Optimal design of the noise transfer function of ΔΣ modulators: IIR strategies, FIR strategies, FIR strategies with preassigned poles

Sergio Callegari [a,c,*], Federico Bizzarri [b,c]

[a] DEI, University of Bologna, Bologna, Italy
[c] ARCES, University of Bologna, Bologna, Italy
[b] DEIB, Politecnico di Milano, Milan, Italy

Received 19 September 2014 Received in revised form
1 February 2015
Accepted 2 February 2015 Available online 20 February 2015

1. Introduction

ΔΣ modulators are widely adopted in tasks such as A/D, D/A and D/D conversion [1,2], frequency synthesis [3], waveform storage [4], signal processing [5] and more. Recently, it has been suggested that they may even be used as heuristic optimizers for special classes of optimization problems [6,7]. In all these applications, their attractiveness mainly derives from the ability to perform *noise shaping*. Such property is achieved through a nonlinear feedback architecture and is typically described by means of an approximated linear model through the so-called noise transfer function (NTF) [8]. Fig. 1(a) illustrates the structure of a classic modulator including a single integrator on the feedforward path, as well as the substitution applied onto the nonlinear quantizer to derive the approximated linear model. In this basic setup, the noise transfer function is $NTF(z) = X(z)/E(z)$ for $U(z) = 0$ (capital letters systematically indicate the z-transforms of the corresponding uncapitalized quantities) and necessarily takes a first-order high pass (HP) form. This is suitable for modulators where the useful signal occupies the lower part of the frequency range.

In order to improve the performance and flexibility of the modulator, it is desirable to adopt NTFs that are higher order, not necessarily HP, and tuned to the specific application. To this aim, two major strategies exist (and in rare cases may even be combined):

* Corresponding author at: DEI, University of Bologna, Bologna, Italy.
E-mail addresses: sergio.callegari@unibo.it (S. Callegari), federico.bizzarri@polimi.it (F. Bizzarri).

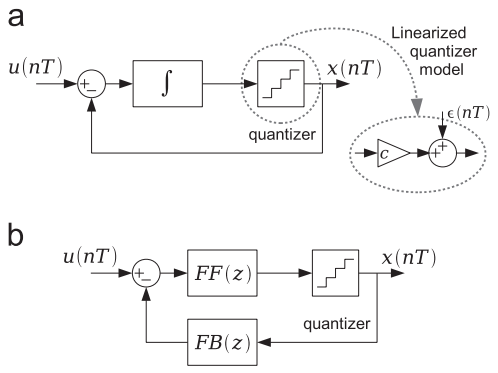

Fig. 1. Block diagram of a classic discrete-time ΔΣ modulator (a) and of its generalization (b). The derivation of the approximated linear model is also illustrated.

(i) move to the setup in Fig. 1(b), where $FF(z)$ and $FB(z)$ are generic filters;
(ii) move to arrangements where multiple ΔΣ loops are cascaded.

In (ii), multiple quantizers are used and the quantization error of each stage is passed to the subsequent one. Then, the outputs of all the quantizers are combined, leading to the so-called multi-stage noise shaping (MASH) [8]. As long as it relies on basic internal loops, the MASH approach is substantially immune from stability issues [9]. However, it may constrain the NTF to specific forms and suffer from mismatch among its stages (at least in analog implementations). Most important, it necessarily delivers a multi-bit output due to the composition of the signals from the individual quantizers. This is ill suited for applications where very low sample depths are required, e.g., the direct drive of power bridges. Even if workarounds exist for this problem, such as the interposition of pulse width modulation stages, they may not always be desirable.

For these reasons, single-quantizer structures such as that in Fig. 1(b) remain essential. This implies an interesting signal processing problem with respect to the design of their inner filters, because the analysis of the approximated linear model is not sufficient to guarantee stability. As a matter of fact, until 1987 it was common belief that single-bit modulators with orders higher than 2 were impossible to reliably stabilize [10]. Today, many issues have been overcome and design flows for these modulators exist [8,11–15]. They are centered around the NTF selection, from which $FF(z)$ and $FB(z)$ can be determined once the signal
transfer function (STF), namely $STF(z) = X(z)/U(z)$ for $E(z) = 0$, is assigned. The key idea is that stability can be achieved by appropriately constraining the NTF. Even if the constraints are often semi-empirical, meaning that stability cannot be formally guaranteed, the concept is known to work in most practical cases [16]. Different types of constraints exist, the most famous ones being: the Agrawal and Shenoi Power Gain Criterion [17]; the Lee Criterion [18]; the Kenney No-overload Criterion [11]; and the Anastassiu Criterion [19]. Among them, the Lee criterion, that bounds the peak gain of the NTF, is probably the most popular one for modulators with 2- or 3-level quantizers [8].

A key point to note is that, once structural and stability constraints are defined, the remaining degrees of freedom are available to optimize the NTF features under some *cost function*. The most classic one is in-band quantization noise power [8], but other indexes exist such as application-perceived noise power [12], worst-case noise power density [14], etc. In some of them, adaptation to the embedding application is inherent [12,13,20], while other indexes completely neglect the matter. This is interesting, because the noise shaping practiced by the modulator should in principle be matched to the noise removal abilities of the embedding environment [12]. Neglecting this aspect is equivalent to assuming that the matching should be delegated to the embedding environment [21]. Yet, this may not always be possible or convenient [12,22].

Attempting a constrained NTF optimization can be complex. The problem of optimally determining the coefficients of a discrete time filter is still a hot topic in many application fields (e.g., see [23–25] for a few representative cases). The complication induced by the different treatment necessitated by the numerator and denominator of rational polynomial functions often requires approximations, heuristic techniques, or specific assumptions on the filter structure [23]. Restriction to finite impulse response (FIR) arrangements can often help achieve exact solutions [13,14,24,25]. Conversely, the presence of constraints defined in the frequency domain *for every possible frequency*, implies *semi-infinite programming* [26] and may make the optimization harder to tackle. In the case at hand, the problem is always non-linear, possibly characterized by local optima, and the management of the constraints is non-trivial, even for empirical ones enjoying simple formulations such as Lee's. Furthermore, one cannot ignore implementation costs. In fact, the NTF determines the modulator internal filters, so that the implementation effort scales with its order. For analog modulators, very high order NTFs can be particularly impractical.

For these reasons, there cannot be a catch-all optimization approach. To keep the optimization manageable, the NTF search is generally restricted to specific filter classes, with different classes being better or worse suited at different combinations of cost-functions and constraints. Two major options exists, where the NTF is respectively sought within
(a) a *particular* infinite impulse response (IIR) form;
(b) a *generic* FIR form.

Only quite recently (in fact while this paper was in review), a proposal has been made for a *free form* IIR design strategy [15]. Yet, this targets just the combination of a specific merit factor with a specific constraint arrangement (more in the next pages). For other cases, practically adopted IIR forms include Chebyshev Type II pole-zero arrangements [10], or forms where the NTF zeros all lie on the unit circle, while the poles take maximally flat arrangements [27]. A rationale for using a specific IIR form is that the NTF parametrization can be reduced. For instance, consider an N-th order IIR NTF. In the most general case it takes $2N+1$ coefficients. Working with such a large number of parameters when ties to merit factors and constraints are non-linear can be hard. Restricting to a Type II Chebyshev form [28], reduces the

parameters to just 3 (in-band gain, stop-band minimal attenuation, and stop-band edge frequency) and then structural requirements from the loop architecture (detailed in the following pages) enable further reduction to just 1, regardless of N. Similarly, restricting to forms where the NTF zeros all lie on the unit circle, while the poles take a max-flat form, restricts the parametrization to $N/2$ frequencies associated to the zeros and to a single coefficient for the poles. Such problem-size reductions can be a determining factor for the feasibility of the optimization. Furthermore, restricted forms can make constraints specified at every possible frequency immediately collapse into constraints for a single frequency (e.g., think of a bound on a peak gain and its implications where the peak frequency cannot or can be known in advance). Conversely, the idea underlying the use of FIR forms is not to reduce the parametrization (actually FIR transfer functions require lots of parameters to achieve detailed features), but to enable the use of efficient optimization codes. With FIR forms, the ties between the NTF parameters, the merit factors and constraints can typically be expressed through standard convex forms. For instance, the Lee constraint can be translated into a linear matrix inequality (LMI) via the Kalman–Yakubovich–Popov (KYP) lemma [29] so that the resulting optimization problem can be solved via semidefinite programming (SDP) and interior point methods [30,31].

It is worth underlining that a fundamental difference exists between restricting to a specific IIR form and a generic FIR one. In the latter case, no limitation is placed on the *features* that the NTF can get, provided that a sufficiently high order can be selected. Indeed, any transfer function can be approximated arbitrarily well by an FIR form, as long as it takes a sufficiently high order. Conversely, restricting to a *particular* IIR form severely limits the NTF features. For instance, Chebyshev forms cannot approximate arbitrary magnitude responses, regardless of their order. This has important implications. In the FIR case, as the order is increased, one certainly tends to the best possible NTF shape and multiple merit factors can be tackled. Conversely, some restricted IIR form may work well for a merit factor and bad for another. Furthermore, the behavior cannot generally tend to the best possible one as the order is increased. Indeed, restricted IIR forms may be better performing than FIR forms at very low orders, but may be paradoxically characterized by performances that *worsen rather than improving* as one lets them take higher orders.

From a designer point of view, *asymptotic* behaviors achievable for extremely large NTF orders may often be irrelevant. Yet, it is certainly relevant to know if there are *critical cases* where conventional restricted IIR forms can exhibit an opposite-to-expected order-performance relationship, since this may misguide into wrong design choices or hinder the reach of the targeted performance levels. To the best of the authors' knowledge, a thorough answer to this question has not yet been given. Then, there is a second question, namely if, by taking less constrained forms, IIR strategies can be made more flexible, in order to avoid these critical situations. This question has not yet been answered either. The purpose of this work is to start attacking the two points. To this aim (i) multiple design strategies are profiled against different sets of merit factors. Cases where the sub-optimal nature of conventional IIR strategies can be critical are identified; (ii) FIR strategies are extended to work with pre-assigned pole structures; (iii) this extension is used to devise more flexible IIR strategies where a fixed pole structure is coupled with a non-restricted search of the zeros. The last point is particularly interesting as it provides a way to work around the limitation of current IIR design flows and particularly to design low order IIR NTFs with better adaptation to the modulator embedding environment. In the discussion, the results are also used to assess the quality of zero placement in conventional design methods for IIR NTFs.

2. Background

When a modulator such as that in Fig. 1 is linearized, it gets described by an STF from input $u(nT)$ to output $x(nT)$ and an NTF from quantization-noise $\epsilon(nT)$ to $x(nT)$, where T is the sampling period. The relationships with the loop filters $\mathrm{FF}(z)$ and $\mathrm{FB}(z)$ are

$$\begin{cases} \mathrm{NTF}(z) = \dfrac{1}{1+c\mathrm{FF}(z)\mathrm{FB}(z)} \\ \mathrm{STF}(z) = \dfrac{c\mathrm{FF}(z)}{1+c\mathrm{FF}(z)\mathrm{FB}(z)} \end{cases}, \quad \begin{cases} \mathrm{FF}(z) = \dfrac{\mathrm{STF}(z)}{c\mathrm{NTF}(z)} \\ \mathrm{FB}(z) = \dfrac{1-\mathrm{NTF}(z)}{\mathrm{STF}(z)} \end{cases} \tag{1}$$

where c is the equivalent quantizer gain, customarily assumed to be 1. As long as the STF is preassigned, as it is typically the case, the choice of the NTF determines the modulator design.

Information on $\epsilon(nT)$ can be derived from the classical model of quantization (CMQ) which states that a uniform quantizer can be approximately modeled by the superposition of white noise independent from the quantized signal and uniformly distributed within $[-\Delta/2, +\Delta/2]$, where Δ is the quantization step. This holds relatively well whenever the modulator input signal is "busy" [8]. Following CMQ, the average power of $\epsilon(nT)$ is $\sigma_\epsilon^2 = \Delta^2/12$, and its power spectral density (PSD) is $\Psi_\epsilon(f) = \Delta^2/12$, where $f \in [-1/2, +1/2]$ is normalized frequency, namely the real frequency over the update frequency $f_\Phi = 1/T$. In this paper, PSDs are defined as *two-sided* over *normalized frequency*, consequently, the quantization noise component at the modulator output has a PSD

$$\Psi_n(f) = \frac{\Delta^2}{12}\left|\mathrm{NTF}\left(\mathrm{e}^{\mathrm{i}2\pi f}\right)\right|^2. \tag{2}$$

2.1. Constraints

The NTF choice is subject to some constraints. First of all, the modulator loop cannot be algebraic. Thus, the loop function $c\mathrm{FF}(z)\mathrm{FB}(z) = (1-\mathrm{NTF}(z))/\mathrm{NTF}(z)$ must include some delay. For this condition to be satisfied, the NTF impulse response needs a unitary zero lag coefficient (or, equivalently, the NTF should be a bi-proper rational transfer function with monic numerator and denominator). Secondly, the modulator must be stable. As mentioned in the Introduction, this requirement cannot be enforced looking at the stability of the linear model, which is approximated. Namely, it is not sufficient to design a stable NTF and many criteria exist to practically ensure the loop stability. Among them, this work focuses on the Lee criterion [18], which is the most commonly

adopted one for very low depth quantizers [8]. Some of the proposed conclusions may hold for other criteria, though.

The Lee criterion limits the peak NTF gain. Formally, the following inequality is enforced

$$\|\mathrm{NTF}\|_\infty = \max_{f \in [0,1/2]} \left|\mathrm{NTF}\left(e^{i2\pi f}\right)\right| < \gamma \tag{3}$$

where γ is a constant depending on the quantizer resolution. Binary quantizers need $\gamma \leq 2$ and $\gamma = 1.5$ is often used. A justification of the criterion comes from the consideration that $\Delta\Sigma$ loops can easily lead to approximated linear models showing *conditional stability*. In other words, once FF(z) and FB(z) are assigned, the linear model can be stable for $c=1$, but unstable for lower c values. Recalling that c is the equivalent quantizer gain, one can expect its value to reduce when the quantizer input gets large (or so to say, when the modulator is *overloaded*). From Fig. 1 and Eq. (1), one sees that, in the ideal case, the quantizer input is $\mathrm{STF}(z)U(z)+(\mathrm{NTF}(z)-1)E(z)$. Namely, the quantizer input can grow large even if $u(nT)$ is appropriately bounded, due to the $(\mathrm{NTF}(z)-1)E(z)$ component. The worst case is obviously encountered if one has a burst where $\epsilon(nT)$ looks like a tone whose frequency corre-sponds to a peak in the magnitude response of NTF(z).
Thus, the precaution of limiting $\|\mathrm{NTF}\|_\infty$ is justified. Even if the Lee criterion is neither a necessary nor a sufficient condition for stability, experiments reported in [18,32] and subsequent experience have shown its effectiveness.

2.2. Merit factors

The quality of an NTF can be evaluated by different criteria. Three notable ones are reported here and used in the following discussion.

2.2.1. In-band quantization noise power

The most obvious way to define a cost function is to look at the in-band quantization noise at the modulator output. This is [8]

$$P_B = \int_{\mathcal{B}} \Psi_n(f)\mathrm{d}f \tag{4}$$

where $\mathcal{B}$ is the set of frequencies in the signal band.[1] For low pass (LP) modulators, this is

$$P_B = \int_{-1/(2OSR)}^{1/(2OSR)} \Psi_n(f)\mathrm{d}f = 2\int_0^{1/(2OSR)} \Psi_n(f)\mathrm{d}f \tag{5}$$

where OSR is the oversampling ratio $f_\Phi/(2B)$ when B indicates the signal bandwidth, namely half the measure of $\mathcal{B}$. Evidently, P_B is improved when the NTF magnitude response is kept low *on average* in the whole signal-band of the modulator.

2.2.2. Worst-case quantization noise power density

Another approach is to look at the worst case noise power density for any possible in-band frequency [14]

$$P_{dM} = \max_{f \in \mathcal{B}} \Psi_n(f). \tag{6}$$

[1] Including negative frequencies, since two-sided PSDs are considered.

This merit factor makes particular sense when the input signal is made of multiple components. Say that there are m sub-bands $\mathcal{B}_1, \ldots, \mathcal{B}_m$ with bandwidths $B_1, \ldots, B_m$ and that the input power P_u spreads uniformly among them (so that $\mathcal{B}_i$ gets power $P_u B_i/B$). Clearly, designing the NTF, one should avoid favoring SNR in a sub-band at the detriment of some other. Thus, the optimization should be aimed at worst case SNR, namely

$$\mathrm{SNR}_{\mathrm{worst}} = \min_{i \in \{1,\ldots,m\}} \frac{\frac{P_u B_i}{B}}{\int_{\mathcal{B}_i} \Psi_n(f)\,\mathrm{d}f}. \tag{7}$$

If m is large, the bands are thin and Eq. (7) is easily approximated by assuming that $\Psi_n(f)$ stays approximately constant within each of them. With this, one gets

$$\mathrm{SNR}_{\mathrm{worst}} \approx \frac{P_u}{2B \max_{f \in \mathcal{B}} \Psi_n(f)} = \frac{P_u}{2BP_{dM}} \tag{8}$$

which is maximized by minimizing P_{dM}. Evidently, P_{dM} is improved when the peak value of the NTF magnitude response within the signal band of the modulator is kept low. Empiric evidence shows that this typically implies having the NTF magnitude response as flat as possible in the signal band of the modulator [14].

2.2.3. Weighted quantization noise power (application-perceived noise power)

Finally, one may want to look at the quantization noise PSD integrated over the whole available bandwidth, after some weighting [13]. Namely

$$P_W = 2\int_0^{1/2} \Psi_n(f)w(f)\mathrm{d}f \tag{9}$$

where $w(f)$ is the weighting function. This merit factor is particularly useful when the modulator is followed by a filter in charge of removing the quantization noise. By setting $w(f)$ to the squared magnitude response of the filter, P_W returns the power of the noise that leaks through it [12,20]. This quantity can be used to compute the SNR that is actually perceived by the application embedding the modulator. Note that the filter following the modulator need not be electronic. For instance, in an audio application, the weighting function can also account for the psycho-acoustic filtering provided by the listener auditory system [33]. Finally, note that the in-band merit factor P_B is a special case of P_W with $w(f)=1$ for $f \in \mathcal{B}$ and null otherwise. Empiric evidence shows that having a good P_W typically involves getting an NTF magnitude response capable of "compensating" $w(f)$, namely capable of making the product $\Psi_n(f)w(f)$ as flat as possible [13].

2.3. Conventional IIR design flows for the NTFs

As representative examples of conventional design flows based on restricted IIR forms, two cases are worth reporting. One uses Chebyshev Type II forms and the other one is Schreier's sophisticated `synthesizeNTF` method [27]. As a justification for the particular choice of these two examples, it is worth underlining that (i) representative constructs with significantly fewer constraints are lacking, as they would imply a non-convex optimizations too complex to be exactly solved for every possible case. The best that is currently

available is a quite recent result [15] that is completely free-form, but that only targets the P_{dM} merit factor (which, according to our evaluations, is a not too critical case for the two sample methods above); (ii) these methods are what most designers are accustomed to and use, since they are present in widespread electronic design automation (EDA) tools. Both methods are illustrated for the LP case and can be adapted to band pass (BP) modulators as well. Variants of these strategies exist, but similar considerations apply, so that the results presented in this work also hold for them.

2.3.1. IIR design flow based on Chebyshev Type II forms

In an LP modulator, the NTF needs to be HP. An HP Chebyshev Type II filter with unitary in-band gain is defined by three parameters: the order N; the stop-band edge frequency f_{st}; the maximum out-of-band gain $R<1$ [34]. Once the order is assigned and f_{st} is set to $1/(2\mathrm{OSR})$, to assure that quantization noise is strongly attenuated in the signal band, one remains with the single parameter R. Now recall that NTF(z) needs to be monic in its numerator and denominator. After the gain is adjusted to satisfy this condition, R becomes the ratio between the peak out-of-band gain and the in-band gain. Rising R reduces the in-band gain and increases the out-of-band gain and viceversa. To satisfy the Lee criterion, it is enough to rise R until the in-band gain is reduced to γ. Rising R more than that should be avoided as it increases the NTF in the signal band for no reason. Obviously, adjusting R in this way minimizes merit factor P_B for this class of filters. The role of R and how P_B gets optimized is well illustrated by the plots in Fig. 2. In this example, as through all those in this work, OSR=64 and $\gamma = 1.5$. Even if OSR can vary significantly across practical applications, the chosen quantity is within the range of adopted values, helps obtaining well readable plots, and is close to the defaults used in some EDA tools [27].

2.3.2. IIR design flow forcing zeros on the unit circle (Schreier's `synthesizeNTF`)

In this design flow, the NTF is expressed as $B(z)/A(z)$, where $A(z)$ and $B(z)$ are monic. The roots of $B(z)$ are constrained to lay on the unit circle, while those of $A(z)$ are chosen so that $A(z)$ is HP and maximally flat at low frequencies. Since $A(z)$ must be monic, this leaves a single degree of freedom for its definition, a quantity that shall be indicated as α, related to the stop-band edge frequency [8, Section 4.3]. Specifically, the pole placement is chosen according to the solution of

$$1+\alpha(z-1)^N\left(z^{-1}-1\right)^N=0. \tag{10}$$

where N is the modulator order. This returns $2N$ values, but only those inside the unit-circle are retained. As α is increased, the stop-band edge frequency is reduced, the low-frequency gain is reduced and the high-frequency gain is increased. With this, the design can be based on three steps, where the last two may be iterated multiple times:

1. The roots of $A(z)$ are all initially set at zero.
2. The roots of $B(z)$ are chosen to minimize P_B. This leads to an HP $B(z)$.
3. The value of α is adjusted until the peak magnitude response of $B(z)/A(z)$ is equal to γ, to respect the Lee constraint. This adjustment is relatively easy to perform, since the particular choice in the forms of $B(z)$ and $A(z)$ always makes $B(z)/A(z)$ peak at $f=1/2$. The design flow may already stop here. Alternatively, for better accuracy, it can repeat from step 2.

Clearly, this procedure leads to a minimization of P_B for this specific class of filters. The visual aspect of the NTF magnitude responses obtained by this procedure is very similar to that obtained by the Chebyshev method, already illustrated in Fig. 2b. However, in many cases the minimal P_B for this class of filters is slightly better than the minimal P_B for the class based on Chebyshev Type II forms. This happens because this procedure is less constrained than the Chebyshev one. In other words, Schreier's `synthesizeNTF` in many cases can do better because, at comparable order, the search space provided by its restricted IIR filter class is somewhat larger than that provided by Chebyshev Type II forms and can consequently get closer to the ideally optimal NTF.

2.4. FIR design flows for the NTF based on the KYP lemma

NTF design methods based on FIR forms are relatively recent [12–14]. In their original statements, each of them considers a different merit factor, yet many common traits exist. A key point is that since there are no restrictions on

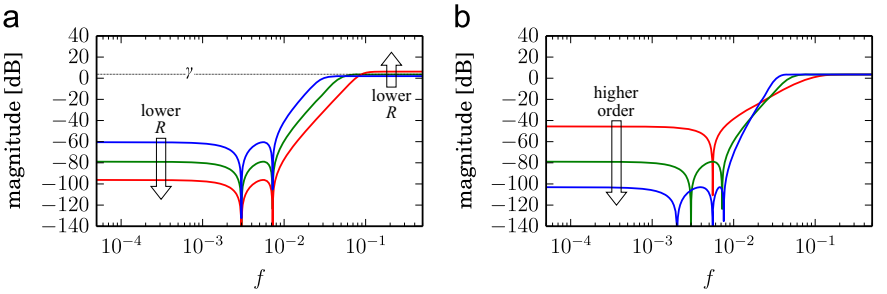

Fig. 2. Effect of changing parameter R and modulator order in an NTF design procedure based on an IIR Chebyshev Type II form. The NTF magnitude response is plotted. In (a), curves corresponding to different R values are shown for a fourth-order modulator: optimal R value (curve starting at approximately −80 dB); 10-fold higher and lower R values (curves starting at approximately −60 dB and −100 dB, respectively). Optimal value makes peak gain exactly equal to γ. In (b), curves corresponding to different orders are shown for the optimal R value: orders 2 (curve starting at approximately −40 dB), 4 (starting at approximately −80 dB) and 6 (starting at approximately −100 dB). All plots obtained with OSR=64 and $\gamma=1.5$.

the features that the NTF can take, it is impossible to known in advance where it will peak. With this, the Lee criterion translates into an infinite set of constraints $|\mathrm{NTF}(e^{i2\pi f})| < \gamma$, one for every possible frequency f. Since such a universal qualification is clearly unmanageable, all the FIR based strategies reformulate the constraint via the KYP lemma into an existential qualification. The details of such procedure can be found in [12,14]. Its main lines are the following.

First of all, a controllable state space form is derived for the NTF, such as

$$\begin{cases} \boldsymbol{\xi}(n+1) = \boldsymbol{A}\boldsymbol{\xi}(n) + \boldsymbol{B}\epsilon(n) \\ \nu(n) = \boldsymbol{C}\boldsymbol{\xi}(n) + \boldsymbol{D}\epsilon(n) \end{cases} \tag{11}$$

where $\boldsymbol{\xi}$ is the state vector, ν is the output, ϵ is the input, and matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are $N \times N$, $N \times 1$, $1 \times N$ and 1×1 respectively. By using a controllable canonical state space form, it can be assured that the to-be-determined FIR coefficients of the NTF all go in matrix $\boldsymbol{C}$, while $\boldsymbol{D}$ contains the first NTF coefficient, that is known to be 1. Matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ are independent of the NTF coefficients, with $\boldsymbol{A}$ being upper diagonal and $\boldsymbol{B}$ being a column vector with all zeros, but for a single 1 as its last entry.

With this, the KYP lemma assures that the relation $\|\mathrm{NTF}\|_\infty < \gamma$ holds if

$$\exists \boldsymbol{P} \in \mathbb{R}^{N\times N} \text{ s.t. } \boldsymbol{P} \geq 0 \quad \text{and} \quad \begin{pmatrix} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B} & \boldsymbol{C}^{\mathrm{T}} \\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{B} - \gamma^2 & \boldsymbol{D} \\ \boldsymbol{C} & \boldsymbol{D} & -1 \end{pmatrix} \leq 0 \tag{12}$$

where the inequalities, applied to matrices, indicate positive or negative semidefiniteness and the superscript T indicates transposition.

Under this premise, the NTF selection problem is transformed into an optimization problem where some merit factor such as P_B, P_{dM}, or P_W in (4), (6) and (9) needs to be minimized, while respecting (12). Interestingly, P_B and P_W can be expressed as positive definite quadratic forms on the unknown FIR coefficients [13] and P_{dM} can be minimized by defining an LMI on them [14]. Furthermore, the condition (12) can also be expressed through an LMI in the unknowns (the entries of $\boldsymbol{P}$ and the FIR coefficients). Hence, altogether one has a convex optimization problem that can be efficiently tackled by SDP and interior point methods [35]. Note that in this approach the requirement that the NTF takes an FIR form is crucial. If the NTF were IIR, it would not be possible to state the merit factors P_B, P_{dM}, or P_W as a convex expression in its coefficients. Similarly, it would not be possible to express (12) as an LMI in the unknowns.

As an example, Fig. 3 shows some NTF magnitude responses obtained by design flows based on FIR forms, the KYP lemma and SDP. These plots illustrate how the method can be seamlessly adapted to work with all merit factors.

3. Pros and cons of conventional IIR design strategies

As it should be clear from the examples in the previous sections, methods based on IIR NTF forms can perform quite well when the merit factor to be optimized is P_B. Some quantitative comparison is provided in Table 1 (ignore for now the rows corresponding to methods not yet introduced). From the data, it is evident that, when it comes to P_B, IIR strategies can often achieve better behavior than the FIR SDP strategy at much lower order. Quantitative data also confirms the slight advantage of the `synthesizeNTF` IIR strategy over the Chebyshev one. Yet, the previous discussion also hints at the limitations of conventional IIR strategies. First of all, they are all inherently single-band. Specifically, their most basic imple-mentation leads to HP NTFs, suitable for LP modulators. They can be extended to band-stop NTFs suitable for BP modulators by frequency transformations, but it is hard to go further than that. This is because the restriction to a specific IIR form excludes the possibility to design arbitrary

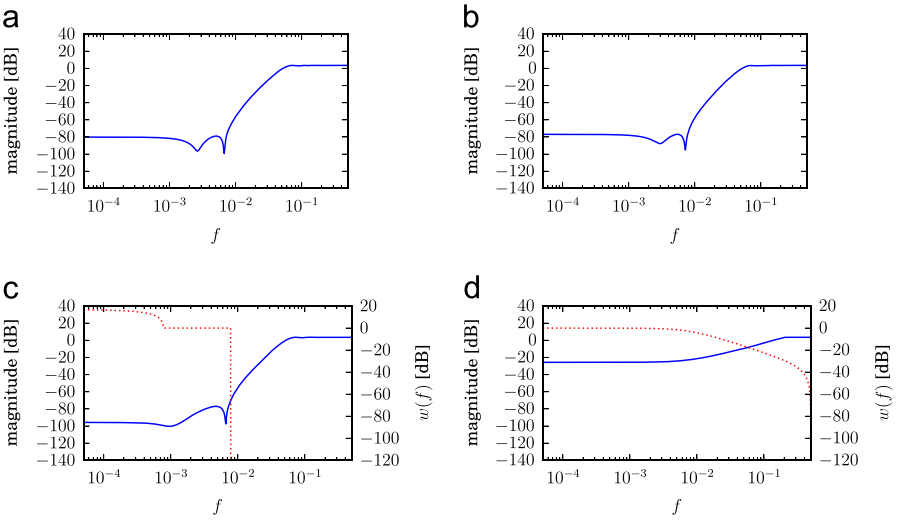

Fig. 3. Sample magnitude responses of NTFs designed by FIR strategies optimizing P_B, P_{dM} and P_W at order 32. In (a), optimization of P_B. In (b), optimization of P_{dM}. In (c), optimization of P_w for a weighting profile (red, dotted) that abruptly becomes negligible out of the signal band. In (d), optimization of P_w for a weighting profile (red, dotted) that does not become abruptly negligible out of the signal band. Plots obtained for OSR=64, and $\gamma = 1.5$.

features, such as multiple stop-bands. By contrast, FIR based strategies can seamlessly deliver NTFs for multi-band modulators, because no restriction is imposed *a priori* on the NTF features. The limitation in the number of signal bands that can be treated may not appear very important as most practical applications of $\Delta\Sigma$ modulator are single-band. Yet, it may be the case that the lack of good multi-band design tools is among the reasons why multi-band applications are rare.

A second limitation, which is more subtle but possibly more important from an application point of view, is that conventional IIR design flows may have difficulties in properly dealing with merit factors other than P_B. This is detailed in the following sections.

3.1. Limitations of the IIR Chebyshev Type II design flow

As mentioned in Section 2.3.1, design flows based on Chebyshev IIR forms are characterized by a single degree of freedom R which is strictly related to the gain of the NTF in its in-band and out-of band frequency regions. Since the out-of-band region of the NTF needs to coincide with the $\Delta\Sigma$ modulator signal band, R has an immediate influence on P_B. As it has already been illustrated, to optimize P_B it is sufficient to select the minimum possible R value that keeps the NTF gain in its in-band frequency region not larger than γ.

One can now consider what happens trying to practice the optimization with respect to other merit factors, so as to fully profile the method and to highlight if critical cases exist. From what has been said so far, one can expect the method to work decently as long as one considers merit factors that appreciate the NTF attenuation in a well-defined LP signal band only. In this case, one can make the NTF stop-band coincident with such band, and then rely on R to basically *scale* the NTF magnitude response there. Specifically, the peaks of the NTF magnitude response in the stop-band are immediately affected by R (by the very role of this control parameter), so one can expect the method to work well with P_{dM} in addition to P_B. For P_W the situation is more complex though. First, the parametrization of Chebyshev NTFs is from the very beginning insufficient to provide any adaptation to a specific weighting profile. There is no "knob" to turn to trade the NTF gain at some frequency neighborhood for the gain at some other frequency neighborhood. Second and most important, the weighting function on which P_W is based may or may not succeed in clearly defining a signal band. When P_W is adopted, depending on $w(f)$ two major cases may occur. In the first scenario, a transition frequency $\tilde{f} \ll 0.5$ exists such that $w(f)$ is clearly non-negligible before it and clearly negligible after it. In the second scenario $w(f)$ stays non negligible through the whole of the $[0, 1/2]$ frequency range or over a very large part of it. In the first case, one can expect Chebyshev type NTFs to be able to somehow cope with P_W, although not optimally. Conversely, in the second case, one can expect the features of the NTF and $w(f)$ to be so badly matched that performance is compromised. In order to verify if these intuitive considerations correspond to actual behaviors, experiments can be run.

For what concerns P_{dM}, from a comparison of Figs. 2(b) and 3(b) one can immediately see that Chebyshev NTFs do not have an optimal in-band shape. In fact, they do not provide a leveled in-band magnitude response as the optimal FIR design strategy. Actually, this would simply be impossible, as Chebyshev type II form are by definition characterized by ripple in their stop-band. In this respect, it is worth recalling that the requirement that NTFs are monic both in the numerator and the denominator means that it is impossible to lower their gain at some frequency interval without raising it somewhere else. This is why the leveling of the NTF is inherent in the P_{dM} optimization: to lower the NTF peaks in the signal band, one ends up raising the valleys. Still, if one looks at quantitative data, it is evident that this *defect* of Chebyshev forms with respect to the P_{dM} cost function does not affect performance significantly. As an example, see Table 2, where the Chebyshev method can provide rather good P_{dM} values at relatively low orders. The reason why this is possible is quite obvious. As the order is increased, the NTF roll-off is improved. Thus, even without a leveled behavior, the NTF peaks in the signal band can be reduced by rising the gain in the initial part of the noise-band (see Fig. 2(b)). To summarize, as expected from the informal considerations at the beginning of this section, the optimization of P_{dM} does not represent a critical case for methods based on IIR Chebyshev forms.

A very similar situation occurs for P_W in conjunction with weightings that abruptly fall to zero above a certain

Table 1
Comparison of different NTF design techniques with respect to the P_B merit factor, for OSR=64, $\gamma = 1.5$.

Design method	Order	P_B (dBm)
FIR, optimized for P_B	32	−79.9
FIR, optimized for P_B	40	−91.5
IIR, Chebyshev Type II	2	−47.8
IIR, Chebyshev Type II	4	−80.9
IIR, Chebyshev Type II	6	−105.0
IIR, Schreier	2	−48.4
IIR, Schreier	4	−81.7
IIR, Schreier	6	−105.4
IIR, Hybrid	2	−48.4
IIR, Hybrid	4	−81.7
IIR, Hybrid	6	−105.5

Table 2
Comparison of different NTF design techniques with respect to the P_{dM} merit factor, for OSR=64, $\gamma = 1.5$.

Design method	Order	P_{dM} (dBm)
FIR, optimized for P_{dM}	32	−54.7
FIR, optimized for P_{dM}	40	−66.3
IIR, Chebyshev Type II	2	−23.4
IIR, Chebyshev Type II	4	−56.7
IIR, Chebyshev Type II	6	−80.8
IIR, Schreier	2	−20.4
IIR, Schreier	4	−51.4
IIR, Schreier	6	−73.9
IIR, Schreier modified for P_{dM}	2	−22.8
IIR, Schreier modified for P_{dM}	4	−56.3
IIR, Schreier modified for P_{dM}	6	−79.6

small $\tilde{f}$. In this case, $\tilde{f}$ can be taken to mark the top of the signal band. As an example, weightings of this sort are inherent in all audio applications, where $w(f)$ frequently includes the psycho-acoustic response of the human auditory system, whose sensitivity drops abruptly above approximately 16 kHz [13,22,33]. In this case, the ideal NTF magnitude response should follow (actually "compensate") the weighting, namely try to make $\Psi_n(f)w(f)$ as flat as possible in the signal band. For instance, Fig. 3(b) shows an FIR NTF fully optimized for a sample arbitrary weighting profile. Again a *flattening* is inherent in the optimization process that Chebyshev forms cannot adapt to. As an example, Table 3 reports data relative to the weighting profile in Fig. 3(c) (column "case I").

Also in this case, design flows based on Chebyshev IIR forms are quite competitive against optimal FIR methods, in the sense that they can easily reach the same performance levels (and typically do so at a low order). The reasons why this happens are substantially the same that have just been illustrated for P_{dM}. To summarize, as expected from the informal considerations at the beginning of this section, the optimization of P_W does not represent a critical case for methods based on IIR Chebyshev forms *as long as one has weightings that fall abruptly to zero at a transition frequency much lower than* $1/2$.

Eventually, one may consider what happens for P_W and weightings that never or only smoothly fade to zero as f is increased. This case is quite common when $w(f)$ is based on the actual profile of a filter in charge of removing quantization noise that is present in the modulator embedding environment, since filters designed to be inexpensive often provide only a moderate roll-off. Also in this case, a fully optimal design should try to "compensate" $w(f)$ by flattening $\Psi_n(f)w(f)$. Yet, now this action should extend to the whole of the $[0, 1/2]$ frequency range. This is something that Chebyshev forms simply cannot accomplish. As an example, consider Fig. 4(a) which proposes the ideal NTF profile for a weighting based on a first-order Butterworth filter. Actually, this figure merely reproduces Fig. 3(d) highlighting the "compensation" effect. The corresponding Chebyshev profile is in Fig. 4(b).

Evidently, here there is no chance of working around the inherent sub-optimality of the restricted NTF shape by rising its order. Conversely, an order rise may deteriorate the match between the NTF and the weighting profiles even more. Quantitative data confirms this impression, as illustrated in Table 3 (case II). Clearly, design flows based on IIR Chebyshev forms cannot reach the same performance levels provided by the optimal FIR method. Furthermore, the situation is subtle, because, contrarily to usual expectations, rising the modulator order worsens the performance rather than improving it. This phenomenon can easily mislead a designer into wrong conclusions unless he/she is well aware of it. Obviously, the reasons lie in the fact that the higher the order, the worse the Chebyshev Type II transfer function can follow a smooth weighting in its transition range. What happens is better evidenced in Fig. 5 that shows the $|\mathrm{NTF}(e^{2i\pi f})|^2 w(f)$ product. The plotted quantity is proportional to the integrand $\Psi(f)w(f)$ appearing in the definition (9) of P_W and as such should be as small as possible. However, as evident from the plots, for the IIR design strategy it is always larger than for the optimal FIR one, precisely at frequencies slightly above the top of the signal band.

Following the latter considerations, one might be tempted to conclude that the case is not actually critical, since one has just to pick a low modulator order, rather than a large one, to get acceptable performance. Unfortunately, this conclusion would be wrong and the approach

Table 3
Comparison of different NTF design techniques with respect to the P_W merit factor, for OSR=64, $\gamma = 1.5$ and two different weighting functions: case I uses the *abrupt falling* weighting in Fig. 3(c), while case II uses the *smoothly fading* weighting in Fig. 4.

Design method	Order	P_W (dBm) (case I)	P_W (dBm) (case II)
FIR, optimized for P_W	32	−78.2	−9.16
FIR, optimized for P_W	40	−90.0	−9.17
IIR, Chebyshev Type II	2	−39.8	−7.48
IIR, Chebyshev Type II	4	−73.2	−4.42
IIR, Chebyshev Type II	6	−94.5	−2.55
IIR, Schreier	2	−42.4	−7.52
IIR, Schreier	4	−75.7	−4.48
IIR, Schreier	6	−95.5	−2.70
IIR, Schreier modified for P_W	2	−44.9	−7.58
IIR, Schreier modified for P_W	4	−80.0	−4.61
IIR, Schreier modified for P_W	6	−96.8	−3.01
IIR, Hybrid	2	−45.7	−8.08
IIR, Hybrid	4	−80.0	−8.00
IIR, Hybrid	6	−96.8	−7.99

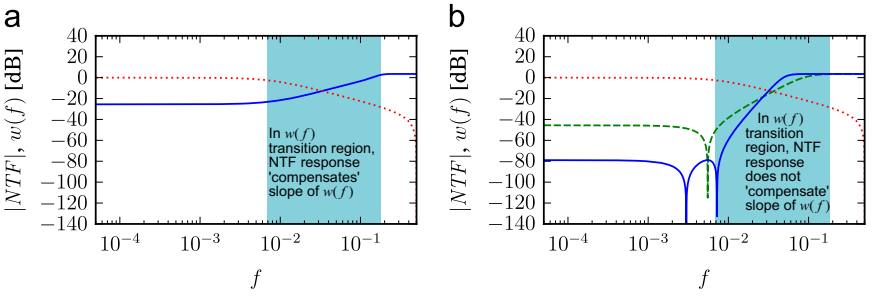

Fig. 4. Comparison of IIR design technique (based on Chebyshev Type II form) and optimal FIR technique with respect to merit factor P_W. In (a), optimal NTF magnitude response (blue, solid line) obtained with an FIR strategy (order 12). In (b), magnitude responses for second- and fourth-order IIR NTFs (dashed green and solid blue line, respectively). Both plots also show the weighting function (red, dotted line). Plots obtained for OSR=64, and $\gamma = 1.5$.

would represent an *ad hoc* solution working just for very plain weightings. For instance, a weighting including a flat zone in the transition range would be impossible to trace for a Chebyshev filter at any order. To summarize, as expectable from the initial parts of this section, a very critical case for design flows based on IIR forms can be identified for the use of the P_W merit factor in conjunction to non-fading or smoothly fading weightings.

3.2. Limitations of the IIR zeros-on-unit-circle design flow

Very similar considerations to those made for the IIR technique based on Chebyshev Type II forms can be made for Schreier's IIR technique. Yet, here there are some additional degrees of freedom thanks to the possibility of arranging the zeros at will, as long as they lie *on the unit circle, within the signal band*. Such freedom can be used to attempt an active optimization of P_{dM} or P_W. This is quite easy to practice, since it is sufficient to substitute P_{dM} or P_W for P_B at step 2 in the algorithm described in Section 2.3.2. For P_{dM} the change is minimal and the results remain substantially aligned to those of the method based on Chebyshev Type II forms. This is well illustrated by the plot in Fig. 6(a) and by the data in Table 2.

For P_W, the results can be far more interesting, as long as the weighting is defined only for the signal band and taken to be negligible elsewhere. As an example, see Fig. 6(b) that can be compared to Fig. 3(c), as well as the data in Table 3 (case I). For the specific application example, the advantage of adapting the zero placement to the specific weighting curve used for the computation of P_W can reach 2.5 dB. As a matter of fact, a similar approach is exploited in [22] to design a psycho-acoustically optimal modulator for audio applications.

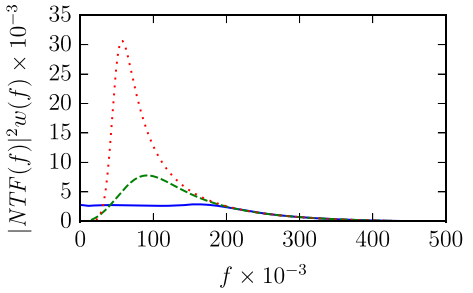

Fig. 5. Comparison of IIR design technique (based on Chebyshev Type II form) and optimal FIR technique with respect to merit factor P_W. The product $|\mathrm{NTF}(e^{2i\pi f})|^2 w(f)$ is illustrated for a FIR strategy with order 12 (blue solid line), as well as for a fourth-order Chebyshev NTFs (dotted, red) and a second-order Chebyshev NTFs (dashed, green). Plots are derived from those in Fig. 4(a) and (b).

Unfortunately, the results become once again quite disappointing when a weighting that does not fall sharply to zero outside the signal band is considered. As an example, Fig. 7 shows an attempt at designing a modulator using a weighting function derived from a first-order LP filter (like that used for the examples in Fig. 4). In Fig. 7(a), the magnitude of the achieved NTF is shown, both for the standard Schreier's method and for the method modified for P_W. In this second case, the method tries to push the zeros at slightly higher frequencies, in order to compensate the weighting profile a bit better at frequencies just above the signal band. Unfortunately, the results are not satisfactory, as evident from Fig. 7(b) that provides plots of the $|\mathrm{NTF}(e^{2i\pi f})|^2 w(f)$ product. The benefit of the method modified for P_W over the standard one is minimal and the performance remains very far away from that achievable by the FIR approach shown in Fig. 5. This is also quantitatively shown in Table 3 (case II). As for the IIR approach based on Chebyshev Type II forms, performance decreases when rising the modulator order.

3.3. Summary of critical cases for conventional IIR design methods

From the previous discussion, it is evident that conventional IIR design methods may show issues for merit factors other than P_B. In fact, there is no guarantee that these methods are *asymtpotically* optimal, namely that they converge to the best possible NTF as the modulator order is grown (in fact, for some cost functions there are strong indications that the contrary happens). Practically, these design methods remain sufficiently successful for the P_{dM} merit factor. Similarly, the behavior remains good for P_W as long as the weighting function rapidly becomes negligible out of the signal band. In fact, in all these cases the merit factors are only concerned with what happens in the signal band. Specifically they look at the *height* of the

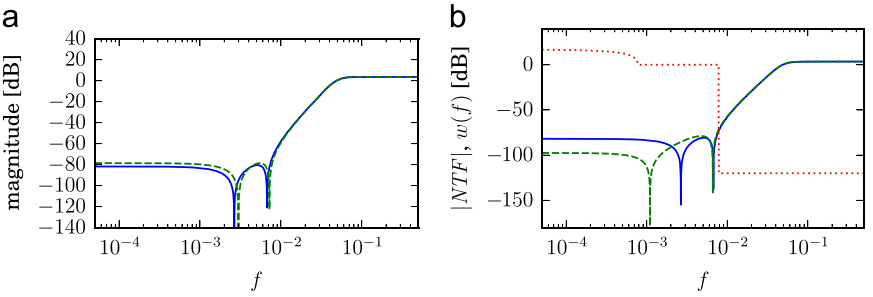

Fig. 6. Comparison of standard Schreier's procedure to the procedure modified for the optimization of P_{dM} and P_W. In (a), comparison of standard procedure (blue solid line) and procedure modified for optimizing P_{dM} (green dashed line), for an order 4 modulator. The two plots are almost coincident. In (b), comparison of standard procedure (blue solid line) and procedure modified for optimizing P_W (green dashed line), for an order 4 modulator. The adopted weighting (red, dotted line) is negligible out of the signal band. In all plots, $OSR=64$, $\gamma = 1.5$.

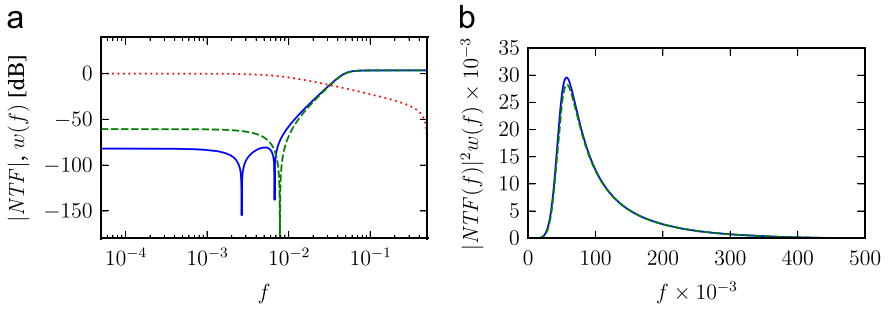

Fig. 7. Behavior of Schreier's IIR NTF design technique when tackling merit factor P_W with a weighting profile that does not fall sharply to zero out of the signal band. In (a), the NTF magnitude response, both for the standard Schreier's method (blue solid line) and for the method modified to explicitly deal with P_W instead of P_B (green dashed line). The weighting function is also given as a reference (red, dotted line). In (b), the $|\mathrm{NTF}(e^{2i\pi f})|^2 w(f)$ product both for the standard (blue solid line) and the modified (green dashed line) method.

magnitude response of the NTF here, observed via some indicator (maximum peak, weighted average of square). All these indicators improve if the NTF scales down as a whole in the signal band. Since IIR design methods let one "shift" down the whole of the NTF magnitude response in the signal band by merely rising the NTF order, they make it possible to improve these indicators even if they do not return the best possible NTF *shape* in this zone.

Yet, critical situations may emerge when one considers merit factors that also look at what happens in the neighborhood of the signal band. This is the case for the merit factor P_W in conjunction with weighting functions that do not fall abruptly to zero above some (low) threshold frequency. In this case, conventional IIR design methods fail because they simply do not offer sufficient degrees of freedom to provide adequate NTF features in the transition region between the modulator signal and noise bands. This means that performance levels provided by FIR strategies, that are *asymptotically optimal*, namely certain to reach the best possible NTF shape as the order is sufficiently increased, are simply unreachable. Paradoxically, the cases where IIR strategies based on restricted forms are ill suited tend to coincide with the scenarios where optimal FIR methods are at their best ease, not requiring very large orders. Critical situations can be immediately spotted from merit factors that decrease while rising the NTF order.

4. Extension of FIR strategies to IIR forms with pre-assigned poles

After critical cases for conventional IIR design techniques have been identified, one may address the question whether such criticality can be avoided by introducing further degrees of freedom in the design flow. Clearly, the ideal case would be to operate with *free form* IIR expressions, namely to have the possibility to use all poles and zeros as degrees of freedom. As hinted in the Introduction, research is active in the area, but the problem is not yet solved. The best currently available is a quite recent algorithm operating with *free form* IIR expressions, specifically targeting the P_{dM} merit factor [15]. This operates by identifying a sequence of convex optimization problems, eventually delivering the best NTF shape. For other merit factors (including P_W), or to rely on techniques with a lower computational cost, a different trade off between flexibility and complexity must be identified. To this aim recall that both the IIR method based on Chebyshev Type II forms and Schreier's `synthesizeNTF` method compel the NTF zeros to fall on the unit circle. In this regard, Schreier's method is already more flexible than the method based on Chebyshev Type II forms, since it allows the zeros to fall anywhere on the unit circle, rather than forcing them to fixed relative positions. This suggests that a good compromise in flexibility can be achieved by making the zero positions selectable over the whole unit disk.

Interestingly, such a result can be easily obtained by extending the fully optimal FIR techniques to take a set of pre-assigned poles. By doing so, one ultimately obtains an hybrid IIR design technique where the poles are as in standard IIR methods, while the zeros are fully optimized. Assume that

$$\mathrm{NTF}(z) = \frac{B(z)}{A(z)} = \frac{\sum_{i=0}^{N_B} b_i z^{-i}}{\sum_{i=0}^{N_A} a_i z^{-i}} \tag{13}$$

where the order is $N = \max\{N_A, N_B\}$, and $a_0 = b_0 = 1$ since the polynomial must be monic. Furthermore assume that $a_1, \ldots, a_{N_A}$ are pre-assigned coefficients, and $b_1, \ldots, b_{N_B}$ are N_B parameters to be found. Let $a_i = 0$ for $i > N_A$ and $b_i = 0$ for $i > N_B$ in order to pad the coefficient vectors to N elements. With this, a controllable state space model equivalent to (13) can be obtained with a structure such as that in Eq. (11) with

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_N & -a_{N-1} & -a_{N-2} & \cdots & a_1 \end{pmatrix}$$
$$\boldsymbol{B} = (0\ 0\ 0\ \cdots\ 0\ 1)^{\mathrm{T}}$$
$$\boldsymbol{C} = (d_N\ d_{N-1}\ d_{N-2}\ \cdots\ d_{N_1}\ 1)$$
$$\boldsymbol{D} = (1) \tag{14}$$

where $d_i = b_i - a_i$ [34]. When (14) is substituted in (12), one gets an expression that is affine in all the unknowns (the b_i coefficients and the entries of $\boldsymbol{P}$), so that it can be reformulated into an LMI that is manageable by practical optimization tools.

Merit factors can be expressed in convex terms too. Here, the discussion is restricted to P_W for which critical situations exist. Yet, P_B is implicitly considered as well, being a special case of the P_W (though in the optimization of P_B critical

situations do not occur). Furthermore, the discussion can also be extended to P_{dM}, but this development is omitted here for brevity and the interested reader is referred to [14]. One has

$$P_W = \frac{\Delta^2}{12}\int_{-1/2}^{1/2}\left|\frac{B(e^{i2\pi f})}{A(e^{i2\pi f})}\right|^2 w(f)\mathrm{d}f. \tag{15}$$

Since $A(z)$ is pre-assigned, this can be rewritten by defining a new weighting function

$$\hat{w}(f) = \frac{w(f)}{|A(e^{i2\pi f})|^2} \tag{16}$$

so that

$$P_W = \frac{\Delta^2}{12}\int_{-1/2}^{1/2}\left|B\left(e^{i2\pi f}\right)\right|^2 \hat{w}(f)\mathrm{d}f \tag{17}$$

where $B(z)$ is FIR, so that the theory in the Literature can be reused. Specifically, following [13] one can take

$$q_{ij} = \int_0^{1/2} \hat{w}(f)e^{i2\pi f(i-j)}\,\mathrm{d}f \tag{18}$$

for $i,j = 0,\ldots,N_B$ as the entries of an $(N_B+1)\times(N_B+1)$ matrix $\boldsymbol{Q}$ and see that P_W is proportional to $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{b}$ where $\boldsymbol{b} = (1, b_1, \ldots, b_{N_B})^{\mathrm{T}}$. Since it can be easily shown that $\boldsymbol{Q}$ must be positive definite, $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{b}$ is necessarily a convex quadratic form that can be used as an optimization goal.

5. Applications

5.1. Hybrid design strategy

The approach that has just been presented as an extension of [12,13] can have a significant application value. In fact, it provides a hybrid design strategy where IIR NTFs can take pole arrangements as in conventional methods and zero arrangements that are fully optimized for any merit factor *given the pre-assigned pole arrangement*. This can get the best of both realms. Namely, it can deliver low order NTFs such as conventional IIR strategies and at the same time provide good performance (almost as good as fully optimal FIR strategies that would be expensive due to high order) for merit factors hard to tackle by conventional IIR strategies. To validate this expectation, test cases similar to those examined previously can be considered. Tables 1 and 3 also indicate the performance for a hybrid IIR design strategy based on the `synthesizeNTF` pole arrangement.[2] As it could be expected, the hybrid strategy has no or negligible advantage in all those cases where conventional IIR methods already work well. However, there is a significant advantage in the critical case where merit factor P_W is considered together with a weighting function that does not go rapidly to zero out of the signal band. For the example under consideration, the hybrid strategy consistently returns results that are very close to the FIR one (just 1 dB worse) and prevents performance from strongly degrading when the NTF order is risen.

[2] Even if it is possible, a hybrid strategy has not been coded for the P_{dM} case, to avoid a cumbersome programming exercise for a case where no benefit was anyway expected.

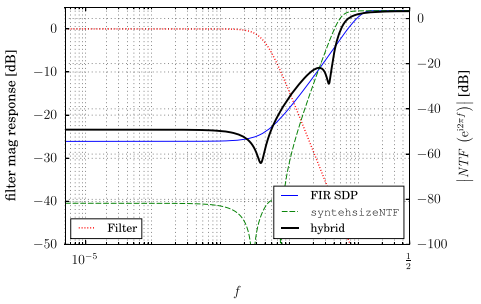

Fig. 8. Exemplification of the hybrid NTF design strategy, taking poles by a conventional method and using SDP to pick optimal zeros in order to minimize the residual quantization noise after filtering.

Since a single example is insufficient to say that a procedure is successful (it could be a lucky case), another test condition is here provided, now thoroughly examining what happens. For this additional example, the values OSR=64 and $\gamma = 1.5$ are kept, but the weighting function is now derived from the hypothesis of having a second-order Butterworth reconstruction filter, with cut-off frequency at the edge of the modulator signal band. The corresponding $w(f)$ is shown as a dotted red line in Fig. 8.

The same figures also show the optimal NTF profile obtained by an FIR strategy with SDP optimization and order set to 16 (above that value performance does not change, so this is approximately the asymptotic optimum to be taken as a reference) as a solid blue line. Furthermore, the figure shows the NTF magnitude response obtained with the `synthesizeNTF` method and the proposed hybrid method, when the order is set to 4 (green dashed line and black thick line, respectively). From the figure, it is straightforward to see that the hybrid methods can deliver an NTF whose magnitude response traces much better the ideal profile, particularly with reference to the roll-off region of the reconstruction filter just above the modulator signal band.

The advantage is confirmed by the P_W data. For the ideal NTF profile, one gets −51.5 dBm. This is a hard bound for the setup, being the plateau at which the optimal FIR method converges at large orders. Schreier's `synthesizeNTF` gives −46.6 dBm, namely a large 5 dB worse than the best possible performance. The proposed hybrid method gives −49.4 dBm, only 2 dB worse than the best possible performance and 3 dB better than the conventional IIR strategy.

5.2. Validation of design choices in conventional methods

Another application of the approach in Section 4 is to *validate* the choices made in conventional IIR strategies in those cases where they are anyway known to work relatively well. Specifically, both Chebyshev and Schreier's design strategies fix the zeros on the unit circle. This is known to give good results when designing for P_B. Yet, backed as it is by many intuitive motivations, this choice is arbitrary. Would placing the zeros elsewhere be better?

A preliminary answer to this question comes from the data in Table 1. Since the performance for Schreier's method and the hybrid one is the same, in this case Schreier's zero placement must be optimal.

Yet, this may be a lucky case and it is sensible to observe the matter in better detail. To this aim, it is worth focusing on the max-flat pole arrangement used in Schreier's `synthesizeNTF`. Two different types of evaluations are reported: (i) tests with different oversampling ratios and different NTF orders using the same pole assignment as `synthesizeNTF`; (ii) tests with maximally flat pole arrangements obtained by varying the α parameter in Eq. (10). The second type of test is useful to also determine if the *pole* assignment done by `synthesizeNTF` under the constraint of a maximally flat pole form is optimal. Results are consistent in showing that

(a) Using a max-flat pole arrangement *and* the α setting delivered by `synthesizeNTF`, the optimal zero placement is on the unit circle and identical to that delivered by `synthesizeNTF`.

(b) α values just slightly smaller than those provided by `synthesizeNTF` (e.g., 1–3%), together with a zero arrangement very close to the unit circle, but not quite on it, may succeed in delivering a marginally better attenuation of the in-band quantization noise, but the advantage is negligible.

The conclusion is that also the α value delivered by `synthesizeNTF` is substantially optimal. In other words, *under the sole assumption that the poles are in a maximally flat arrangement*, `synthesizeNTF` delivers the best pole and zero structure when optimizing for P_B.

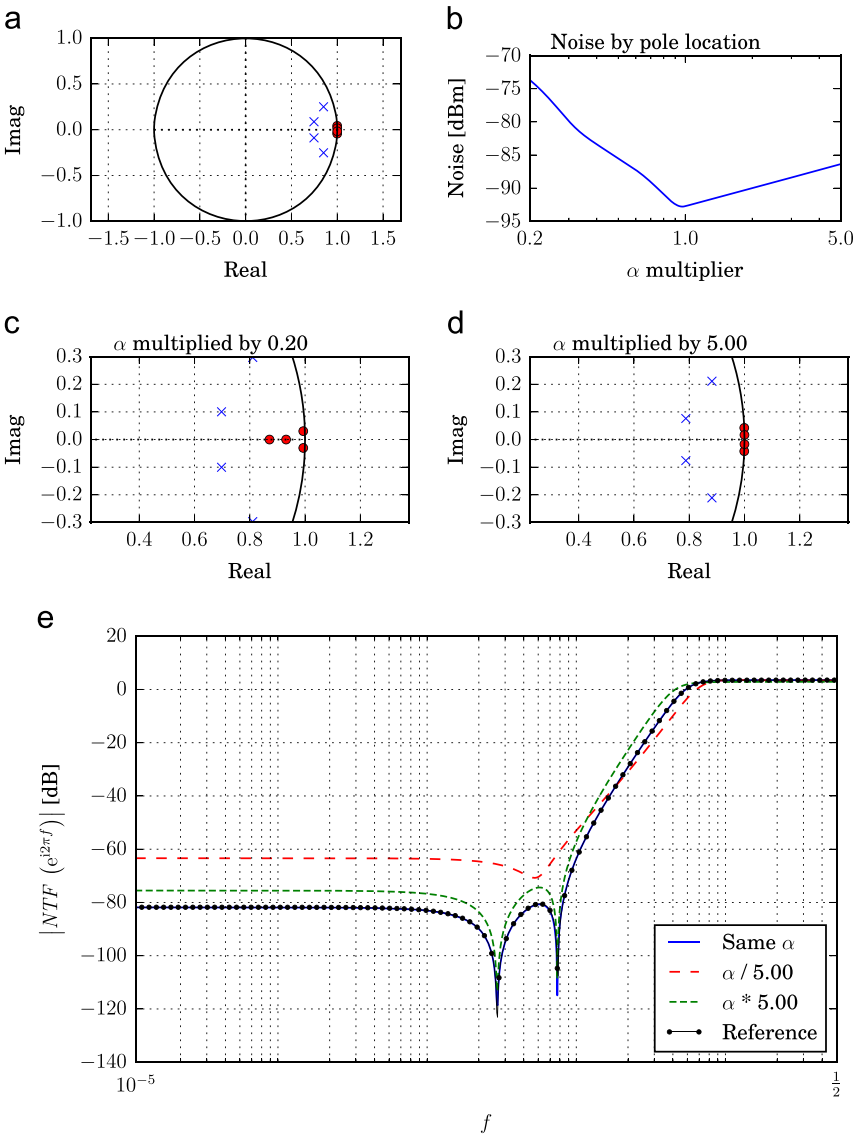

Fig. 9. Validation of design choices in Schreier's `synthesizeNTF` for a sample fourth-order LP modulator with *OSR*=64. Plot (a): pole-zero placement delivered by `synthesizeNTF`. Plot (b): P_B variation as one operates with an NTF with fully optimized zeros and a pre-assigned pole placement given by Eq. (10) when α is varied around the value delivered by `synthesizeNTF`. Plots (c) and (d): pole zero placements corresponding to the cases in plot (b) for the smallest and largest α values under test. Plot (e): NTF magnitude response corresponding to the cases in plot (b) for the smallest and largest α values under test.

For the sake of illustration, at least one of the examined cases is worth a full reporting. Fig. 9 details the findings for a fourth-order binary modulator for LP signals, with OSR$=64$ and $\gamma = 1.5$. Here, `synthesizeNTF` determines $\alpha = 11,662.34$ as the best parameter value. The corresponding pole-zero placement is shown in Fig. 9(a). The corresponding P_B value is -92.71 dBm. Keeping the pole placement from this α value, while determining the zeros by SDP, delivers the very same zeros as `synthesizeNTF`. However, using poles obtained with a perturbed α causes SDP to deliver a zero arrangement with some zeros pushed inside the unit circle, particularly when α is lowered. This is shown in Fig. 9(c) and (d) for the reference α divided and multiplied by 5. The pole-zero arrangements for the perturbed α values are typically worse than those delivered by `synthesizeNTF`, as shown in Fig. 9(b) and (e). However, the best possible setting is not at the α value returned by `synthesizeNTF`, but slightly lower (11,292.92), with P_B at -92.78 dBm. The change is so small that it can be considered negligible, though.

6. Conclusions

In this paper, some of the constraints and merit factors that can be used in the design of $\Delta\Sigma$ modulators have been reviewed, together with some common design strategies. With this, it has been shown that only the most modern strategies based on SDP can correctly deal with the whole range of merit factors herein presented. However, these strategies are generally more expensive than conventional ones in implementation terms, because in most cases they can only deliver FIR NTFs that typically require a high-order to provide the desired features. Then, it has been shown that these modern strategies can be adapted to work with pre-assigned pole arrangements. This result has two valuable practical applications. First of all, it can be exploited to devise *hybrid* design methods that take conventional pole arrangements and fully optimize the zeros. Such arrangements may represent an interesting compromise between different design strategies, since they can deliver low-order IIR NTFs and at the same time provide the flexibility of SDP methods in picking different optimization goals. We expect this to be a useful *ad interim* solution, until other approaches based on IIR forms and imposing even lower constraints can be developed. Secondly, the result can be useful to validate some choices made in conventional design strategies. In the paper, it has been used to validate the substantial optimality of Schreier's `synthesizeNTF`. Code is provided to replicate all the provided results via the PyDSM toolbox (http://pydsm.googlecode.com).

References

[1] P.M. Aziz, H.V. Sorensen, J. Van der Spiegel, An overview of sigma–delta converters, IEEE Signal Process. Mag. 13 (1) (1996) 61–84, http: //dx.doi.org/10.1109/79.482138.

[2] F. Harris, Sigma–delta converters in communication systems, in: J.G. Proakis (Ed.), Wiley Encyclopedia of Telecommunications, vol. IV, John Wiley & Sons, Inc., 2003, pp. 2227–2247, http://dx.doi.org/10.1002/0471219282.eot177.

[3] P.-E. Su, S. Pamarti, Fractional-*N* phase-locked-loop-based frequency synthesis: a tutorial, IEEE Trans. Circuits Syst. II 56 (12) (2009) 881–885, http://dx.doi.org/10.1109/TCSII.2009.2035258.

[4] E. Janssen, D. Reefman, Super-audio CD: an introduction, IEEE Signal Process. Mag. 20 (4) (2003) 83–90, http://dx.doi.org/10.1109/MSP.2003.1226728.

[5] M. Marvin Freedman, D.G. Zrilić, Nonlinear arithmetic operations on the delta sigma pulse stream, Signal Process. 21 (1) (1990) 25–35, http://dx.doi.org/10.1016/0165-1684(90)90024-S.

[6] S. Callegari, F. Bizzarri, R. Rovatti, G. Setti, On the approximate solution of a class of large discrete quadratic programming problems by $\Delta\Sigma$ modulation: the case of circulant quadratic forms, IEEE Trans. Signal Process. 58 (12) (2010) 6126–6139, http://dx.doi.org/10.1109/TSP.2010.2071866.

[7] S. Callegari, F. Bizzarri, A heuristic solution to the optimisation of flutter control in compression systems (and to some more binary quadratic programming problems) via $\Delta\Sigma$ modulation circuits, in: Proceedings of the ISCAS'10, Paris, FR, 2010, pp. 1815–1818.http://dx.doi.org/10.1109/ISCAS.2010.5537729.

[8] R. Schreier, G.C. Temes, Understanding Delta–Sigma Data Converters, Wiley-IEEE Press, 2004.

[9] S.R. Norsworthy, R. Schreier, G.C. Temes (Eds.), Delta–Sigma Data Converters: Theory, Design, and Simulation, Wiley-IEEE Press, 1996.

[10] R.W. Adams, The design of high-order single-bit $\Delta\Sigma$ ADCs, in: S.R. Norsworthy, R. Schreier, G.C. Temes (Eds.), Delta–Sigma Data Converters: Theory, Design, and Simulation, Wiley-IEEE Press, 1996, pp. 165–192 (Chapter 5).

[11] J.G. Kenney, L.R. Carley, Design of multibit noise-shaping data con-verters, Analog. Integr. Circuits Signal Process. 3 (1993) 259–272.

[12] S. Callegari, F. Bizzarri, Output filter aware optimization of the noise shaping properties of $\Delta\Sigma$ modulators via semi-definite programming, IEEE Trans. Circuits Syst. I 60 (9) (2013) 2352–2365, http://dx.doi.org/10.1109/TCSI.2013.2239091.

[13] S. Callegari, F. Bizzarri, Noise weighting in the design of $\Delta\Sigma$ modulators (with a psychoacoustic coder as an example), IEEE Trans. Circuits Syst. II 60 (11) (2013) 756–760, http://dx.doi.org/ 10.1109/TCSII.2013.2281892.

[14] M. Nagahara, Y. Yamamoto, Frequency domain min–max optimization of noise-shaping delta–sigma modulators, IEEE Trans. Signal Process. 60 (6) (2012) 2828–2839.

[15] X. Li, C.B. Yu, H. Gao, Design of delta–sigma modulators via generalized Kalman–Yakubovich–Popov lemma, Automatica 50 (10) (2014) 2700–2708, http://dx.doi.org/10.1016/j.automatica.2014.09.002.

[16] R. Schreier, M. Snelgrove, Stability in a general $\Sigma\Delta$ modulator, in: 1991 International Conference on Acoustics, Speech, and Signal Processing, 1991. ICASSP-91, vol. 3, 1991, pp. 1769–1772.

[17] B.P. Agrawal, K. Shenoi, Design methodology for $\Delta\Sigma$, IEEE Trans. Commun. COM-31 (3) (1983) 360.

[18] W.L. Lee, A novel high order interpolative modulator topology for high resolution oversampling A/D converters (Master's thesis), Mas-sachusetts Institute of Technology, 1987.

[19] D. Anastassiou, Error diffusion coding for A/D conversion, IEEE Trans. Circuits Syst. 36 (9) (1989) 1175–1186, http://dx.doi.org/10.1109/ 31.34663.

[20] S. Callegari, F. Bizzarri, Should $\Delta\Sigma$ modulators used in ac motor drives be adapted to the mechanical load of the motor?, in: Proceedings of IEEE ICECS 2012, Seville (ES), 2012, pp. 849–852. http://dx.doi.org/10.1109/ICECS.2012.6463619.

[21] S.S. Abeysekera, Z. Zang, Optimal Laguerre filters for sigma–delta demodulator circuits, Signal Process. 80 (1) (2000) 205–209, http: //dx.doi.org/10.1016/S0165-1684(99)00122-X.

[22] C. Dunn, M. Sandler, Psychoacoustically optimal sigma delta modulation, J. Audio Eng. Soc. (AES) 45 (4) (1997) 212–223.

[23] L.J. Hollmann, R.L. Stevenson, Pole-zero placement algorithm for the design of digital filters with fractional-order rolloff, Signal Process. 107 (2014) 218–229, http://dx.doi.org/10.1016/j.sigpro.2014.05.007.

[24] C. Wu, D. Gao, K.L. Teo, A direct optimization method for low group delay FIR filter design, Signal Process. 93 (7) (2013) 1764–1772, http: //dx.doi.org/10.1016/j.sigpro.2013.01.015.

[25] P. Apostolov, Method for FIR filters design with compressed cosine using Chebyshev's norm, Signal Process. 91 (11) (2011) 2589–2594, http://dx.doi.org/10.1016/j.sigpro.2011.05.014.

[26] C.Y.-F. Ho, B.W.-K. Ling, J.D. Reiss, Y.-Q. Liu, K.-L. Teo, Design of interpolative sigma delta modulators via semi-infinite programming, IEEE Trans. Signal Process. 54 (10) (2006) 4047–4051, http: //dx.doi.org/10.1109/TSP.2006.880338.

[27] R. Schreier, The Delta-Sigma Toolbox, Analog Devices, Release 7.4, also known as "DELSIG", 2011, URL ⟨http://www.mathworks.com/matlabcentral/fileexchange/19-delta-sigma-toolbox⟩.

[28] T.W. Parks, C.S. Burrus, Digital filter design, John Wiley & Sons, 1987.
[29] T. Iwasaki, S. Hara, Generalized KYP lemma: unified frequency domain inequalities with design applications, IEEE Trans. Autom. Control 50 (1) (2005) 41–59.
[30] E. De Klerk, Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications, Applied Optimization, Kluwer, 2002.
[31] S. Boyd, L. Vandenberghe, Convex Optimization, 7th ed. Cambridge University Press, 2009.
[32] W.L. Lee, C.G. Sodini, A topology for higher order interpolative coders, in: Proceedings of the ISCAS 1987, vol. 4, 1987, pp. 459–462.
[33] R.A. Wannamaker, Psychoacoustically optimal noise shaping, J. Audio Eng. Soc. 40 (7/8) (1992) 611–620.
[34] A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, Prentice-Hall, 1989.
[35] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear matrix inequalities in system and control theory, in: SIAM Studies in Applied Mathematics, vol. 15, SIAM, Philadelphia, 1994.