

# An Optimization-Based Approach to One-Bit Quantization

Florian Mayer

FH JOANNEUM - University of Applied Sciences

Graz, Austria

florian.mayer@fh-joaanneum.at

Christian Vogel

FH JOANNEUM - University of Applied Sciences

Graz, Austria

christian.vogel@fh-joaanneum.at

**Abstract**—In many industrial fields, one-bit signals play an important role, such as improving the operational efficiency in a wide range of applications. This paper introduces a mapping function  $\mathcal{F}(\cdot)$  that converts real-valued discrete-time signals into two-level discrete-time signals, i.e., one-bit signals. By introducing an error function, we can evaluate the quality of the mapping and can introduce different optimization strategies to minimize the error. Mathematical analysis and numerical simulations demonstrate that the optimization approach leads to specific characteristics of the quantization. The well-known Sigma-Delta modulator, for example, turns out to be a special case of the introduced optimization. The approach has the potential to introduce new ways of one-bit quantization, which can be tailored to specific applications.

**Index Terms**—one-bit quantization, one-bit signal processing, sigma-delta modulation, pulse width modulation, optimization-based framework

## I. INTRODUCTION

One-bit signals are either “on” or “off” at any given time. The use of one-bit signals makes it easier to encode, decode, and process signals with higher efficiency because the two-level signal characteristic reduces the energy required for signal processing as well as signal transmission [1]–[4].

For example, using one-bit signals in burst-mode RF transmitters allows the power amplifiers to reach peak efficiency and do not waste power during low-level periods, resulting in a higher average efficiency compared to conventional linear power amplifier [5]–[10]. In audio processing, a one-bit signal enables high sampling rates crucial for superior sound quality. It also simplifies audio processing by reducing the need for complex filtering and signal processing [11]–[13].

To convert real-valued data into a one-bit signal, temporal patterns such as pulse width, pulse position, and pulse density modulation are used. Despite their utility, these methods can have limitations in representation, reconstruction, noise sensitivity, and dynamic range [1]. The Sigma-Delta Quantizer (SDQ) achieves high resolution through oversampling and via noise shaping [12], [14], but introduces high-frequency switching that reduces the power efficiency for example in switched-mode power amplifiers. Pulse-Width Modulation (PWM) enables efficient power management but introduces

This work was funded by the Austrian Science Fund (FWF) DFH 5 and the province of District, Styria.

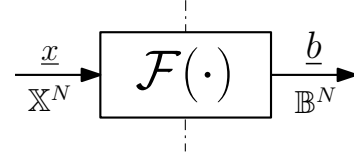


Fig. 1. The real-valued signal  $\underline{x}$  is mapped into a one-bit signal  $\underline{b}$ .

distortion and aliasing due to the nonlinear PWM operator [5]. Aliasing can be circumvented by aliasing-free digital PWM (AFPWM) [5], [15], but at the cost of much higher implementation complexity [6]. Click Modulation produces a one-bit signal through equal clicks, creating a bandpass signal recoverable with a low-pass filter [16], [17]. However, uncertainties in the switching times can affect the performance, i.e., the signal-to-noise ratio, and require high implementation complexity to mitigate them [2].

This paper deals with the feasibility of an optimal one-bit representation for real-valued signals. It introduces an optimization-based one-bit-quantization (OBAQ) approach that converts real signals to their two-level representation of  $\{-1, 1\}$ . This optimization produces an optimal one-bit signal for a real-valued input, despite the problem’s general NP-hard nature. The paper highlights that this framework can be used to derive and explore less complex suboptimal solutions such as iterative methods [18]–[20]. It turns out that our optimization approach easily leads to the sigma-delta modulator as a particular solution. This framework is therefore a starting point for finding other optimal and suboptimal solutions for one-bit quantization with specific properties in frequency and time.

## II. PROBLEM DEFINITION

We consider a function  $\mathcal{F}(\cdot)$  mapping a real-valued discrete-time signal  $\underline{x} = [x_0, x_1, \dots, x_n, \dots, x_{N-1}]^T \in \mathbb{X}^N = [-\alpha, \alpha]^N$  with  $\alpha \in \mathbb{R}_0^+$  and a signal length  $N$ , into a discrete-time one-bit (two-level) signal  $\underline{b} = [b_0, b_1, \dots, b_n, \dots, b_{N-1}]^T \in \mathbb{B}^N = \{-1, 1\}^N$ , that is

$$\underline{b} = \mathcal{F}(\underline{x}). \quad (1)$$

Furthermore, we introduce a measure  $E(\underline{x}, \underline{b})$  between the signal  $\underline{x}$  and the signal  $\underline{b}$ . The measure  $E(\underline{x}, \underline{b})$  could be arbitrarily defined according to the given problem. For many practical problems, however, the physical motivated minimization of an

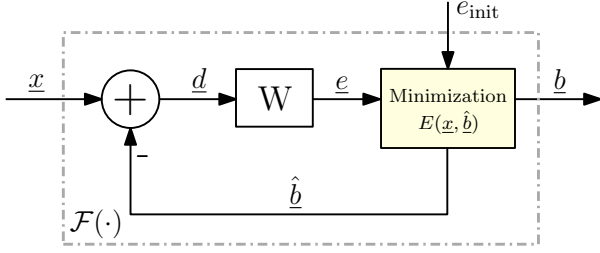


Fig. 2. General structure of the optimization-based mapping function  $\mathcal{F}(\cdot)$ , which generates a one-bit signal from a real-valued discrete-time signal.

error signal energy is used, leading to the squared  $l^2$ -norm. Therefore, for the remaining of this paper, we consider the squared  $l^2$ -norm which is given by

$$E(\underline{x}, \underline{b}) = \|\underline{e}\|_2^2 = \sum_{n=0}^{N-1} |e_n|^2 \quad (2)$$

where the error signal  $\underline{e} \in \mathbb{R}^N$  is defined as the difference between  $\underline{x}$  and  $\underline{b}$ , weighted with matrix  $W$ , consisting of the coefficients  $\underline{w} = [w_0, w_1, \dots, w_n, \dots, w_{N-1}] \in \mathbb{R}^N$ , i.e.,

$$\underline{e} = W \cdot (\underline{x} - \underline{b}). \quad (3)$$

The difference between  $\underline{x}$  and  $\underline{b}$  can be denoted as

$$\underline{d} = \underline{x} - \underline{b} \quad (4)$$

and results with (3) and (2) in

$$E(\underline{x}, \underline{b}) = \|\underline{e}\|_2^2 = \|W \cdot (\underline{x} - \underline{b})\|_2^2 = \|W \cdot \underline{d}\|_2^2. \quad (5)$$

Ideally, the mapping function (1) would minimize the error function given in (5), but for practical cases the mapping results in errors larger than the minimum error, therefore i.e., substituting the mapping function (1) into (5) results in

$$E(\underline{x}, \mathcal{F}(\underline{x})) \geq \min_{\hat{\underline{b}} \in \mathbb{B}^N} E(\underline{x}, \hat{\underline{b}}), \quad (6)$$

where  $\hat{\underline{b}} \in \mathbb{B}^N$  is the one-bit solution obtained by the optimization process.

### III. GENERAL OPTIMIZATION

The mapping defined in (1) can be implemented by an optimization approach as shown in Figure 2.

The function  $\mathcal{F}(\cdot)$  is nonlinear, as it maps an infinite to a finite cardinality, specifically  $\mathbb{R} \rightarrow \mathbb{B}$ . Finding an optimal one-bit representation  $\underline{b}$ , even when dealing with a single signal  $\underline{x}$ , has been shown to be a NP-hard problem [18]. Thus, minimizing  $E(\underline{x}, \hat{\underline{b}})$  with a signal length of  $N$  introduces  $2^N$  possible binary combinations to be considered, resulting in an exponential growth of the search space and high computational complexity. Hence, in practice, even with fast algorithms from discrete optimization [21], [22] we are limited to small signal lengths  $N$ . By using heuristic optimization like genetic algorithms [23] and simulated annealing [24] we may find good local optima, but still need substantial computational resources. To significantly reduce the computational complexity, we need to simplify our optimization problem for finding  $\mathcal{F}(\cdot)$ .

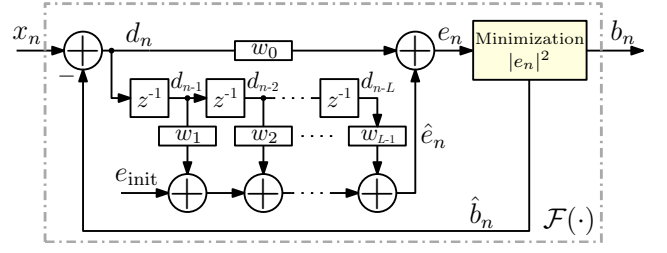


Fig. 3. The representation of  $\mathcal{F}(\cdot)$  sequential optimization process that minimizes the error  $|e_n|^2$  at time instances  $n$ .

### IV. SEQUENTIAL OPTIMIZATION

To obtain more feasible solutions, we replace (2) by the Instantaneous error  $|e_n|^2$  and minimize over  $b_n$ , i.e.,

$$b_n = \arg \min_{\hat{b}_n} |e_n|^2. \quad (7)$$

At time instance  $n$  we assume all  $e_k$  for  $k < n$  have already been optimized and the corresponding  $b_k$  have been determined. In contrast to optimize all elements of  $\underline{b}$  at once, we sequentially optimize the Instantaneous error just by one element  $b_n$  at time  $n$  while, keeping the former coefficients  $b_k$  fixed. To end up with some realizable examples, we further assume a causal time-invariant filter<sup>1</sup> with the coefficient vector  $\underline{w}$  reducing the matrix  $W$  to

$$W = \begin{bmatrix} w_0 & 0 & \cdots & 0 \\ w_1 & w_0 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ w_{N-1} & \cdots & w_1 & w_0 \end{bmatrix} \quad w_0 \neq 0, \quad (8)$$

where all elements above the main diagonal are zero, and arbitrary entries over  $\mathbb{R}$  are located on or below the main diagonal.

With these simplifications the error in (3) reduces to

$$\begin{aligned} e_0 &= w_0 \cdot d_0 + \hat{e}_0 \\ e_1 &= w_0 \cdot d_1 + \hat{e}_1 \\ &\vdots \\ e_n &= w_0 \cdot d_n + \hat{e}_n = w_0 \cdot (x_n - \hat{b}_n) + \hat{e}_n, \end{aligned} \quad (9)$$

where  $\hat{e}_n$  is the convolution of the differences  $d_k = x_k - b_k$ , determined by the already obtained  $b_k$  of previous steps, with the coefficients  $w = [w_1, \dots, w_{N-1}]$  of matrix  $W$  given by

$$\hat{e}_n = \sum_{k=0}^{n-1} w_{n-k} \cdot d_k + e_{\text{init}}, \quad (10)$$

where  $e_{\text{init}}$  is an optional initial offset. Applying (9) and (10) to (7) leads to

$$b_n = \arg \min_{\hat{b}_n \in \mathbb{B}} \left( x_n - \hat{b}_n + \frac{\hat{e}_n}{w_0} \right)^2, \quad (11)$$

<sup>1</sup>While this discussion assumes time-invariant filters for simplicity, the principles can also be applied to time-varying filters.

where  $\hat{b}_n$  was made explicit by dividing  $w_0$ . Since  $\hat{b}_n$  can take only two values (i.e.  $-1$  and  $1$ ), (11) can be shown to be

$$b_n = \text{sgn}_0\left(x_n + \frac{\hat{e}_n}{w_0}\right) = \begin{cases} 1, & \left(x_n + \frac{\hat{e}_n}{w_0}\right) \geq 0 \\ -1, & \left(x_n + \frac{\hat{e}_n}{w_0}\right) < 0 \end{cases}, \quad (12)$$

where each element  $\hat{b}_n$  is updated by mapping  $(x_n + \frac{\hat{e}_n}{w_0})$  to  $\mathbb{B}$ , similarly to a comparator.

The lower-triangular structure of matrix  $W$  as well as the sequential process yields an optimization similar to the Coordinate Descent method outlined in [25]. The minimization of the squared  $l^2$ -norm of the error signal  $e_n$  in (7) contributes to the minimization of the global  $l^2$ -norm of the error signal as given in (2) as well. It might therefore serve as a good initial vector for a further general optimization. In the following, we examine some examples exploiting the introduced sequential optimization.

#### A. Identity Matrix

We investigate the filter matrix  $W$  being the identity matrix  $\mathcal{I}$ , where all coefficients on the main diagonal are 1, and all other elements are 0 such that

$$W_{\mathcal{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}. \quad (13)$$

Employing  $W_{\mathcal{I}}$  in (11) and setting  $e_{\text{init}} = 0$  yields

$$b_n = \arg \min_{\hat{b}_n \in \mathbb{B}} (x_n - \hat{b}_n)^2, \quad (14)$$

where  $\hat{e}_n = 0$ , since all coefficients  $w_{n-k}$  in (10) are zero as well. Consequently, (12) results in

$$b_n = \text{sgn}_0(x_n) = \begin{cases} 1, & x_n \geq 0 \\ -1, & x_n < 0 \end{cases}. \quad (15)$$

This is the simply two-level quantization of the signal  $x$ . It is, however, remarkable that for the identity matrix  $W_{\mathcal{I}}$  the sequential optimization achieves also a global optimum in (6), i.e.,

$$E(\underline{x}, F(\underline{x})) = \min_{\underline{b} \in \mathbb{B}^N} E(\underline{x}, \underline{b}) \quad \text{for } W = W_{\mathcal{I}}, \quad (16)$$

since in (5) any error  $e_n$  only depends on  $b_n$ .

#### B. Lower Triangular Matrix $\Delta$

We consider a lower-triangular matrix  $\Delta$  as filter matrix  $W$  given by

$$W_{\Delta} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & 1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{bmatrix}, \quad (17)$$

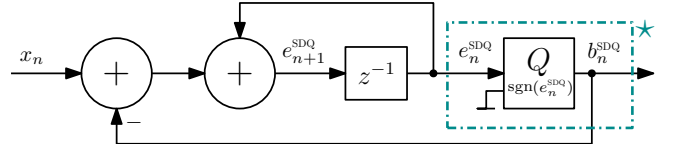


Fig. 4. General structure of a first-order Sigma-Delta Quantizer (SDQ).

where all coefficients on or below the main diagonal are set to 1, all other elements are set to 0. Plugging  $W_{\Delta}$  into (9) leads to

$$e_n = x_n - \hat{b}_n + \hat{e}_n, \quad (18)$$

and into (11) yields

$$b_n = \arg \min_{\hat{b}_n \in \mathbb{B}} (x_n - \hat{b}_n + \hat{e}_n)^2, \quad (19)$$

where the error  $\hat{e}_n$  in (10) is obtained by

$$\hat{e}_n = \sum_{k=0}^{n-1} d_k + e_{\text{init}} = \sum_{k=0}^{n-1} x_k - \sum_{k=0}^{n-1} \hat{b}_k + e_{\text{init}}. \quad (20)$$

Accordingly, by employing  $W_{\Delta}$  to (12) and (11) we get

$$b_n = \text{sgn}_0(x_n + \hat{e}_n) = \begin{cases} 1, & x_n + \hat{e}_n \geq 0 \\ -1, & x_n + \hat{e}_n < 0 \end{cases}. \quad (21)$$

The resulting error of the sequential optimization in (18) and the resulting binary output is identical to that of the first-order Sigma-Delta Quantizer (SDQ) shown in Figure 4 and given by [12]

$$e_{n+1}^{\text{SDQ}} = x_n - b_n^{\text{SDQ}} + e_n^{\text{SDQ}} \quad (22)$$

$$b_n^{\text{SDQ}} = \text{sgn}_0(e_n^{\text{SDQ}}) = \begin{cases} 1, & e_n^{\text{SDQ}} \geq 0 \\ -1, & e_n^{\text{SDQ}} < 0 \end{cases}. \quad (23)$$

By using the current error  $e_n^{\text{SDQ}}$  and current binary value  $b_n^{\text{SDQ}}$  of the Sigma-Delta Quantizer (highlighted by  $\star$  in Figure 4) as the initial value  $\hat{e}_0$  in (18), i.e.,  $\hat{e}_0 = e_{\text{init}} = e_0^{\text{SDQ}} - b_0^{\text{SDQ}}$ , we can rewrite (18) with (22) as

$$\begin{aligned} e_0^2 &= (x_0 + e_0^{\text{SDQ}} - b_0^{\text{SDQ}} - \hat{b}_0)^2 = (e_1^{\text{SDQ}} - \hat{b}_0)^2 \\ e_1^2 &= (e_2^{\text{SDQ}} - \hat{b}_1)^2 \\ &\vdots \\ e_n^2 &= (e_{n+1}^{\text{SDQ}} - \hat{b}_n)^2. \end{aligned} \quad (24)$$

Thus, we obtain the next one-bit sample of the first-order SDQ by applying the sequential optimization at time  $n$ , i.e.,  $\hat{b}_n = b_{n+1}^{\text{SDQ}}$ . Therefore, the first-order SDQ can be seen as the sequential optimization using the weighting matrix  $W_{\Delta}$ . Hence, the optimization approach may lead to new structures, which maybe have for example a better trade-off between noise shaping and filter complexity.

#### V. NUMERICAL SIMULATIONS AND ENHANCEMENTS

To verify and evaluate the derivations for the sequential optimization-based mapping function  $\mathcal{F}(\cdot)$  in Section IV,

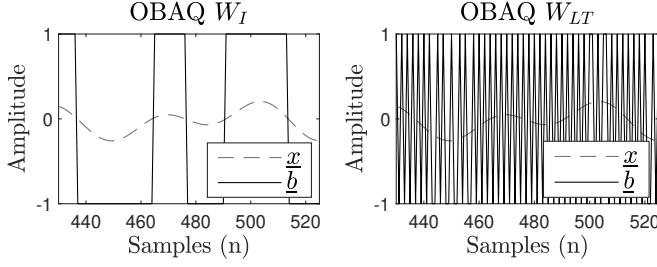


Fig. 5. Applying  $W_I$  we obtain the One-bit Quantizer (left), while applying  $W_{\setminus}$  yields the first-order SDQ, for  $e_{\text{init}} = 0$ .

numerical simulations have been performed. Based on the choice of  $W$  in Section IV-A and Section IV-B we obtain the One-bit Quantizer and the first-order SDQ (as shown in Figure 5) respectively. To demonstrate the optimization-based approach for one-bit-quantization (OBAQ), we exemplify a basic extension of the structure by simply plugging the coefficients of a causal non-ideal FIR lowpass filter  $\underline{h} = [h_0, h_1, \dots, h_{L-1}]^T \in \mathbb{R}^N$ , where the filter length  $L = 83 < N$ , into the matrix  $W$  to create the convolution matrix  $W_H$ . For our simulations, we use four sets of different signal lengths  $N$ , specifically  $N = [2^9, 2^{10}, 2^{11}, 2^{12}]$ . Each set contains 1000 band-limited test signals  $\underline{x} \in [-1, 1]^N$ , comprising frequency bins from 1 to 131, in steps of 3. Therefore, we scale both, the oversampling factor and the number of samples  $N$ . The oversampling factor varies within each set, maintaining the maximum frequency bin at 131. The cut-off for  $\underline{h}$  is defined as  $w_c = \frac{262}{N}\pi$ .

For all implementations,  $e_{\text{init}}$  has been set to 0. The one-bit quantization results are evaluated by investigating the Signal-to-Noise Ratio (SNR) [26], incorporating an ideal reconstruction filter at  $w_r = \frac{264}{N}\pi$  for the SNR calculation. The obtained results are averaged over all 1000 SNR for each length set, as shown in Table I.

TABLE I  
PERFORMANCE OF THE ONE-BIT QUANTIZATION METHODS

|                                          | Average over 1000 results per signal length $N$ |              |              |              |
|------------------------------------------|-------------------------------------------------|--------------|--------------|--------------|
|                                          | $N = 2^9$                                       | $N = 2^{10}$ | $N = 2^{11}$ | $N = 2^{12}$ |
| SNR <sub>OBAQ</sub> $W_{\setminus}$ (dB) | -0.46                                           | 7.16         | 15.79        | 24.76        |
| SNR <sub>OBAQ</sub> $W_H$ (dB)           | 2.92                                            | 9.77         | 17.70        | 26.05        |
| Oversampl. Factor                        | 1.95                                            | 3.91         | 7.82         | 15.63        |

The use of  $W_H$ , in particular the lowpass filter characteristics, of course leads the OBAQ to push the noise outside of the band specified by  $w_c$ . Based on Figure 6 and Table I, it can be seen that even at lower oversampling rates, the SNR of the quantized signal can be further improved.

The results indicate that, in the case of OBAQ, the selection of  $W$  can simulate the behavior of the quantizer without re-modifying the algorithmic structure of the system.

We further examine the use of 3 initial vectors: a random vector  $v_{\text{rand}}$ , OBAQ with  $W_{\setminus}$ , as well as  $W_H$  for the general optimization process, MATLAB's Genetic Algorithm (GA) [27], with a limited processing time of 60 seconds.

Therefore, future works could benefit from a quick first

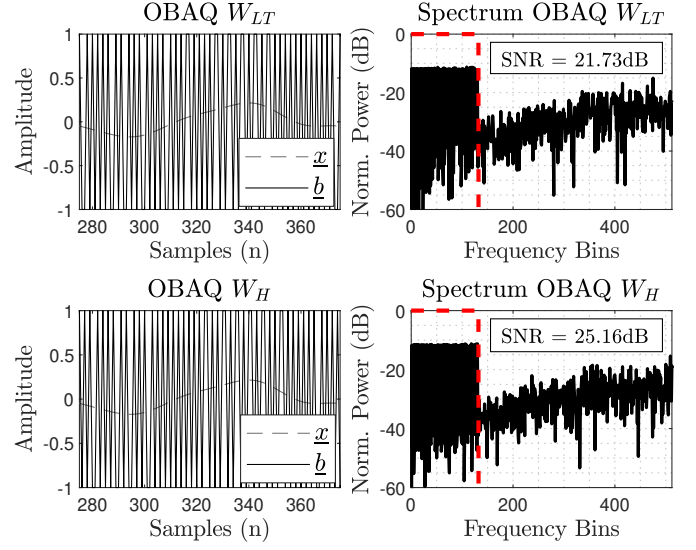


Fig. 6. The outputs of the proposed quantization process using  $W_{\setminus}$  (top) and  $W_H$  (bottom). Comparing both spectra shows that employing  $W_H$  tends to process the noise to the frequency bins greater than  $w_c$ . The red dashed line depicts an ideal reconstruction filter, employed for the SNR calculation.

guess, as the sequentially generated signal  $\underline{b}_{W_H}$  offers an advantageous initial vector for the optimization process, where a robust local minimum can be achieved in a short time that would have taken many times longer using random initialization.

## VI. CONCLUSIONS

We presented a framework, specified as  $\mathcal{F}(\cdot)$ , for mapping real-valued discrete-time signals to their one-bit representations. The framework employs an optimization-based approach that minimizes the error function  $E(\underline{x}, \underline{b})$ , quantifying the difference between the original signal  $\underline{x}$  and its one-bit representation  $\underline{b}$  using the squared  $l^2$ -norm.

To obtain a one-bit signal, discrete optimization toolboxes such as Gurobi, CVX toolbox, for the MATLAB global optimization toolbox can be used, but their suitability for real-time applications may be limited as they need more computational resources to find a good local optimum. In addressing this issue, we present a sequential optimization approach with complexity  $O(N)$ . In general, the presented approach allows to model the characteristic of the quantization process (behavior of the introduced harmonics) just by employing particular structures for filter matrix  $W$ .

Future works could employ the fast sequential quantization result as a initial vector for a general optimization. Furthermore, converting the filter matrix  $W$  to its block recursive form may have the potential to decrease complexity and mitigate the NP-hard problem due to more independent filter coefficients.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the Austrian Science Fund (FWF), grant number DFH 5, and the province of Styria.

## REFERENCES

- [1] N. C. Sevuhtekin, L. R. Varshney, P. K. Hanumolu, and A. C. Singer, "Signal Processing Foundations for Time-Based Signal Representations: Neurobiological Parallels to Engineered Systems Designed for Energy Efficiency or Hardware Simplicity," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 38–50, Nov. 2019.
- [2] L. Stefanazzi, E. Paolini, and A. Oliva, "Click modulation: an off-line implementation," in *2008 51st Midwest Symposium on Circuits and Systems*. Knoxville, TN, USA: IEEE, Aug. 2008, pp. 946–949.
- [3] S. Khobahi and M. Soltanalian, "Model-Based Deep Learning for One-Bit Compressive Sensing," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5292–5307, 2020.
- [4] S. Khobahi, N. Naimipour, M. Soltanalian, and Y. C. Eldar, "Deep Signal Recovery with One-Bit Quantization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2987–2991.
- [5] K. Hausmair, S. Chi, P. Singerl, and C. Vogel, "Aliasing-Free Digital Pulse-Width Modulation for Burst-Mode RF Transmitters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 2, pp. 415–427, Feb. 2013. [Online]. Available: <http://dx.doi.org/10.1109/TCSI.2012.2215776>
- [6] K. Hausmair, P. Singerl, and C. Vogel, "Multiplierless Implementation of an Aliasing-Free Digital Pulsewidth Modulator," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 9, pp. 592–596, Sep. 2013. [Online]. Available: <http://dx.doi.org/10.1109/TCSII.2013.2268431>
- [7] K. Hausmair, Shuli Chi, and C. Vogel, "How to reach 100% coding efficiency in multilevel burst-mode RF transmitters," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. Beijing: IEEE, May 2013, pp. 2255–2258. [Online]. Available: <http://ieeexplore.ieee.org/document/6572326/>
- [8] S. Chi, P. Singerl, and C. Vogel, "Coding efficiency optimization for multilevel PWM based switched-mode RF transmitters," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*. Seoul, Korea (South): IEEE, Aug. 2011, pp. 1–4. [Online]. Available: <http://ieeexplore.ieee.org/document/6026539/>
- [9] Shuli Chi, K. Hausmair, and C. Vogel, "Coding efficiency of bandlimited PWM based burst-mode RF transmitters," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. Beijing: IEEE, May 2013, pp. 2263–2266. [Online]. Available: <http://ieeexplore.ieee.org/document/6572328/>
- [10] Shuli Chi, C. Vogel, and P. Singerl, "The frequency spectrum of polar modulated PWM signals and the image problem," in *2010 17th IEEE International Conference on Electronics, Circuits and Systems*. Athens, Greece: IEEE, Dec. 2010, pp. 679–682. [Online]. Available: <http://ieeexplore.ieee.org/document/5724603/>
- [11] D. Reefman and E. Janssen, "One-bit Audio: An Overview," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 52, Mar. 2004.
- [12] J. D. Reiss, "Understanding Sigma-Delta Modulation: The Solved and Unsolved Issues," *AES: Journal of the Audio Engineering Society*, vol. 56, no. 1, 2008.
- [13] S. Kershaw and M. Sandler, "Sigma-delta modulation for audio DSP," in *IEE Colloquium on Audio DSP - Circuits and Systems (Digest No. 1993/219)*, 1993, pp. 1/1–1/6.
- [14] R. Schreier, S. Pavan, and G. C. Temes, *Understanding Delta-Sigma Data Converters*, 1st ed. Wiley, Apr. 2017.
- [15] H. Enzinger and C. Vogel, "Analytical description of multilevel carrier-based PWM of arbitrary bounded input signals," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. Melbourne VIC: IEEE, Jun. 2014, pp. 1030–1033.
- [16] B. F. Logan, "Click Modulation," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 3, pp. 401–423, Mar. 1984.
- [17] F. Chierchie and E. E. Paolini, "Digital Distortion-Free PWM and Click Modulation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 3, pp. 396–400, Mar. 2018.
- [18] G. Yuan and B. Ghanem, "An Exact Penalty Method for Binary Optimization Based on MPEC Formulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017.
- [19] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
- [20] J. Chae and S. Hong, "Greedy Algorithms for Sparse and Positive Signal Recovery Based on Bit-Wise MAP Detection," *Trans. Sig. Proc.*, vol. 68, pp. 4017–4029, Jan. 2020, publisher: IEEE Press.
- [21] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: <https://www.gurobi.com>
- [22] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," Mar. 2014. [Online]. Available: <http://cvxr.com/cvx>
- [23] M. Srinivas and L. Patnaik, "Genetic algorithms: a survey," *Computer*, vol. 27, no. 6, pp. 17–26, 1994.
- [24] D. Delahaye, S. Chaimatanan, and M. Mongeau, "Simulated Annealing: From Basics to Applications," in *Handbook of Metaheuristics*, M. Gendreau and J.-Y. Potvin, Eds. Cham: Springer International Publishing, 2019, pp. 1–35.
- [25] S. J. Wright, "Coordinate Descent Algorithms," Feb. 2015, arXiv:1502.04759 [math]. [Online]. Available: <http://arxiv.org/abs/1502.04759>
- [26] A. Oppenheim and R. Schaffer, *Discrete-Time Signal Processing: Pearson New International Edition*. Pearson Education, 2013.
- [27] T. M. I. Mathworks, "Optimization Toolbox version: 9.4 (R2022b)," Natick, Massachusetts, United States, 2022. [Online]. Available: <https://www.mathworks.com>