounded C^1-function satisfying (4), and let λ be a fixed parameter satisfying $\lambda > 1$. The single-bit dithered oversampled analog-to-digital converter, $\mathbf{D}^d_{\lambda,\tau}$, is defined as the operator $\mathbf{D}^d_{\lambda,\tau} : \mathcal{C} \to \ell^\infty(\mathbf{Z})$ given by*

$$(\mathbf{D}^d_{\lambda,\tau}f)[n] = \min\{m : m \in \mathbb{Z}, m\tau \in I_n, \text{sgn}[(f+d)(m\tau)] \neq \text{sgn}[(f+d)(m\tau+\tau)]\} - \mu_n$$

where $I_n = (n/\lambda - 1/(2\lambda), n/\lambda + 1/(2\lambda))$, and $\mu_n = \lfloor n/\lambda\tau \rfloor$.

Remarks: 1. The output of the converter is a sequence of indices of sampling intervals where zero-crossings of the dithered signal $f + d$ occur; one zero-crossing within each interval I_n. For simplicity, in the definition of the converter we choose this to be the first zero-crossing in I_n, but any other selection algorithm would work as well. Alternatively, we can define the converter as the superposition $\mathbf{D}^d_{\lambda,\tau} = \mathbf{C}_\lambda \mathbf{S}^d_\tau$ where the operator $\mathbf{S}^d_\tau : L^2(\mathbb{R}) \to \ell^2(\mathbb{Z})$ performs sampling and single-single bit quantization of the dithered input signal, $(\mathbf{S}^d_\tau f)[n] = \text{sgn}[(f + d)(n\tau)]$, and the operator $\mathbf{C}_\lambda$ performs selection and coding, which amounts to providing indices of sign changes. We refer to $\mathbf{D}^d_{\lambda,\tau}$ as a *single-bit* converter since the quantization involved is single-bit quantization.
2. The conditions on d can be relaxed: in addition to the oscillation requirement (4), it is sufficient to require that d is piecewise C^1, *i.e.* that it is C^1 in the open intervals I_n, for all n in $\mathbb{Z}$, and also that $\sup_n \sup_{t\in I_n} |d'(t)| = \Delta < \infty$.

The bit-rate, R, of this conversion scheme is determined by the number of sampling intervals within each interval I_n, the size of which is $1/\lambda$. Here λ will be kept fixed, and τ will typically be significantly smaller than $1/\lambda$; the large rate limit corresponds to $\tau \to 0$. Thus the bit-rate needed for specifying the location of one data change within I_n with precision τ equals $R = \lambda|\log(\tau\lambda)|$.

It remains to discuss how much information about the signal f is contained in the sequence $\mathbf{D}^d_{\lambda,\tau}(f)$, and how accurately f can be reconstructed from this information. For every interval I_n, the $\mathbf{D}^d_{\lambda,\tau}(f)$ gives us the value of a $s_{n,k} = k\tau + n/\lambda - 1/2\lambda \in I_n$ (we assume, for simplicity, that τ divides $1/\lambda$) such that $(f + d)(s_{n,k})$ and $(f + d)(s_{n,k+1})$ have different signs, implying that $(f + d)(t)$ must be zero for some $t \in (s_{n,k}, s_{n,k+1})$. Let us define, $t_n := s_{n,k} + \tau/2$. Since $|f'(x)| \leq \pi$ for all x (this is a consequence of $\|f\|_\infty < 1$ and the bandlimitedness of f), and $|d'(x)| \leq \|d\|_{C^1} =: \Delta$, it follows that $f(t_n) = -d(t_n) + \epsilon_n$, with $|\epsilon_n| \leq (\Delta + \pi)\tau/2$. The bit sequence $\mathbf{D}^d_{\lambda,\tau}(f)$ thus defines an effective sample sequence $(t_n)_{n\in\mathbb{Z}}$ and tells us, within an error proportional to τ, the values of f at these sample points.

The sequence $(t_n)_{n\in\mathbb{Z}}$ is uniformly discrete, i.e. $\inf_{n,k\in\mathbb{Z},n\neq k} |t_n - t_k| > 0$, which follows from $|(f + d)(n/\lambda + 1/(2\lambda))| \geq \gamma - 1$ and $|(f + d)'| \leq \Delta + \pi$. Moreover, the lower uniform density of $(t_n)_{n\in\mathbb{Z}}$ equals $\lambda > 1$. Therefore, $(t_n)_{n\in\mathbb{Z}}$ constitutes a sequence of stable sampling for all the spaces $\mathcal{V}_{\mu\pi}$ for all $\mu < \lambda$ [5], and thus for $\mathcal{V}_\pi$ and $\mathcal{C}$. Hence, any function f in $\mathcal{C}$ can be reconstructed from its samples $(f(t_n))_{n\in\mathbb{Z}}$. As we saw above, however, we know these samples only within a certain error. To make sure that these errors still allow for a stable reconstruction, we need an extra stability analysis, which will be discussed in the next section.

3 Local reconstruction and stability

Not every sequence of stable sampling allows for reconstruction algorithms that are still stable when the sample values are contaminated by errors. For instance, Shannon's classical sampling formula[1]

$$f(t) = \sum_n f(n)\mathrm{sinc}(t-n) \tag{5}$$

does not provide for a stable approximate reconstruction from perturbed samples. Indeed, if each $f(n)$ in (5) is replaced by $f_n^* = f(n) + \varepsilon_n$, with uniformly bounded perturbation, $|\varepsilon_n| \le \varepsilon$ for all n, then this can lead to divergences in some t, because the series $\sum_n \mathrm{sinc}(t-n)$ is not absolutely convergent. It is well known that this instability can be overcome by oversampling. Indeed, if we know, for $f \in \mathcal{V}_\pi$, the sample values $\left(f\left(\frac{n}{\lambda}\right)\right)_{n\in\mathbb{Z}}$, where $\lambda > 1$, then we can write many reconstruction formulas other than (5). For instance, if g is a function such that its Fourier transform $\hat{g}$ is C^∞, and satisfies $|\hat{g}(\omega)| = 0$ for $|\omega| > \lambda\pi, \hat{g}(\omega) = \frac{1}{\sqrt{2\pi}}$ for $|\omega| \le \pi$, and $0 < \hat{g}(\omega) < \frac{1}{\sqrt{2\pi}}$ for $\pi < |\omega| \le \lambda\pi$, then we also have, for all $f \in \mathcal{V}_\pi$,

$$f(t) = \frac{1}{\lambda}\sum_n f\left(\frac{n}{\lambda}\right) g\left(t - \frac{n}{\lambda}\right) \quad . \tag{6}$$

Because g decays faster than any inverse polynomial, the series in (6) is absolutely convergent, so that convergence of (6) holds not only in L^2, but also pointwise. Moreover, if the samples $f(\frac{n}{\lambda})$ in (6) are replaced by perturbed values $f_n^* = f(\frac{n}{\lambda}) + \varepsilon_n$, with $|\varepsilon_n| \le \varepsilon$ for all n, then the resulting sum approximates f within an error proportional to ε, uniformly in t:

$$\left| f(t) - \frac{1}{\lambda}\sum_n f_n^* g\left(t - \frac{n}{\lambda}\right)\right| \le \frac{\varepsilon}{\lambda}\sum_n \left| g\left(t - \frac{n}{\lambda}\right)\right| \le C\varepsilon \quad .$$

In this section we will show that similar stability properties can be proved for *irregular* sampling at densities higher than the Nyquist density. To prove this, we first establish the following lemma which is corollary of a theorem by S. Jaffard [6].

Lemma 3.1 *Suppose $B : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ is a bounded operator, and suppose there exist $K_1 > 0, K_2 < \infty$ such that, for all $c \in l^2(\mathbb{Z})$,*

$$K_1\|c\|^2 \le \|Bc\|^2 \le K_2\|c\|^2 \quad . \tag{7}$$

[1] We use the notation $\mathrm{sinc}(t) = \sin \pi t/\pi t$.

Suppose that, for some $\mu > 0$, the matrix elements $B_{m,n}$ satisfy

$$|B_{m,n}| \leq C_N(1 + |m - \mu n|)^{-N} \quad ,$$

for all $m, n \in \mathbb{Z}$ and all $N \geq 1$, with C_N independent of m, n. Then the operator B is invertible, and the matrix elements of its bounded inverse B^{-1} satisfy a dual inequality, i.e. there exist $C'_N > 0$ such that, for all $m, n \in \mathbb{Z}$ and all $N \geq 1$,

$$|(B^{-1})_{n,m}| \leq C'_N(1 + |\mu n - m|)^{-N} \quad .$$

Proof: The condition (7) on B implies that both B and B^*B are invertible; we have $B^{-1} = (B^*B)^{-1}B^*$. Now, for all $m, n \in \mathbb{Z}$ and $N > 1$,

$$|(B^*B)_{m,n}| \leq C_N^2 \sum_k (1 + |k - \mu m|)^{-N}(1 + |k - \mu n|)^{-N} \leq C_N^2 C(1 + |m - n|)^{-N} \quad .$$

The inverse of B^*B has the same decay off the main diagonal as B^*B [6], that is we have that $|[(B^*B)^{-1}]_{m,n}| \leq C''_N(1 + |m - n|)^{-N}$. It then follows that

$$|(B^{-1})_{n,m}| \leq C''_N C_N \sum_k (1 + |n - k|)^{-N}(1 + |m - \mu k|)^{-N} \leq C'_N(1 + |\mu n - m|)^{-N} \; .$$

We now use this result to prove the following theorem:

Theorem 3.2 *Suppose $(t_n)_{n \in \mathbb{Z}}$ is a uniformly discrete sequence such that $\sup_n |t_n - \frac{n}{\lambda}| < \infty$, where $\lambda > 1$. Then there exist functions $\psi_n \in C^\infty$ and constants $C_N > 0$ satisfying, for all $t \in \mathbb{R}$, all $n \in \mathbb{Z}$ and all $N \geq 1$,*

$$|\psi_n(t)| \leq C_N(1 + |t|)^{-N} \quad ,$$

such that, for all $f \in \mathcal{V}_\pi$,

$$f(t) = \sum_n f(t_n)\psi_n(t - t_n) \quad ,$$

where the convergence holds pointwise, absolutely and uniformly on compact sets, as well as in L^2. The functions ψ_n depend on the particular sequence $(t_n)_{n \in \mathbb{Z}}$, but the constants C_N can be chosen so that they depend only on λ, $\sup_n |t_n - \frac{n}{\lambda}|$ and $\inf_{n \neq k} |t_n - t_k|$.

Proof: 1. Our argument will use a specially constructed space V_g so that $\mathcal{V}_\pi \subset V_g \subset \mathcal{V}_{\lambda\pi}$. To construct this space, we start by choosing a number $\nu \in (1, \lambda)$, and a monotonous C^∞ function $\theta : \mathbb{R} \to [0, 1]$ such that $\theta(x) = 0$ for $x < -1/2$, $\theta(x) = 1$ for $x > 1/2$, and $\theta(x) + \theta(1 - x) = 1$ for all x. We then define the function g by

$$\hat{g}(\omega) = \frac{1}{\sqrt{\pi(1 + \nu)}} \sin\left[\frac{\pi}{2}\theta\left(\frac{1 + \nu - 2|\omega|/\pi}{2(\nu - 1)}\right)\right] \quad .$$

One checks that $\hat{g}(\omega) = 1/\sqrt{\pi(1+\nu)}$ for $|\omega| < \pi$, $\hat{g}(\omega) = 0$ for $|\omega| > \nu\pi$, and

$$\sum_m \left| \hat{g}\left(\omega + \frac{(\nu+1)m}{2}\right) \right|^2 = \frac{1}{\pi(\nu+1)}. \tag{8}$$

From this construction and (8) it follows that functions $g_k(t) := g(t - \frac{2k}{\nu+1})$ $k \in \mathbb{Z}$ are orthonormal. We shall denote by V_g the subspace of $L^2(\mathbb{R})$ spanned by the $(g_k)_{k\in\mathbb{Z}}$.

2. A generic function φ in V_g can be written as $\varphi = \sum_k c_k g_k$, where $c_k =< \varphi, g_k >$, so that $\sum_k |c_k|^2 < \infty$. The Fourier transform $\hat{\varphi}$ of φ can then be written as

$$\hat{\varphi}(\omega) = \sum_k c_k e^{-i2k\omega/(\nu+1)}\ \hat{g}(\omega) \quad ,$$

i.e. $\hat{\varphi}$ is the product of $\hat{g}$ with any $\pi(\nu+1)$-periodic function for which the restriction to one period is square integrable. We can use this observation to show that $\mathcal{V}_\pi \subset V_g$. If $h \in \mathcal{V}_\pi$, then $H(\omega) = \sqrt{\pi(1+\nu)} \sum_m \hat{h}(\omega + \pi m(\nu+1))$ is $\pi(\nu+1)$-periodic, and its restriction to $[-\frac{\pi}{2}(\nu+1), \frac{\pi}{2}(\nu+1)]$ is a multiple of $\hat{h}$, and therefore square integrable. Moreover, because $H(\omega) = 0$ for $\pi < |\omega| < \nu\pi$ and $\hat{g}(\omega) = 0$ for $|\omega| \geq \nu\pi$, we have $\hat{g}(\omega)H(\omega) = \hat{h}(\omega)$; it follows that $h \in V_g$, establishing $\mathcal{V}_\pi \subset V_g$. The inclusion $V_g \subset \mathcal{V}_{\nu\pi}$ follows immediately from $support(\hat{g}) \subset [-\nu\pi, \nu\pi]$.

3. Due to the fast decay of g, the formula $\varphi = \sum_k < \varphi, g_k > g_k$ converges not only in L^2, but also uniformly and absolutely pointwise. If we choose, in particular, to apply this expansion to $f \in \mathcal{V}_\pi \subset V_g$, then $< f, g_k >= \sqrt{\frac{2}{\nu+1}} f\left(\frac{2k}{\nu+1}\right)$ and the expansion reverts to a special case of (6).

4. We now proceed to derive reconstruction formulas for functions in V_g from their samples at the t_n. Since $V_g \subset \mathcal{V}_{\nu\pi}$ and $(t_n)_{n\in\mathbb{Z}}$ is a sequence of stable sampling for $\mathcal{V}_{\nu\pi}$, there exist $K_1 > 0, K_2 < \infty$ so that, for all $\varphi \in V_g$, $K_1\|\varphi\|^2 \leq \sum_n |\varphi(t_n)|^2 \leq K_2\|\varphi\|^2$. If we write this in terms of the $c_k =< \varphi, g_k >$, and introducing $B_{n,k} = g_k(t_n)$, we obtain

$$K_1 \sum_k |c_k|^2 \leq \sum_n |\sum_k B_{n,k} c_k|^2 \leq K_2 \sum_k |c_k|^2 \quad . \tag{9}$$

Because g decays faster than any inverse polynomial, and $|t_n - \frac{n}{\lambda}| \leq C < \infty$, there exist $C_N > 0$ so that for all $k, n \in \mathbb{Z}$ and all $N > 1$,

$$|B_{n,k}| = \left| g\left(t_n - \frac{2k}{\nu+1}\right) \right| \leq C_N \left[1 + \left| \frac{(\nu+1)n}{2\lambda} - k \right| \right]^{-N} \quad .$$

It then follows from (9) and Lemma 3.1 that B is invertible and that the matrix elements $(B^{-1})_{k,n}$ satisfy a similar inequality, i.e. for all $k, n \in \mathbb{Z}$ and all $N > 1$,

$$|(B^{-1})_{k,n}| \leq C'_N \left[1 + \left| \frac{(\nu+1)n}{2\lambda} - k \right| \right]^{-N} \quad ,$$

where C'_N is independent of k, n. Since $< \varphi, g_k >= \sum_n (B^{-1})_{k,n} \varphi(t_n)$, we obtain that

$$\varphi(t) = \sum_k < \varphi, g_k > \; g\left(t - \frac{2k}{\nu+1}\right) = \sum_n \varphi(t_n) \left(\sum_k (B^{-1})_{k,n} \; g\left(t - \frac{2k}{\nu+1}\right)\right) \quad ,$$

where the sums converge absolutely.

5. Define now $\psi_n(t) = \sum_k (B^{-1})_{k,n} g(t + t_n - \frac{2k}{\nu+1})$. For all $n \in \mathbb{Z}, t \in \mathbb{R}$, and $N > 1$

$$|\psi_n(t)| \leq C'_N C''_N \sum_k \left[1 + \left|\frac{(\nu+1)n}{2\lambda} - k\right|\right]^{-N} \left[1 + \left|t + \frac{n}{\lambda} - \frac{2k}{\nu+1}\right|\right]^{-N} \leq C'''_N [1+|t|]^{-N} \quad .$$

On the other hand,

$$\varphi(t) = \sum_n \varphi(t_n) \psi_n(t - t_n) \quad .$$

This holds for all $\varphi \in V_g$; in particular, it holds for $f \in \mathcal{V}_\pi$, which proves our claim.

The following stability result is an immediate corollary.

Corollary 3.3 *Suppose that the sequence $(t_n)_{n \in \mathbb{Z}}$ is uniformly discrete, and that $\sup_n |t_n - \frac{n}{\lambda}| \leq C < \infty$, where $\lambda > 1$. Let $(\alpha_n)_{n \in \mathbb{Z}}$ be a sequence such that there exists a function $f \in \mathcal{V}_\pi$ for which $|f(t_n) - \alpha_n| \leq \varepsilon$ for all $n \in \mathbb{Z}$. Then, for all $t \in \mathbb{R}$, $|f(t) - \sum_n \alpha_n \psi_n(t - t_n)| \leq \varepsilon \sum_n |\psi_n(t - t_n)| \leq C'\varepsilon$, where the ψ_n are the functions of Theorem 3.2, and C' is independent of the particular sequence $(t_n)_{n \in \mathbb{Z}}$.*

The following theorem statement summarizes what we have proved so far.

Theorem 3.4 *Let d be a bounded C^1-function satisfying (4) and $|d'(t)| \leq \Delta$ for all t, and let $\lambda > 1$. Let $\mathbf{D}^d_{\lambda\tau}$ be the corresponding single-bit dithered oversampled A/D converter, as defined in Definition 3.1. Take $f \in \mathcal{C}$, and for every $n \in \mathbb{Z}$, define*

$$t_n = \left(\mu_n + (\mathbf{D}^d_{\lambda,\tau} f)[n] + \frac{1}{2}\right) \tau \quad ,$$

where $\mu_n = \lfloor n/\lambda\tau \rfloor$ Then one can reconstruct an approximation $\tilde{f}$ to f as

$$\tilde{f}(t) = - \sum_{n \in \mathbb{Z}} d(t_n) \psi_n(t - t_n) \quad ,$$

where the ψ_n are the functions associated with the sequence $(t_n)_{n \in \mathbb{Z}}$ as in Theorem 3.2; the pointwise error is then bounded by

$$|f(t) - \tilde{f}(t)| \leq C(\Delta + \pi)\frac{\tau}{2} \quad , \qquad (10)$$

where C does not depend on f or τ.

The distortion of this A/D conversion scheme, defined as the local average of error power, can be bounded by

$$D = \frac{1}{T}\int_{|s-t|\leq T/2} |f(s)-\tilde{f}(s)|^2 ds \leq \frac{C^2}{4}(\Delta+\pi)^2\tau^2 \quad ,$$

uniformly on $\mathcal{C}$. Considering that the bit-rate of this scheme equals $R = \lambda|\log(\tau\lambda)|$ this leads to a rate-distortion characteristic of the form $D \leq C2^{-\gamma R}$.

In the next section we discuss some practical local reconstruction algorithms.

4 Reconstruction with finite interpolation

As an illustration of other reconstruction algorithms, we give one example of a local interpolation algorithm which gives accuracy similar to (10), up to $|\log(\tau)|$ factors.

Given an appropriate dither function with maximum amplitude $\gamma > 1$ and C^1 norm $\Delta := \sup|d'(t)|$, the t_n must be separated by $\delta := \frac{2(\gamma-1)}{\Delta+\pi} - \tau$. For the sake of definiteness, let us assume $\lambda = 2$. For each $m \in \mathbb{Z}$, one can compute an approximation to $f(\frac{m}{2})$ by Lagrange interpolation of the $f(t_{m+l})$ with $|l| \leq L$; we denote this value by $f_{\text{app};m,L}$. The Lagrange interpolation $g_{\text{approx},K}(x)$ based on the values values $g(x_1), \cdots, g(x_K)$ of a function g that is K times continuously differentiable, satisfies the bound

$$|g(x) - g_{\text{approx},K}(x)| \;\leq\; \frac{1}{K!}\sup_y |g^{(K)}(y)| \prod_{k=1}^{K} |t-t_k| \quad .$$

In our case $|f^{(l)}(t)| \leq \pi^l$, because $f \in \mathcal{V}_\pi$ and $|f(t)| \leq 1$. It follows that

$$\left| f\left(\frac{m}{2}\right) - f_{\text{app};m,L}\right| \;\leq\; \frac{\pi^{2L+1}}{4(2L+1)!}\prod_{l=1}^{L}\left(\frac{l}{2}+\frac{1}{4}\right)^2 \;\leq\; \sqrt{L}\left(\frac{\pi}{4}\right)^{2L+1} \quad . \tag{11}$$

This decreases exponentially as L increases. However, as explained earlier, we do our reconstruction not from the exact $f(t_n)$, but from approximate values that are within $(\Delta+\pi)\tau/2$ of the true $f(t_n)$. We therefore also need to estimate the error between $f_{\text{app};m,L}$ and $\tilde{f}_{\text{app};m,L}$, which is the Lagrange interpolation of the approximate values of $f(t_n)$. The explicit form of the Lagrange interpolation formula allows us to bound this as

$$|f_{\text{app};m,L} - \tilde{f}_{\text{app};m,L}| \;\leq\; \frac{\Delta+\pi}{2}\tau \sum_{l=-L}^{L} \prod_{k\neq l; k\in\{-L,L\}} \frac{|m/2 - t_{m+k}|}{|t_{m+l}-t_{m+k}|} \leq C_1\,\frac{(\Delta+\pi)\tau}{2\delta}L^2 \quad . \tag{12}$$

In obtaining the final bound in (12) we used $|m/2 - t_{m+k}| \leq (2|k|+1)/4$, $|t_{m+l} - t_{m+k}| \geq \delta + (|k-l|-1)/2$, and $|t_{m+l} - t_{m+l+l'}||t_{m+l} - t_{m+l-l'}| \geq (\delta + |l'| - 1)|l'|$. The factor C_1 is about 90, and its value would change if we chose another λ. Combining the two estimates, we find

$$\left| f\left(\frac{m}{2}\right) - \tilde{f}_{\text{app};m,L}\right| \;\leq\; \sqrt{L}\left(\frac{\pi}{4}\right)^{2L+1} \;+\; C_1\frac{(\Delta+\pi)\tau L^2}{2\delta} \quad ;$$

optimizing over L leads to

$$\left| f\left(\frac{m}{2}\right) - \tilde{f}_{\mathrm{app};m,L} \right| \;\leq\; C_2 \tau (\log \tau)^{(2+\epsilon)} \quad .$$

We have thus a reconstruction of the regular samples $f(m/2)$ with precision proportional to τ, up to logarithmic factors in τ. From these regularly spaced samples, the whole function f can be reconstructed using standard techniques. Several steps in this bounding estimate are very coarse, and it can no doubt be improved by a more careful analysis. This example is given only to show that one can achieve near optimality very easily, with local and fast algorithms.

More practical local algorithms may well involve trigonometric polynomial interpolation rather than Lagrange interpolation. It should also be noted that departing from "standard" frames, and introducing weight factors for the different samples, depending on their local density [7], would probably lead to better error bounds. In this more adaptive framework, all the zero crossings of $f+d$ could be used, instead of only one per interval I_n, as in Definition 2.1. But all this is work for future research.

Acknowledgments

The authors are grateful to R. DeVore, K. Gröchening and S. Güntürk for fruitful discussions on this topic.

References

[1] W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal.* Vol. 27, pp. 446-472, July 1948.

[2] N. T. Thao and M. Vetterli. Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates. *IEEE Trans. on Signal Processing.* Vol. 42, No. 3, pp. 519-531, March 1994.

[3] Z. Cvetković and M. Vetterli. On simple oversampled A/D conversion in $L^2(\mathbf{R})$. Submitted to *IEEE Trans. on Information Theory.* June 1999.

[4] Z. Cvetković and M. Vetterli. Error-rate characteristics of oversampled analog-to-digital conversion. *IEEE Trans. Information Theory.* Vol. 44, No. 5, pp. 1961-1964, September 1998.

[5] S. Jaffard. A density criterion for frames of complex exponentials. *Michigan Math. Journal.* Vol. 38, pp. 339-348, 1991.

[6] S. Jaffard. Propriétés des matrices 'bien localisées' près de leur diagonale et quelques applications. *Ann. Ints. Henri Poincaré.* Vol. 7, pp. 461-476, 1990.

[7] K. Gröchenig. Irregular sampling, toeplitz matrices, and the approximation of entire functions of exponential type. *Mathematics of Computation.* Vol. 68, No. 226, pp. 749-765, April 1999.