Quantization Noise Minimization in $\Sigma\Delta$-modulation based RF Transmitter Architectures

Ulf Gustavsson, *Student Member, IEEE,* Thomas Eriksson and Christian Fager, *Member, IEEE*

***Abstract*—This paper describes an optimization method for minimization of quantization noise in $\Sigma\Delta$-based RF transmitters. The aim of the method is to enable the use of reconstruction filters with wider passband, or alternatively, a lower switch-rate.**

The method uses a general representation of the $\Sigma\Delta$-converters in combination with a differentiable approximation of the quantizer. Based on this, a Monte-Carlo based algorithm is developed around the damped Gauss-Newton iteration. As a result of the suggested algorithm, the residual quantization noise after reconstruction filtering is significantly decreased.

Finally, simulations using a bandlimited signal with a Gaussian distribution are used to demonstrate the capabilities of the suggested algorithm when applied with the proposed $\Sigma\Delta$-modulator representation. The resulting performance is compared to several cases of $\Sigma\Delta$-converters designed using traditional methods, demonstrating significant improvements in terms of reduced reconstruction normalized mean square error (NMSE). This implicates that the transmitter efficiency can be improved with minor changes in the modulator implementation.

***Index Terms*—$\Sigma\Delta$-modulation, Pulse-Density Modulation (PDM), Noise-Shaped Coding (NSC), Quantization Noise, Gauss-Newton iteration, Monte-Carlo based algorithms, RF transmitter architectures.**

I. INTRODUCTION

MODERN wireless systems use advanced, high order modulation schemes to maximize the capacity. As a consequence, the signal has very large peak to average power ratio. Traditional amplifiers need therefore to be operated in a backed off, low efficiency, region to satisfy the linearity requirements. However, a large number of efficiency enhancement techniques have been proposed in order to circumvent this problem and techniques like the Doherty amplifier [1], Chireix outphasing systems [2] and envelope tracking (ET) [3] have therefore gained popularity during the last decade.

Manuscript received August 17, 2009; revised December 22, 2009, March 8, 2010 and May 6 2010. This research has been carried out in GigaHertz Centre in a joint project financed by the Swedish Governmental Agency for Innovation Systems (VINNOVA), Chalmers University of Technology, Ericsson AB, Infineon Technologies Austria AG, and NXP Semiconductors BV. This paper was recommended by Associate Editor Andreas Demosthenous.

U. Gustavsson is with Ericsson AB, Stockholm, Sweden (e-mail: ulf.gustavsson@ericsson.com).

C. Fager is with the department of Microtechnology and Nanoscience, Microwave Electronics Laboratory, Chalmers University of Technology (e-mail: christian.fager@chalmers.se).

T. Eriksson is with the department of Signals and Systems, Communication Systems Group, Chalmers University of Technology (e-mail: thomase@chalmers.se).

Digital Object Identifier 10.1109/TCSI.2010.2052512

Other promising techniques proposed to achieve high efficiency, while maintaining high linearity, are based upon the use of 1-bit quantization in different forms with the amplifiers [4]. This means that the PA is operated only in either of its two most efficient points: in deep compression or completely off. These types of quantization schemes usually map the amplitude and phase information of the signal to either the width and position of each pulse (pulse width modulation, PWM), or to the density of pulses with a fixed duration (pulse density modulation, PDM).

Unfortunately, the 1-bit quantized representation of the signal contains large amounts of undesired distortion. Due to regulations imposed upon all wireless communication systems one can not simply transmit the quantized signal. Therefore, reconstruction filters, in general of bandpass type, are needed to avoid violations of the spectral mask. The required fractional bandwidth of these filters are, however, very small which causes the insertion loss in the pass-band to be large for the practical implementation [5]. This reduces the power delivered to the load, thus decreasing the power efficiency of the system considerably. The amount of quantization noise produced within the filter passband can be reduced by increasing the pulse rate. However, this leads to increased switch losses in the power amplifier circuit, and very high clock rates in the digital signal processing units.

One possible remedy for this problem is to apply pulserates at a moderate level in combination with *Noise Shaped Coding* (NSC) [6]. NSC maps the signal by PDM in a manner where the quantization noise is minimized in a specific part of the spectrum. The most common type of implementation of NSC mappings are so called $\Sigma\Delta$-modulators [7], [8]. In this type of modulator, the NSC properties are directly determined by the coefficients of the loop-filters inherent to the structure. As shown in Fig. 1, the key idea is to minimize the energy of the quantization noise within a specified frequency band, BW_F, considerably larger than the bandwidth of the signal which is to be quantized, BW_s. Increasing BW_F while keeping the quantization noise within BW_F to a minimum, can potentially relax the bandwidth requirements for the reconstruction filters used and therefore also reduce their losses.

Determining the loop-filter coefficients for desired NSC is however far from trivial, since the loop comprises an extreme nonlinearity in the form of a quantizer. Regular linear systems theory lacks straight-forward methods to compute these coefficients for arbitrary input signal statistics. Thus, an optimization-based method needs to be deployed in which the coefficients derived from the linear models can serve as good

1549–8328/00$26.00 © 2010 IEEE.

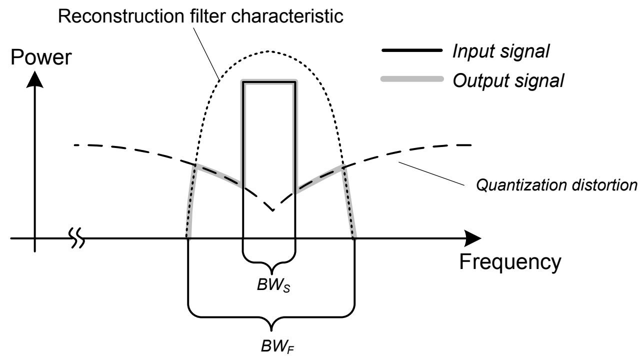


Fig. 1: Illustration of the optimization criteria in the frequency domain.

start-points.

In this paper we first survey some state of the art analysis approaches for $\Sigma\Delta$-modulators and their use in high efficiency RF transmitters. We then suggest an algorithm in which a differentiable approximation of the quantizer is introduced. In combination with a generalized IIR-filter based representation of the $\Sigma\Delta$-modulator, we enable the use of gradient-based search algorithms. As a result, it is possible to reduce the quantization noise within BW_F, significantly compared to the common modulator implementations.

The paper is organized as follows. First, we will review the state-of-the-art in $\Sigma\Delta$-modulator analysis and $\Sigma\Delta$-modulator based transmitter architectures in section II. This serves as a background to section III, which introduces a highly generalized mathematical representation of the $\Sigma\Delta$-modulator. In section IV we move toward describing a working optimization algorithm. The equations for implementing the damped Gauss-Newton iteration are then described in section IV-A. Further on, a differentiable approximation of the quantizer is suggested in section IV-B, which enables the use of the damped Gauss-Newton iteration. An optimization algorithm is then suggested in section IV-C, using the damped Gauss-Newton in combination with a Monte-Carlo based technique for reducing overall start-value sensitivity of the algorithm. The outcome of the algorithm in combination with the generalized $\Sigma\Delta$-representation is then benchmarked against regular integrator-based solutions as well as against modulators comprising Noise Transfer Functions (NTF) optimized with traditional methods. This benchmark is performed by experimental simulations in section V. Finally, conclusions are drawn in section VI.

II. State of the art

A. $\Sigma\Delta$-modulator analysis

One of the most common topologies used in $\Sigma\Delta$-modulation is the integrator based low-pass modulator. This topology is commonly used in applications where a very large oversampling ratio (OSR) is feasible, e.g. for audio coding. For wideband RF applications, however, the headroom for using high OSR is not as generous. An oversampling ratio below 20 is typically used when it comes to baseband $\Sigma\Delta$-modulation [9]. The need for optimized NSC is therefore of paramount importance in pulsed RF transmitters.

Several methods of mathematical analysis of the quantization noise have been suggested in [10], [11]. However, these approaches are mainly aimed for linear and nonlinear reconstruction approaches in digital systems and are highly complex. Further on, they leave little insight of how to determine the modulator parameters in order to achieve desired NSC, e.g. desired spectral shape of the quantization noise from a given criterion. Another recently developed approach for optimizing the NSC properties of a $\Sigma\Delta$-modulator is presented in [12]. Here, the 1-bit quantization problem is analyzed as a maximum likelihood sequence detection, for which the Viterbi-algorithm is the optimal solution.

The vast majority of published analysis methods are however based on empirical methods [13], [14], [15], [16]. These methods are typically derived under the assumption that the quantization noise is a stochastic process with a uniform or Gaussian distribution, usually modeled as independent of the quantizer input. From this assumption, the quantizer is replaced with a noise source in order to enable regular linear analysis (thus the term quantization "noise"). These types of models are useful for determining the filter-coefficients for a subclass of simplified problems, as for example the low-pass integrator-based $\Sigma\Delta$-modulators.

B. $\Sigma\Delta$-modulator based transmitter architectures

The use of 1-bit quantization for efficiency enhancement of power amplifiers has been demonstrated using different types of transmitter system topologies. Some applications suggests that the quantization should be performed at RF-rate or above, as for example RF pulse-width modulation (RF-PWM, [4]) or band-pass $\Delta\Sigma$-modulation (BP$\Sigma\Delta$ [17], [18]). These types of architectures are in general very difficult to implement since the harmonic content of the pulse train ranges over several multiples of the carrier frequency and therefore put extreme requirements on the modulator implementation, as well as the bandwidth of the interconnect between the modulator and the power amplifier circuits. Systems performing the quantization on a baseband level have also been suggested, either in Cartesian or polar form, [19], [20]. The quantized signal can then be applied by either modulating the RF input of the PA [21], [22], but it can also be applied by switching the DC voltage supply [9].

A common baseband type of pulsed transmitter architecture is called Cartesian pulse modulation, which is illustrated in Fig. 2. Here, the signal is treated in Cartesian form, e.g. on the orthogonal quadrature components I and Q. These are then separately quantized by two $\Sigma\Delta$-modulators before they are recombined and up-converted to the RF-carrier frequency [20]. The figure also shows the reconstruction filter in form of a bandpass-filter connected at the PA output. In all of these cases, the residual quantization noise needs to be suppressed in order to comply with spectral requirements of the output signal.

Further on, there have been several methods suggested to cope with the filtering issue by selective cancellation of

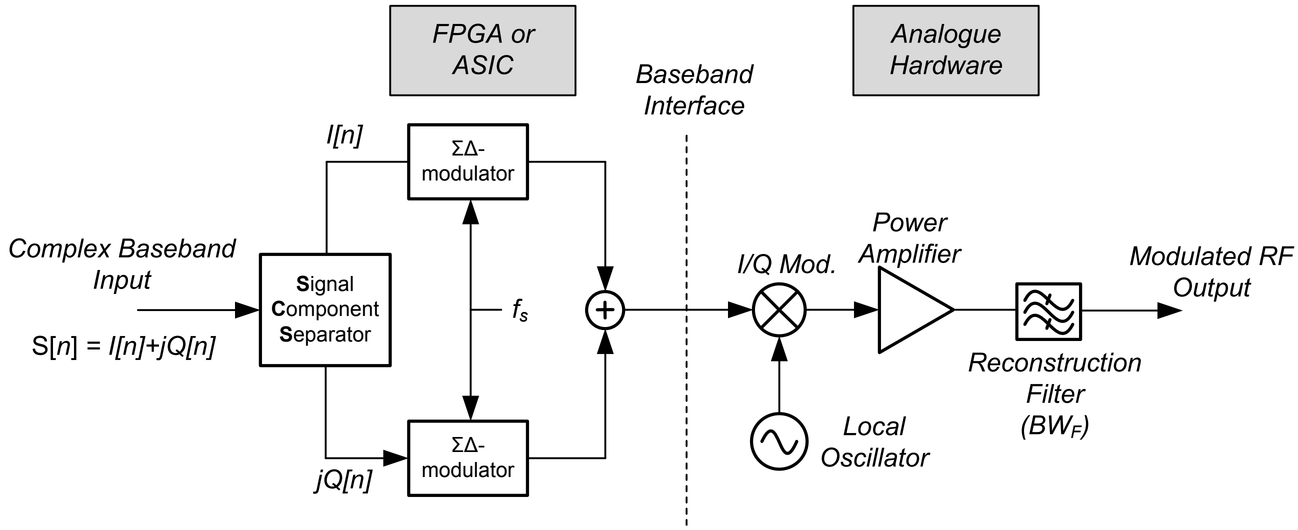


Fig. 2: Simplified schematic view of a Cartesian pulse modulated system, using $\Sigma\Delta$-modulators for quantization of the quadrature components of the signal. The orthogonal I and Q components are then combined and upconverted to RF, creating a constant envelope drive signal for high efficiency operation of the power amplifier. The reconstruction filter is then used to recreate the bandlimited communication signal.

quantization noise close to the carrier. In [23] a feed-forward method quite commonly used to perform linearization in regular RF PA systems is suggested, to cancel the quantization noise close to the modulated RF-carrier. An alternative to this is suggested in [24], which requires no additional hardware, but suppresses the adjacent quantization noise by imposing an amplitude component onto the pulses.

A quite different use for the $\Sigma\Delta$-modulator related to high efficiency RF transmitters is presented in [25], where a asynchronous, continuous-time $\Sigma\Delta$-modulator is used for stabilizing a very wideband feedback-loop used for pre-distorting a RF PA.

We now move on to describe the suggested, generalized $\Sigma\Delta$-modulator structure, and the algorithm designed to optimize its NSC performance.

III. Generalized $\Sigma\Delta$-modulator representation

In order to enable a gradient-based method to search for the coefficients that provides the desired NSC properties, we need a highly generalized structure able to represent as many $\Sigma\Delta$-modulator implementations as possible. The quantizer function $Q(\cdot)$ can be arbitrarily defined, but within the scope of this paper we will consider a 1-bit quantizer only. The generalized $\Sigma\Delta$-modulator representation suggested in this paper is shown in Fig. 3, where we H and G are considered to be generic IIR-filters with coefficients

$$H \triangleq \{a_1, \dots, a_{Q_H}, b_0, \dots, b_{P_H}\} \quad (1)$$

$$G \triangleq \{c_1, \dots, c_{Q_G}, d_0, \dots, d_{P_G}\} \quad (2)$$

P_G and P_H are the feedforward orders and Q_G, Q_H are the feedback orders of G and H, respectively. Note that we can obtain the case of FIR-filters by simply setting $Q_G = Q_H = 0$. The equations governing this system, at time instant n and at each node of the system, are described in (3) - (6).

$$p_n = x_n - r_n \quad (3)$$

$$r_n = \frac{1}{c_0}\left(\sum_{k=0}^{P_G} d_k q_{n-k-1} + \sum_{k=1}^{Q_G} c_k r_{n-k}\right) \quad (4)$$

$$z_n = \frac{1}{a_0}\left(\sum_{k=0}^{P_H} b_k p_{n-k} + \sum_{j=1}^{Q_H} a_j z_{n-j}\right) \quad (5)$$

$$q_n = Q(z_n) \quad (6)$$

The constant scaling factor α included with the general representation in Fig. 3 is applied to accommodate arbitrary input signal variance σ_x^2. This particular representation is capable of representing a very large set of different implementations which makes is suitable for use in the forthcoming analysis. For example, by setting H and G to

$$H(z) = \frac{1}{1-z^{-1}} \quad (7)$$

$$G(z) = 1 \quad (8)$$

we obtain the NSC-properties of a first order single loop, integrator-based implementation of a low-pass $\Sigma\Delta$-modulator shown in Fig. 4 (a). Further on, by setting

$$H(z) = \frac{1}{(1-z^{-1})^2} \quad (9)$$

$$G(z) = b + (a-b)z^{-1} \quad (10)$$

where a and b are the constant gain coefficients of the two feedback branches as shown in Fig. 4 (b), we can achieve the NSC-property of a second order dual loop, integrator-based low-pass $\Sigma\Delta$-modulator. Both of these implementations are

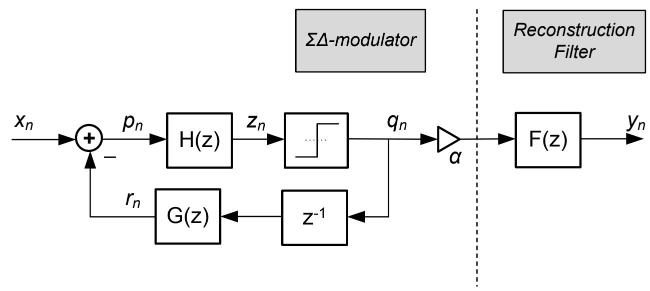


Fig. 3: A general representation of a $\Sigma\Delta$-modulator. The reconstruction filter, $F(z)$, determines the frequency-band over which the optimization is to be performed

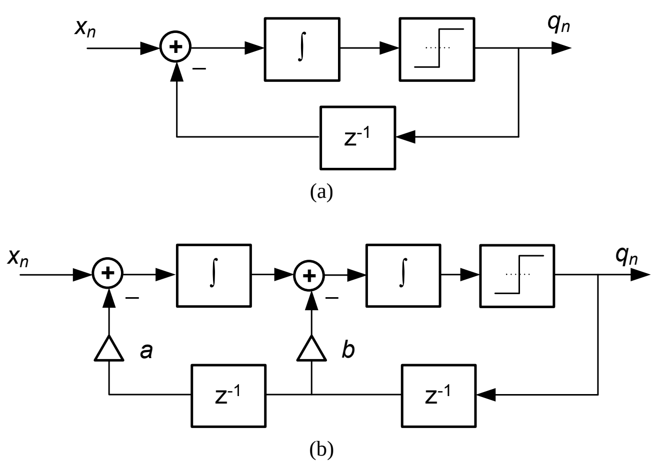


Fig. 4: Examples of common implementations of $\Sigma\Delta$-modulators. (a) 1^{st} order low-pass $\Sigma\Delta$-modulator. (b) 2^{nd} order low-pass $\Sigma\Delta$-modulator.

described in [26]. Analogously, the generalized representation in Fig. 3 can, via simple analysis, be used to describe the NSC-properties of higher order $\Sigma\Delta$-modulators, both low-pass or even band-pass.

IV. Optimization of the noise-shaped coding properties

We now proceed by showing how a differentiable approximation of the quantizer can enable a gradient based search over the parameter space of the generalized $\Sigma\Delta$-representation. This is illustrated by deriving the equations for the damped Gauss-Newton iteration, as described in [27]. Further on, it will also be shown that by using the search repeatedly while iteratively letting the approximation approach the quantizer by incrementing the slope-factor λ, we can move toward one set of parameters $\boldsymbol{\theta}^*$ that provides a good minimum for our reconstruction error criteria.

A. Damped Gauss-Newton iteration

In this subsection we derive the expressions necessary to form the damped Gauss-Newton iteration applied to the particular problem described in previous sections. By introducing the reconstruction filter, $F(z)$, we can use the reconstructed output sequence y_n to formulate the optimization criterion in the terms of mean square error:

$$\{\alpha^*, \boldsymbol{\theta}^*\} = \arg\min_{\{\alpha,\boldsymbol{\theta}\}} \frac{1}{N}\sum_{n=0}^{N-1} |x_n - y_n(\alpha,\boldsymbol{\theta})|^2 \tag{11}$$

where $\boldsymbol{\theta} \triangleq \{H,\ G\}$ and α is the scalar constant used to adapt the structure to arbitrary signal variance σ_x^2. In this case, the reconstruction-filter is represented by the impulse response $F \triangleq \{f_k\}_{k=1}^{L}$ of a linear phase filter with selected pass-band BW_F. The expression for the output sequence y_n can now be formed as

$$y_n(\alpha,\boldsymbol{\theta}) = \alpha\sum_{k=0}^{L-1} f_n q_{n-k}(\boldsymbol{\theta}) \tag{12}$$

Since convolution and differentiation are both linear operators and F is linear, it is straight forward to calculate the partial derivatives $\partial y_n/\partial\theta_i$ by filtering the partial derivative $\partial q_n/\partial\theta_i$ as

$$\frac{\partial y_n}{\partial\theta_i} = \alpha\sum_{k=0}^{L-1} f_k \frac{\partial q_n}{\partial\theta_i} \tag{13}$$

From here, it is now easy to apply this in any gradient-based search of choice. For this paper we have selected the damped Gauss-Newton method, where the parameter vector for the $k+1$:th iteration is calculated as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mu\mathbf{H}^{\dagger}\left[\mathbf{x} - \mathbf{y}(\alpha^{(k)},\boldsymbol{\theta}^{(k)})\right] \tag{14}$$

where μ is the step-size and

$$\mathbf{x} \;=\; [x_1\ x_2\ \ldots\ x_N]^T \tag{15}$$

$$\mathbf{y}(\alpha^{(k)},\boldsymbol{\theta}^{(k)}) \;=\; \left[y_1(\alpha^{(k)},\boldsymbol{\theta}^{(k)})\ \ldots\ y_N(\alpha^{(k)},\boldsymbol{\theta}^{(k)})\right]^T \tag{16}$$

are the $N \times 1$ vectors containing the input data sequence and reconstructed output data sequence, respectively. Finally, † denotes the Moore-Penrose generalized matrix inverse,

$$\mathbf{H}^{\dagger} = \left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T \tag{17}$$

Further on, $\mathbf{H}$ is the $N \times P$ Jacobian matrix for parameter vector $\boldsymbol{\theta}^{(k)}$, written as

$$\mathbf{H} = \begin{bmatrix} \frac{\partial y_1(\alpha^{(k)},\boldsymbol{\theta}^{(k)})}{\partial\theta_1^{(k)}} & \cdots & \frac{\partial y_1(\alpha^{(k)},\boldsymbol{\theta}^{(k)})}{\partial\theta_P^{(k)}} \\ \vdots & & \vdots \\ \frac{\partial y_N(\alpha^{(k)},\boldsymbol{\theta}^{(k)})}{\partial\theta_1^{(k)}} & \cdots & \frac{\partial y_N(\alpha^{(k)},\boldsymbol{\theta}^{(k)})}{\partial\theta_P^{(k)}} \end{bmatrix} \tag{18}$$

where $P = \dim\boldsymbol{\theta}$ and N is the data record length. The elements needed to calculate $\mathbf{H}$ are given in (19)-(22), which

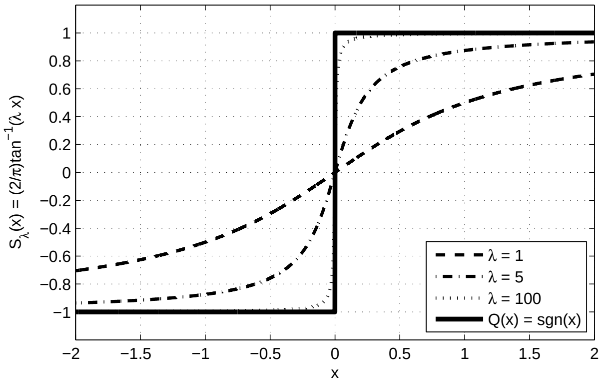


Fig. 5: Approximation of the ± 1 quantizer with different settings of the slope-factor λ.

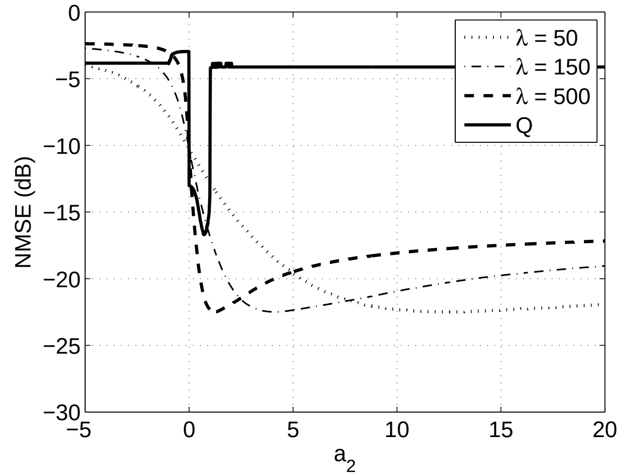


Fig. 6: One-dimensional error-surface over the H-filter coefficient a_2, with three different settings of the slope-factor λ and the quantizer $Q(\cdot)$.

forms a closed, recursive expression.

$$\frac{\partial p_n}{\partial \theta_i} = -\frac{\partial r_n}{\partial \theta_i} \tag{19}$$

$$\begin{aligned}\frac{\partial r_n}{\partial \theta_i} &= \frac{1}{c_0}\sum_{k=0}^{P_G}\left(\frac{\partial d_k}{\partial \theta_i}q_{n-k-1} + d_k\frac{\partial q_{n-k-1}}{\partial \theta_i}\right)\\ &+ \frac{1}{c_0}\sum_{j=1}^{Q_G}\left(\frac{\partial c_j}{\partial \theta_i}r_{n-j} + c_j\frac{\partial r_{n-j}}{\partial \theta_i}\right)\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{\partial z_n}{\partial \theta_i} &= \frac{1}{a_0}\sum_{k=0}^{P_H}\left(\frac{\partial b_k}{\partial \theta_i}p_{n-k} + b_k\frac{\partial p_{n-k}}{\partial \theta_i}\right)\\ &+ \frac{1}{a_0}\sum_{j=1}^{Q_H}\left(\frac{\partial a_j}{\partial \theta_i}z_{n-k} + a_j\frac{\partial z_{n-k}}{\partial \theta_i}\right)\end{aligned} \tag{21}$$

$$\frac{\partial q_n}{\partial \theta_i} = 0 \tag{22}$$

A consequence of (22) is that the Jacobian $\mathbf{H}$ in (18) becomes a zero-matrix when a true quantizer is used. Gradient based search algorithms can therefore not be directly used to find $\boldsymbol{\theta}^*$. In the next section we describe a method where the quantizer is replaced with a differentiable approximation to circumvent this problem.

B. Differentiable quantizer approximation

Within the framework of this paper we will only consider the case of the ± 1-quantizer, but the analysis below is easy to generalize for arbitrary choice of quantizer. The proposed method solves the issue pointed out in (22) by a simple differentiable approximation of $Q(\cdot)$. In this work, we have used the quantizer approximation described by (23), which is illustrated in Fig. 5 over several slope-factor values, denoted λ.

$$\mathcal{S}_\lambda(x) = \frac{2}{\pi}\tan^{-1}(\lambda x) \tag{23}$$

By using (23), q_n can now be approximated as

$$q_n = S_\lambda(z_n) \tag{24}$$

Since $S_\lambda(z_n)$ now is a differentiable expression approximating the quantizer, we can derive the expressions leading up to the elements in the Jacobian matrix $\mathbf{H}$ without problems. This is done by using the equations in the previous section, derived from Fig. 3, but with (6) replaced with (24) which results in a recursive calculation, i.e. $\partial q_n/\partial\theta_i$ is calculated as

$$\frac{\partial q_n}{\partial \theta_i} = \frac{\partial S_\lambda(z_n)}{\partial \theta_i} = \frac{2}{\pi}\frac{\partial z_n}{\partial \theta_i}\frac{\lambda}{1+(\lambda z_n)^2} \tag{25}$$

which after replacing (22) will enable the use of a gradient-based optimization approach.

The necessity of this differential approximation is clearly visible in Fig. 6, where a 1-dimensional error-surface is plotted, using several cases of the approximation $\mathcal{S}_\lambda$. The error surface is created by iteratively running the modulator with a predetermined signal sequence x_n over several settings of a parameter of choice. In this case, $a_2 \in H$ is swept. The same error-surface is further on shown for the case of a quantizer, which illustrates the fact stated in (22), e.g. that the partial derivative $\partial q_n/\partial\theta_i = 0$ almost everywhere. One observation easily made, is the locus of the local minimum as $\lambda \to \infty$. Thus, a progression over a sequence of λ-values is necessary for the algorithm to converge to a good, final minimum.

We will now continue with the description of the Monte-Carlo based approach in which the start-value sensitivity of the damped Gauss-Newton iteration is reduced. Further on, an optimization algorithm designed for selection of the set $\{\boldsymbol{\theta}, \alpha\}$ improving the NSC-performance of the $\Sigma\Delta$-modulator given a certain reconstruction-filter F is described and simulated.

C. Monte-Carlo based optimization algorithm

In this section we aim to develop an optimization algorithm based on the damped Gauss-Newton iteration. It is a commonly known fact that gradient-based search algorithms can be quite sensitive to local minima when it comes to convergence. In particular, the number of local minima will grow rapidly if the dimensionality of the problem is large. Identifying the global minimum in an optimization problem as nonlinear as the one described here could be considered practically impossible. Therefore, a simple Monte Carlo-based approach to the algorithm is developed in order to approach a good enough local minima. Further on, for computational simplicity and increased robustness, the damped Gauss-Newton is split up to be performed iteratively over three partitions of the parameter space, instead of over the complete space.

The proposed algorithm described here is illustrated in Fig. 7. The algorithm starts off by scattering M different parameter vectors using the selected start vector, $\boldsymbol{\theta}^{(1)}$ and Gaussian noise. This scattering provides the set $\boldsymbol{\Theta}^{(1)}$ containing M parameter vectors.

$$\boldsymbol{\Theta}^{(1)} = \{\boldsymbol{\theta}_i^{(1)}\}_{i=1}^{M} = \{\boldsymbol{\theta}^{(1)} + \mathbf{w}_i^{(1)}\}_{i=1}^{M} \tag{26}$$

where $\mathbf{w}_i^{(1)}$ is drawn from a Gaussian distribution

$$\mathbf{w}_i^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{w}^{(1)}}) \tag{27}$$

with variances

$$\mathbf{C}_{\mathbf{w}^{(1)}} = \operatorname{diag}\left(\sigma^2_{{w_1}^{(1)}}, \sigma^2_{{w_2}^{(1)}}, \ldots, \sigma^2_{{w_P}^{(1)}}\right) \tag{28}$$

Note that the notation $\boldsymbol{\theta}_i^{(k)}$ refers to the i:th parameter vector in $\boldsymbol{\Theta}^{(k)}$, of the k:th iteration of the algorithm. If the index i is left out, as in $\boldsymbol{\theta}^{(k)}$, the notation refers to the parameter vector used generating the set $\boldsymbol{\Theta}^{(k)}$, as in (26).

From each of the M generated vectors, we now perform the same optimization sequence shown in Fig. 7 until convergence. As further depicted in Fig. 7, the damped Gauss-Newton is executed sequentially over partitions of the parameter set $\boldsymbol{\theta}$, i.e. on H and G separately, to further improve the convergence properties. The optimization of α is then done separately by solving

$$\alpha_i^{(k)} = \arg\min_{\alpha} \frac{1}{N}\sum_{n=1}^{N} |x_n - y_n(\alpha, \boldsymbol{\theta}_i^{(k)})|^2 \tag{29}$$

This routine is then repeated until convergence. After these procedures are done for all parameter vectors $\boldsymbol{\theta}_i^{(k)} \in \boldsymbol{\Theta}^{(k)}$, the i:th parameter vector and scaling factor, $(\alpha_i^{(k)}, \boldsymbol{\theta}_i^{(k)})$, $1 \le i \le M$, that minimizes the normalized mean square error (NMSE), is selected as the next start vector,

$$(\alpha^{(k+1)}, \boldsymbol{\theta}^{(k+1)}) = \min_{1\le i\le M} \text{NMSE}(\alpha_i^{(k)}, \boldsymbol{\theta}_i^{(k)}) \tag{30}$$

where the NMSE is defined after reconstruction as

$$\text{NMSE}(\alpha, \boldsymbol{\theta}) = \frac{\sum_{n=1}^{N} |x_n - y_n(\alpha, \boldsymbol{\theta})|^2}{\sum_{n=1}^{N} |x_n|^2} \tag{31}$$

From this parameter vector, we then generate the set $\boldsymbol{\Theta}^{(k+1)}$ as described in (26) and repeat the procedure until we reach a k:th iteration of the algorithm that satisfies the convergence criteria

$$\text{NMSE}(\alpha^{(k-1)}, \boldsymbol{\theta}^{(k-1)}) - \text{NMSE}(\alpha^{(k)}, \boldsymbol{\theta}^{(k)}) < \varepsilon \tag{32}$$

for any sufficiently small ε. Further on, as the algorithm is moving toward a good minima, the variance $\mathbf{C}_{\mathbf{w}^{(k)}}$, is continuously decreased as k increases,

$$\mathbf{C}_{\mathbf{w}^{(k+1)}} \preceq \mathbf{C}_{\mathbf{w}^{(k)}} \tag{33}$$

where $\preceq$ denotes the component-wise matrix inequality. For numerical simplifications and in order to reduce computation time, the number of new start points generated, M, can be reduced for each iteration as well. To further decrease the overall computational time it is possible to start with a small number of samples N_1 for a large M, and sequentially increase the number of samples, $N^{(k)} < N^{(k+1)}$, used in each damped Gauss-Newton iteration while also decreasing the number of scattered vectors M in each iteration.

After convergence in NMSE is reached, the slope-factor λ is then increased and the optimization routine is reset to start over again by generating M new vectors from $\boldsymbol{\theta}^*$ according to (26). The algorithm is performed until a λ large enough is reached, i.e. when the approximation error is arbitrary small, thus making it possible to replace the approximation with the quantizer with good performance.

V. Simulation results

A. Prerequisites

In order to verify the performance of the suggested algorithm, the optimization routine and the generalized, IIR-filter based $\Sigma\Delta$-modulator representation was implemented in MATLAB using direct form 2 (DFII) IIR-filters, along with the damped Gauss-Newton search. The DFII-filter based modulator has a computational complexity of 1 delay-tap, 1 addition and 1 multiplication per parameter.

The input signal was generated using a low-pass filtered Gaussian noise sequence $\{x_n\}_{n=1}^{N} \sim \mathcal{N}(0, \sigma_x^2)$, where $\sigma_x^2 = 0.1$, resulting in $BW_S = f_s/10$, i.e. an effective OSR of 10. Further on, a wideband reconstruction filter with bandwidth $BW_F = f_s/4$ was used to reconstruct the signal (see Fig. 1). The optimization used iterations of λ in coarse steps, from $\lambda = 1$ up to $\lambda = 10^4$, after which the quantizer could be reinserted. The results for the generalized $\Sigma\Delta$ modulator are based on simulations using $P_H = 4$, $Q_H = 4$, $P_G = 4$ and $Q_G = 4$.

The simulations evaluating the final results are put in terms of reconstructed SQNR $= 1/\text{NMSE}$, since this is the most common figure of merit throughout the literature [26]. The performance of the proposed implementation is then compared against both 1^{st} and 2^{nd} order integrator-based lowpass $\Sigma\Delta$-modulators, as well as two $\Sigma\Delta$-modulators with optimized NTFs. The NTFs of these modulators has been optimized using the $\Sigma\Delta$-toolbox [28], described in detail in [26]. These two $\Sigma\Delta$-modulators will be labeled $\Sigma\Delta\#1$ and $\Sigma\Delta\#2$ from here on. The NTF of $\Sigma\Delta\#1$ were optimized using the constraint $\|NTF(z)\|_\infty \le 1.5$ and the NTF of $\Sigma\Delta\#2$ were optimized using the constraint $\|NTF(z)\|_\infty \le 1.75$. In order to achieve a fair comparison, the order of the NTFs in both cases were set to 7, which is the order of the resulting NTF in the proposed modulator.

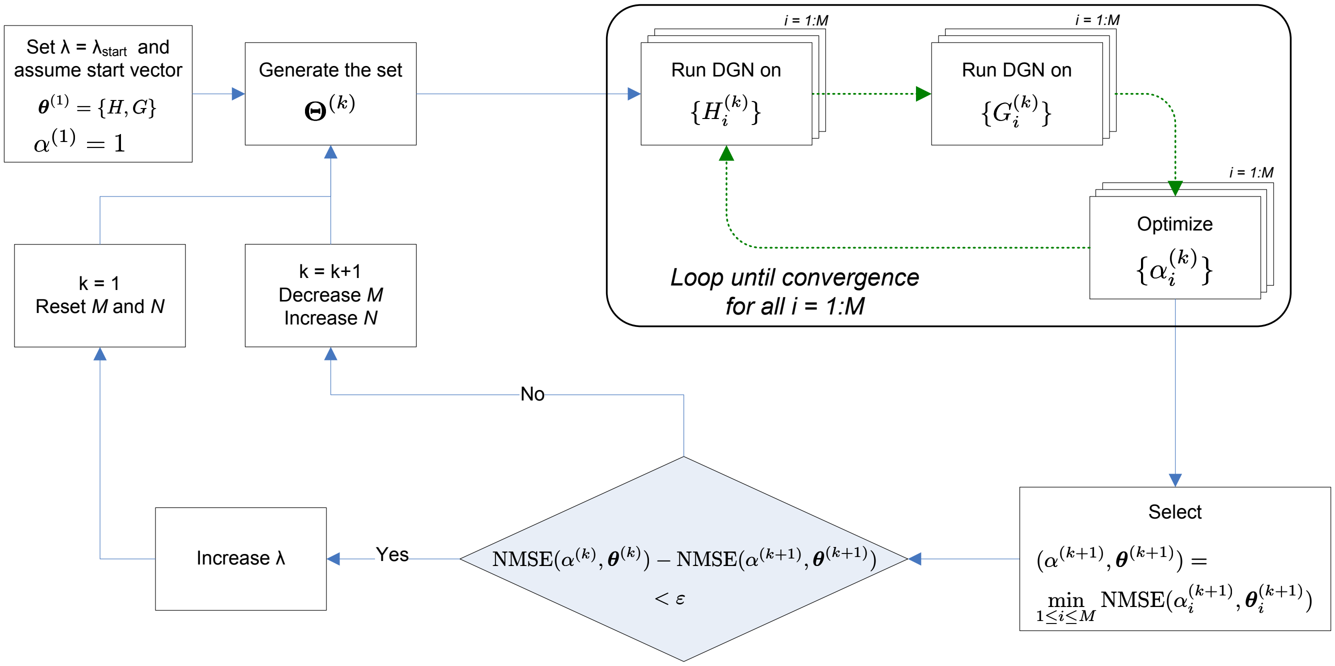


Fig. 7: Block diagram of the proposed algorithm which sequentially uses the damped Gauss-Newton (DGN), in combination with a Monte-Carlo based approach, to minimize the quantization noise within the selected bandpass region.

B. Simulations of the proposed algorithm

Fig. 8 shows one example of the proposed method described in section IV-C. The figure illustrates, over three iterations of k at $\lambda = 5$, how the suggested algorithm overcomes the start-value sensitivity by iteratively forcing the set of parameters toward a good minimum. Note the decreasing number of start-points per iteration, M, and the increasing sequence length, N, as presented in Table I. For simplicity, the same variance is used for all positions when generating the M vectors, e.g. $\sigma^2_{w_1(k)} = \sigma^2_{w_2(k)} = \ldots = \sigma^2_{w_P(k)}$ (given by $\sigma^2_{\mathbf{w}}$ in Table I). Simulations clearly illustrate how the algorithm results in reduced variance of the estimate and thus a reduction of the start-value sensitivity. The figure shows that, for this λ setting, the minimum NMSE that can be obtained is approximately -25 dB. It should be noted that, in the following iterations of the algorithm, as $\lambda \to \infty$ and approaches the true quantizer, the minimum NMSE is higher.

A spectrum plot of the final optimized NSC is shown in Fig. 9, where it is compared to the other modulator implementations as described in V-A. It is clearly shown in Fig. 9, as well as within the SQNR-calculations in Table II, that the quantization noise power within the reconstruction filter passband is significantly reduced in comparison with the other modulators.

TABLE I: Summary of simulation parameters for Fig. 8

Iteration k	Sequence length N	Nr of scattered initial vectors M	Variance $\sigma^2_{\mathbf{w}}$
1	100	300	1
2	500	200	0.8
3	1000	100	0.4

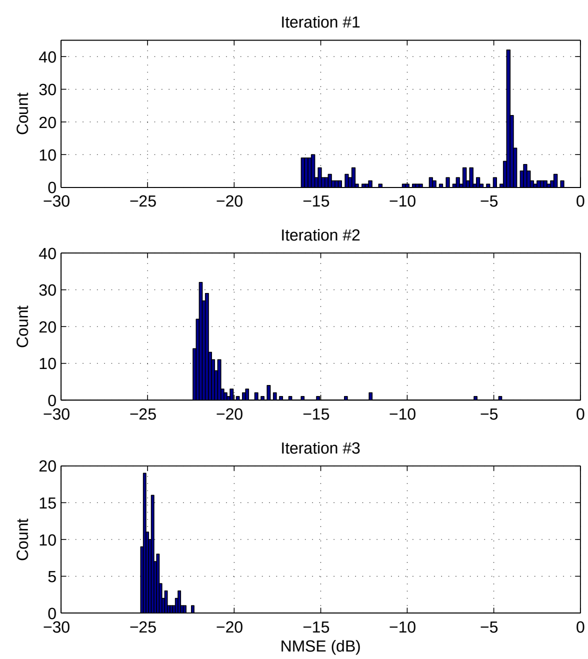
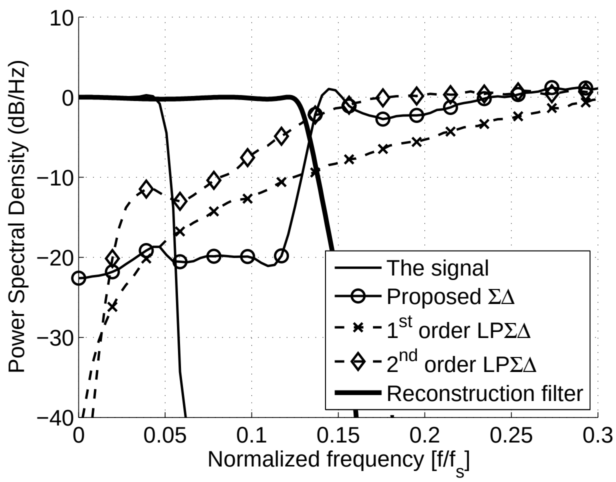


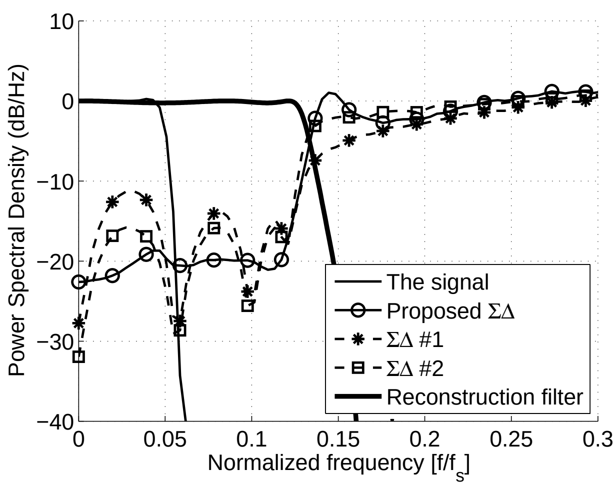
Fig. 8: One example of three iterations in the Monte-Carlo based algorithm at $\lambda = 5$. Decreasing noise-variance $\mathbf{C}_{\mathbf{w}^{(k)}}$ and increased sequence length is used according to Table I.

C. Signal- and Noise-Transfer Functions

Using a linear Gaussian noise approximation of the quantizer [15], the Signal Transfer Function (STF) and the Noise Transfer Function (NTF) of the generalized $\Sigma\Delta$ modulator can be derived. Given that $H(z) = B(z)/A(z)$ and $G(z) =$



(a)



(b)

Fig. 9: A comparison of the PSD of the quantization noise produced by the proposed $\Sigma\Delta$-modulator in comparison with (a) regular integrator-based 1^{st} and 2^{nd} order $\Sigma\Delta$-modulators, and with (b) $\Sigma\Delta\#1$ and $\Sigma\Delta\#2$ as described in V-A. Both plots also shows the original signal as well as the magnitude frequency response of the reconstruction filter.

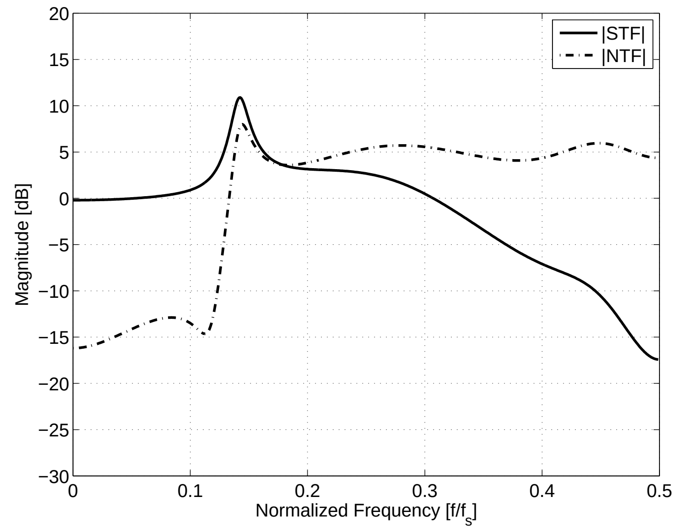


Fig. 10: Calculated magnitude signal- and noise-transfer functions, |STF| and |NTF|, for the proposed modulator.

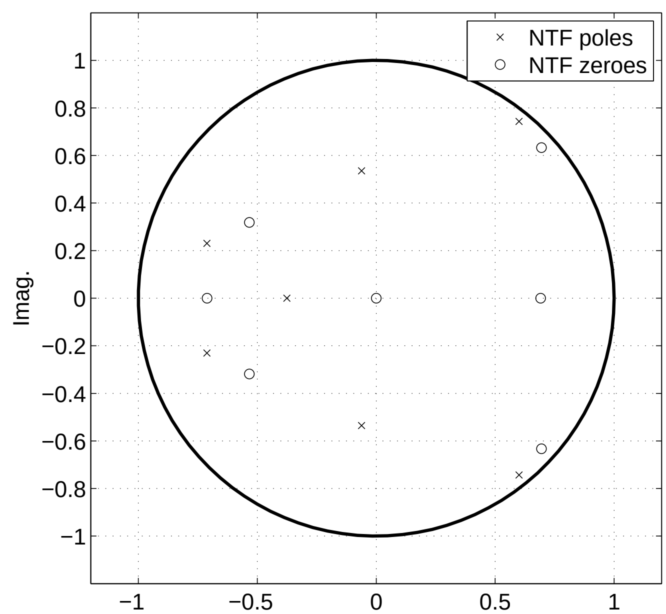


Fig. 11: Poles and zeroes of the NTF for the proposed modulator.

$D(z)/C(z)$, we end up with the expressions

$$\begin{aligned} \text{STF}(z) &= \frac{1}{1+z^{-1}H(z)G(z)} \\ &= \frac{A(z)C(z)}{A(z)C(z)+z^{-1}B(z)D(z)} \end{aligned} \tag{34}$$

and

$$\begin{aligned} \text{NTF}(z) &= \frac{H(z)}{1+z^{-1}H(z)G(z)} \\ &= \frac{B(z)C(z)}{A(z)C(z)+z^{-1}B(z)D(z)} \end{aligned} \tag{35}$$

from which we can calculate STF and NTF using the optimized generalized $\Sigma\Delta$ parameter-set, $\boldsymbol{\theta}^*$. The result is presented in the magnitude plot Fig. 10 and the pole-zero diagram in Fig. 11. As can be seen, the calculated NTF shows resemblance to the quantization noise presented in Fig. 9, which indicates that a linear analysis could be useful for providing starting values to the optimization algorithm and thus further improve its robustness.

D. Signal to Quantization Noise Ratio and Modulator Robustness Simulations

In order to extract the SQNR characteristics for the optimized modulator, several new band-limited data-sets $\{x_n\}_{n=1}^N \sim \mathcal{N}(0, \sigma_x^2)$ of length $N = 100000$ were generated with σ_x^2 ranging from 10^{-3} to 1. These data-sets were encoded and reconstructed using both the proposed generalized $\Sigma\Delta$

with parameters $\boldsymbol{\theta}^*$, as well as the modulator implementations listed in V-A. In the case of the 2nd order integrator-based modulator, the constant gain-factors were set to $a = b = 1$ (see Fig. 4).

The simulated results in terms of maximum SQNR after reconstruction is shown in Table II, where a 1.7 dB improvement of SQNR is shown compared to $\Sigma\Delta\#2$. The SQNR characteristics are shown in Fig. 12 as a function of the input signal variance, σ_x^2. The figure also shows the points at which $\Sigma\Delta\#2$ and the 2^{nd} order $\Sigma\Delta$-modulator becomes instable. The 1^{st} order integrator-based modulator, $\Sigma\Delta\#1$ and the generalized $\Sigma\Delta$-modulator stays stable, even when severely overdriven. Despite the high order of modulator, the generalized $\Sigma\Delta$-structure has a relatively large stable input amplitude-range due to the use of the optimization algorithm. The algorithm inherently eliminates parameter-sets $\boldsymbol{\theta}$ causing instability since these will score poorly with the optimization cost-function.

The need of an optimization based approach can be further motivated by calculating the maximum NTF-gain, or $\|NTF(\omega)\|_\infty$, of the optimized modulator. Generally, $\|NTF(\omega)\|_\infty < 1.5$ is required for a stable 1-bit modulator [26], [29]. Calculating the NTF stated in (35) for the optimized modulator, we get that $\|NTF(\omega)\|_\infty \approx 2.5$. This illustrates how the linear approximation-based design methods are likely to rule out high performance modulators such as the one proposed in this paper, thus providing a strong motivation for an optimization based approach.

VI. Conclusion

A generalized structure able of representing a large set of $\Sigma\Delta$-modulator implementations has been suggested. Using this representation, and a differentiable approximation of the quantizer, an algorithm for minimization of the quantization noise within a custom frequency band has been proposed. The algorithm combines this approximation with a Monte Carlo approach in order to decrease the start value sensitivity.

Simulations of a generalized, low OSR $\Sigma\Delta$-modulator have been used to demonstrate that significant improvements in reconstructed SQNR can be obtained with a optimized, generalized $\Sigma\Delta$-modulator. These results were compared to the regular 1st and 2nd order modulators, as well as against two modulators with NTFs optimized using the classical AWGN-approximation. Increased robustness in terms of modulator stability were also shown by studying the SQNR performance over a wide range of input signal variance.

These results implies that when used in pulsed RF transmitter architectures, the requirements for narrowband reconstruction filters can be relaxed, or the switching frequency reduced. In either case the result is that the transmitter efficiency can be improved at the cost of a small increase in modulator complexity.

TABLE II: Summary of simulated results.

Modulator Type	Max. SQNR (dB)
Optimized, generalized $\Sigma\Delta$ modulator	16.6
$\Sigma\Delta\#1$	13.5
$\Sigma\Delta\#2$	14.9
1st order $\Sigma\Delta$	12.2
2nd order $\Sigma\Delta$	7.3

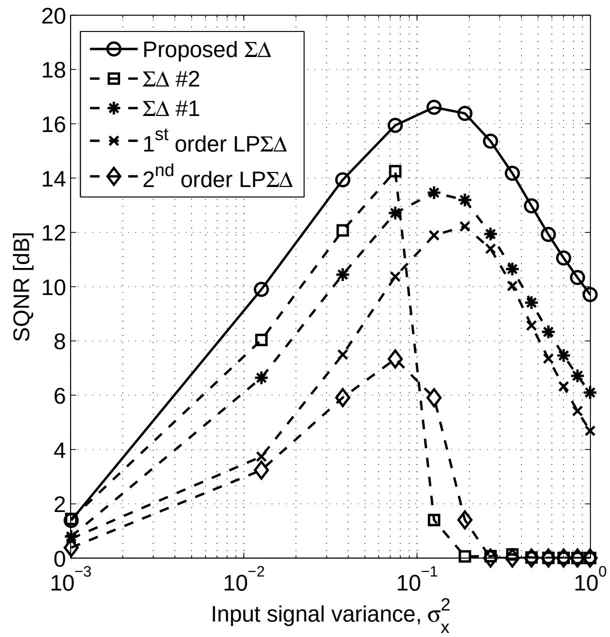


Fig. 12: Signal to Quantization Noise Ratio (SQNR) vs input signal variance (σ_x^2) for the benchmarked $\Sigma\Delta$ implementations.

Acknowledgment

The authors of this paper would like to thank professor Tomas McKelvey for the fruitful discussions around the damped Gauss-Newton and other optimization-related issues.

References

[1] W. Doherty, “A new high efficiency power amplifier for modulated waves,” *Proceedings of the IRE*, vol. 24, no. 9, pp. 1163–1182, Sept. 1936.

[2] H. Chireix, “High power outphasing modulation,” *Proceedings of the IRE*, vol. 23, no. 11, pp. 1370–1392, Nov. 1935.

[3] F. Raab, “Split-band modulator for kahn-technique transmitters,” *Microwave Symposium Digest, 2004 IEEE MTT-S International*, vol. 2, pp. 887–890 Vol.2, June 2004.

[4] ——, “Radio frequency pulsewidth modulation,” *IEEE Trans. Commun.*, vol. 21, no. 8, pp. 958– 966, Aug. 1973.

[5] T. Blocher and P. Singerl, “Coding efficiency for different switched-mode rf transmitter architectures,” *Circuits and Systems, Midwest Symposium on*, vol. 0, pp. 276–279, 2009.

[6] R. Schreier, “Noise-shaped coding,” Ph.D. dissertation, University of Toronto, Toronto, Canada, May 1991. [Online]. Available: http://www.dissonance.com/archive/phd/schreier.pdf

[7] N. Jayant and P. Noll, *Digital coding of waveforms*. Prentice Hall, Signal Processing Series, ISBN 0-13-211913-7, 1984.

[8] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Data Converters Theory, Design and Simulation*. New York: IEEE Press., 1992.

[9] J. Choi, J. Yim, J. Yang, J. Kim, J. Cha, D. Kang, D. Kim, and B. Kim, “A $\Delta\Sigma$ - Digitized Polar RF Transmitter,” *IEEE Trans. Microwave Theory Tech.*, vol. 55, no. 12, pp. 2679–2690, Dec. 2007.

[10] C. Gunturk and N. Thao, “Refined error analysis in second-order $\Sigma\Delta$ modulation with constant inputs,” *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 839–860, May 2004.

[11] N. Thao, "Overview on a new approach to one-bit nth order $\Sigma\Delta$ modulation," *Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on*, vol. 1, pp. 623–626 vol. 1, May 2001.

[12] R. Gopalan and O. M. Collins, "An optimization approach to single-bit quantization," *IEEE Trans. Circuits Syst. I*, vol. 56, no. 12, pp. 2655–2668, December 2009.

[13] R. Schreier, "An empirical study of high-order single-bit $\Delta\Sigma$ modulators," *IEEE Trans. Circuits Syst. II*, vol. 40, no. 8, pp. 461–466, Aug 1993.

[14] R. Gray, W. Chou, and P. Wong, "Quantization noise in single-loop $\Sigma\Delta$ modulation with sinusoidal inputs," *IEEE Trans. Commun.*, vol. 37, no. 9, pp. 956–968, Sep 1989.

[15] R. Gray, "Spectral analysis of quantization noise in a single-loop $\Sigma\Delta$ modulator with dc input," *IEEE Trans. Commun.*, vol. 37, no. 6, pp. 588–599, Jun 1989.

[16] S. Pamarti, J. Welz, and I. Galton, "Statistics of the quantization noise in 1-bit dithered single-quantizer digital deltasigma modulators," *IEEE Trans. Circuits Syst. I*, vol. 54, no. 3, pp. 492–503, March 2007.

[17] S. Jantzi, R. Schreier, and M. Snelgrove, "Bandpass $\Sigma\Delta$ analog-to-digital conversion," *IEEE Trans. Circuits Syst. II*, vol. 38, no. 11, pp. 1406–1409, Nov 1991.

[18] T. Johnson and S. Stapleton, "Comparison of bandpass $\Sigma\Delta$ modulator coding efficiency with a periodic signal model," *IEEE Trans. Circuits Syst. II*, vol. 55, no. 11, pp. 3763–3775, Dec. 2008.

[19] F. H. Raab, P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and microwave power amplifier and transmitter technologies - part 5," *High Frequency Electronics*, vol. 3, no. 1, pp. 46 – 54, Jan. 2004.

[20] P. Kenington, *RF and Baseband Techniques for Software Defined Radio*. Boston: Artech House Publishers, 2005.

[21] C. Berland, I. Hibon, J. Bercher, M. Villegas, D. Belot, D. Pache, and V. Le Goascoz, "A transmitter architecture for nonconstant envelope modulation," *IEEE Trans. Circuits Syst. II*, vol. 53, no. 1, pp. 13–17, Jan. 2006.

[22] A. Dupuy and Y. Wang, "High efficiency power transmitter based on envelope $\Delta\Sigma$ modulation (edsm)," *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 3, pp. 2092–2095 Vol. 3, Sept. 2004.

[23] T. Matsuura and H. Adachi, "A high efficiency transmitter with a Delta-Sigma modulator and a noise cancellation circuit," in *European conference on wireless technology*, 2004.

[24] U. Gustavsson, T. Eriksson, and C. Fager, "A general method for passband quantization noise suppression in pulsed transmitter architectures," in *Microwave Symposium Digest, 2009. MTT '09. IEEE MTT-S International*, June 2009, pp. 1529–1532.

[25] T. Johnson, K. Mekechuk, D. Kelly, and J. Lu, "Asynchronous modulator for linearization and switch-mode rf power amplifier applications," in *Radio Frequency Integrated Circuits Symposium, 2009. RFIC 2009. IEEE*, June 2009, pp. 185–188.

[26] R. Schreier and G. Temes, *Understanding Delta-Sigma Data Converters*. New Jersey: Wiley Interscience, 2005.

[27] S. Kay, *Fundamentals of statistical signal processing - Volume 1: Estimation*. New Jersey: Prentice Hall Signal Processing Series, 1993.

[28] R. Schreier, "MATLAB Delta Sigma Toolbox," Available at http://www.mathworks.com/matlabcentral/fileexchange/19.

[29] S. Norsworthy, R. Schreier, and G. Temes, *Delta-Sigma Data Converters - Theory, Design, and Simulation*. New Jersey: John Wiley and Sons, Inc., 1997.