Discrete Programming Entailing Circulant Quadratic Forms: Refinement of a Heuristic Approach Based on ΔΣ Modulation

Discrete Programming Entailing Circulant Quadratic Forms: Refinement of a Heuristic Approach Based on ΔΣ Modulation

Sergio Callegari, *Senior Member, IEEE,* Federico Bizzarri, *Senior Member, IEEE,* and Enrico Malaguti, *Member, IEEE*

***Abstract*—A recent result on the potential of ΔΣ modulators (ΔΣMs) as heuristic optimizers for circulant unconstrained discrete quadratic programming (C-UDQP) is revisited, bridging it with current developments on the design of ΔΣMs by semi-definite programming (SDP). This provides an efficient strategy by which one can design a ΔΣM and its input signal from a C-UDQP specification so that the solution of the C-UDQP problem can be found in the ΔΣM output, all with almost no manual intervention. The proposed concept is validated by simulation-based experiments on a benchmark case, comparing the new strategy to previous results and exact optimization techniques.**

***Index Terms*—Delta-Sigma Modulation, Semi-definite Programming, Integer Programming, Optimization.**

I. Introduction

ΔΣ modulators (ΔΣMs) [1], [2] have recently been interpreted as heuristic optimizers for filtered approximation (FA) problems [3]–[5] (Fig. 1). Additionally, it has been established that a specific sub-class of unconstrained discrete quadratic programming (UDQP), namely circulant-UDQP (C-UDQP), exists whose problems can be transformed into a subset of FA, i.e., periodic FA (P-FA) [6]. These results allow ΔΣMs to be used as heuristic solvers for C-UDQP problems not originating from the electronics or signal processing fields. For instance, in [7] a specially crafted ΔΣM was applied to the optimization of a turbomachine with the aim to minimize flutter phenomena. This suggests the possibility to devise hardware solvers for a niche of difficult optimization problems and may help advance the understanding of the fundamental operation of ΔΣMs.

On a related note, the past few years have also witnessed significant developments in the design of ΔΣMs, including approaches based on the Kalman–Yakubovich–Popov (KYP) lemma [8], semi-definite programming (SDP), and efficient interior-point methods [9] for convex optimization [10], [11]. The new strategies [12]–[14] let one formally maximize SNR-related merit factors while enforcing stability constraints expressed via the semi-empiric Lee criterion [2], [15].

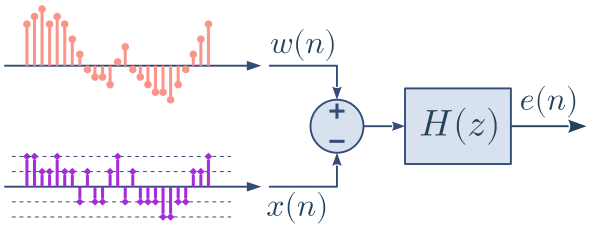

Figure 1. FA setup: signal $x(n)$ is optimized to be as similar as possible to $w(n)$ after filtering through an assigned $H(z)$.

This brief bridges between the two accomplishments above, revisiting the optimization-via-ΔΣ-modulation approach in [6] to incorporate the ΔΣM design techniques in [12], [13] with specific reference to the formalization in [16]. In this way, the need for manual adjustment and expertise that characterizes the method in [6] (notwithstanding its algorithmic formulation and good results) is mostly avoided. The resulting procedure starts at a C-UDQP problem and returns a digital ΔΣM design and a signal so that when the latter is used as the ΔΣM input, the solution of the original problem can be found in the ΔΣM output stream. In comparison to [6], that does not rely on an optimized ΔΣM design, this approach trades one type of optimization, i.e. the original C-UDQP, for another one, namely the SDP used for the modulator design. On large problems, the exchange can be advantageous because the complexity of SDP grows polynomially with the problem size, while C-UDQP is conjectured to be *hard*, with exponentially growing complexity for the known solution methods.

A notable feature of the revisited strategy is that it does not merely cascade a C-UDQP to P-FA transformation [6] with a P-FA to ΔΣM design step [13]. It merges the two, making the explicit derivation of the P-FA filter superfluous. It is fair to anticipate that not *all* the limitations in [6] can be removed. Marginal cases may exist where the ΔΣM fails to correctly operate as a good optimizer. If so, this brief still helps relate the situation to specific features of the original C-UDQP problem.

The proposed concepts are validated by simulation using the same benchmark case as in [7], comparing the revisited strategy to the original one and exact optimization results.

The first two lines here are a placeholder for the manuscript dates (“Manuscript received March 2, 2020; etc.”) included for length estimation. The work of S. Callegari and E. Malaguti has been supported by the project *Metodi di Ottimizzazione e ispirazione biologica nella Rappresentazione del SEgnale (MORSE)* in the framework of the *ALMA IDEA 2017* funding initiative by the University of Bologna.

S. Callegari is with the Advanced Research Center on Electronic Systems “E. De Castro” (ARCES) and the Department of Electrical, Electronic and Information Engineering “Guglielmo Marconi” (DEI) at the University of Bologna, Italy. E-mail: sergio.callegari@unibo.it.

F. Bizzarri is with the Dipartimento di Elettronica, Informazione e Bioingegneria at Politecnico di Milano, Italy and the Advanced Research Center on Electronic Systems “E. De Castro” (ARCES) at the University of Bologna, Italy. E-mail: federico.bizzarri@polimi.it.

E. Malaguti is with the Department of Electrical, Electronic and Information Engineering “Guglielmo Marconi” (DEI) at the University of Bologna, Italy. E-mail: enrico.malaguti@unibo.it.

II. Definitions

A. FA and P-FA problems

In FA, a finite portion of a signal $w(n)$ is given together with a filter $H(z)$ and a set of values $\mathcal{A}$, and one needs to find a portion of a signal $x(n)$ with values in $\mathcal{A}$ so that its difference with respect to $w(n)$ evaluated after the filter $H(z)$ as $e(n)$ has as little energy as possible (Fig. 1—see [6] for a formal definition). The set $\mathcal{A}$ generally has low cardinality, often being binary, as in $\{-1, 1\}$, or ternary, as in $\{-1, 0, 1\}$. To avoid ambiguity, the filter must be specified together with its initial conditions or a warm-up procedure may be prescribed.

P-FA is a variant of FA where one takes $w(n)$ to be periodic on a period N_w and $x(n)$ to be periodic on a period $N = kN_w$ with $k \in \mathbb{N}^+$. The problem is defined over a portion of $w(n)$ and $x(n)$ including N samples, assuming steady-state operation from the filter. This saves the need to specify initial or warm up conditions.

B. UDQP and C-UDQP problems

A UDQP problem is defined by an $N \times N$ real symmetric matrix $\boldsymbol{Q}$, an N-entry real vector $\boldsymbol{L}$, and a set of real values $\mathcal{A}$ as

$$\arg\min_{\boldsymbol{x} \in \mathcal{A}^N} \ \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{L}^{\mathrm{T}} \boldsymbol{x} \tag{1}$$

under the assumption that $\boldsymbol{Q}$ is positive definite.

C-UDQP is a variant of UDQP where $\boldsymbol{Q}$ is circulant [17].

C. ΔΣMs

A ΔΣM is a nonlinear feedback system operating as a smart quantizer for oversampled signals. As shown in Fig. 2, its most basic setup involves a single static quantizer $q(\cdot)$ and a couple of filters $FF(z)$ and $FB(z)$. Modeling the quantizer by the superposition of its quantization error $\varepsilon(n)$, as on the right of the figure, and making the (incorrect, but generally acceptable) assumption that $\varepsilon(n)$ is independent of the ΔΣM input $w(n)$, one can describe the modulator behavior via a signal transfer function $STF(z)$ and a noise transfer function $NTF(z)$ so that $X(z)$, i.e., the transform of the output $x(n)$ is $STF(z)W(z) + NTF(z)E(z)$. Note that in the previous expression and in the rest of this brief, capitalized symbols indicate the Z-transform of the corresponding small-case quantities and vice-versa. Also note that not only $STF(z)$ and $NTF(z)$ are univocally defined from $FF(z)$ and $FB(z)$, but the inverse also holds letting the modulator filters be fully determined from its $STF(z)$ and $NTF(z)$ as long as $ntf(0) = 1$ [2]. Hence, when $STF(z) = 1$ as common, the ΔΣM ends up being fully defined from $q(\cdot)$ and $NTF(z)$. Typically, one wants $NTF(z)$ to have a very low gain in the band of $w(n)$ so that most of the quantization noise at the output $x(n)$ can be filtered away without hindering the information content.

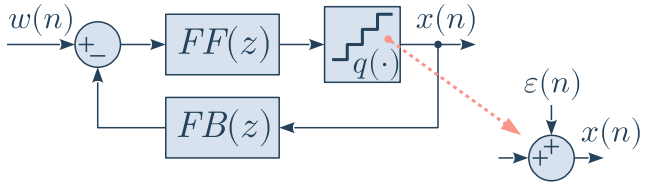

Figure 2. Basic ΔΣM architecture (left) and modeling of its quantizer $q(\cdot)$ as the superposition of a quantization error (right).

III. C-UDQP via ΔΣMs

A. Relationships and properties

1) P-FA and C-UDQP problems: any FA problem is also a UDQP problem. For an intuitive view, collect the entries of $w(n)$ and $x(n)$ under consideration in N-length vectors $\boldsymbol{w}$ and $\boldsymbol{x}$. As long as initial conditions are ignored, their filtering by $H(z)$ can then be represented by multiplication by a suitable $M \times N$ matrix $\boldsymbol{H}$, so that $\boldsymbol{e} = \boldsymbol{H}(\boldsymbol{w} - \boldsymbol{x})$ holds the relevant entries of $e(n)$ and the error power to minimize becomes

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = (\boldsymbol{w}^{\mathrm{T}} - \boldsymbol{x}^{\mathrm{T}})\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{w} - \boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{w} \tag{2}$$

where $\boldsymbol{Q} = \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$. Observe that $\boldsymbol{Q}$ is by construction $N \times N$, symmetric and positive definite. Now, let

$$\boldsymbol{L} = -2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q} \ . \tag{3}$$

With this, the problem is reduced to the form in (1). If the original problem is P-FA, $\boldsymbol{Q}$ turns out circulant [6]. Specifically, once one has defined the discrete Fourier transform (DFT)–inverse DFT (IDFT) pair as

$$\mathsf{a}_k = \sum_{j=0}^{N-1} a_j \, \mathrm{e}^{-\mathrm{i}2\pi \frac{jk}{N}} \qquad a_j = \frac{1}{N} \sum_{k=0}^{N-1} \mathsf{a}_k \, \mathrm{e}^{\mathrm{i}2\pi \frac{jk}{N}} \ , \tag{4}$$

where a is the DFT of $\boldsymbol{a}$, $\boldsymbol{Q}_0$, i.e., the first row of $\boldsymbol{Q}$, turns out to be the IDFT of a vector with entries $|H(\mathrm{e}^{\mathrm{i}2\pi k/N})|^2$ for $k = 0, \ldots, N-1$. Being $\boldsymbol{Q}$ circulant, $\boldsymbol{Q}_0$ defines it all.

More interesting is to show that a C-UDQP problem is also a P-FA problem. Following [6], this can be done *assuming* that a P-FA problem equivalent to some given C-UDQP problem exists and looking at the properties that its input $w(n)$ and filter $H(z)$ must have. Regarding $w(n)$, from (3) one can express its portion of interest as a vector

$$\boldsymbol{w} = -\boldsymbol{Q}^{-1}\boldsymbol{L}/2 \ . \tag{5}$$

In (5), the invertibility of $\boldsymbol{Q}$ is guaranteed by its positive definiteness. For what concerns the filter, the previous properties let one indicate as $\boldsymbol{\lambda}$ the vector with the DFT of $\boldsymbol{Q}_0$, to then note that its entries must be linked to the magnitude response of $H(z)$ at specific frequency values

$$\left| H\left(\mathrm{e}^{\mathrm{i}2\pi \frac{k}{N}}\right) \right| = \sqrt{\lambda_k} \ . \tag{6}$$

In (6), the entries of $\boldsymbol{\lambda}$ are certainly positive since they are also eigenvalues of $\boldsymbol{Q}$, thanks to the circulant property. The previous considerations guarantee that there are not just *one* but infinitely many P-FA problems equivalent to a given C-UDQP one. In [6], a C-UDQP to P-FA conversion procedure is proposed taking an $H(z)$ whose magnitude response is a particular interpolation of the points in (6), and specifically the N-tap finite impulse response (FIR) filter with coefficients

$$h_j = \frac{1}{N} \sum_{k=0}^{N-1} \sqrt{\lambda_k} \, \mathrm{e}^{\mathrm{i}2\pi \frac{jk}{N}} \ . \tag{7}$$

Yet, *any* filter whose power response interpolates the values λ_k passing through them at the appropriate frequencies according to (6) would be adequate [18].

2) FA, P-FA problems and ΔΣMs: FA problems can be solved by ΔΣMs. This can be readily seen for an FA problem where $H(z)$ is "on-off". In this case, it suffices to take a quantizer with output levels in $\mathcal{A}$, $STF(z) = 1$ (at least) in the "on" frequencies of $H(z)$, and an $NTF(z)$ with negligible power response in those same frequencies. The idea can be adapted to FA problems with more general features by using a ΔΣM with $STF(z) = 1$ and an $NTF(z)$ capable of making $\varepsilon(n)$ minimal-power once filtered through $NTF(z)$ and $H(z)$. Given that $\varepsilon(n)$ can be assumed to be white, this means picking $NTF(z)$ such that

$$\int_0^\pi \left|NTF\left(\mathrm{e}^{\mathrm{i}\omega}\right)\right|^2 \left|H\left(\mathrm{e}^{\mathrm{i}\omega}\right)\right|^2 \mathrm{d}\omega \tag{8}$$

is minimal. Intuitively, this entails making $NTF(\mathrm{e}^{\mathrm{i}\omega})$ small in modulus at those frequencies ω where $H(\mathrm{e}^{\mathrm{i}\omega})$ is large.

In [6], this consideration is used to solve P-FA problems by setting the modulator $NTF(z)$ to $\alpha/\hat{H}(z)$, where $\hat{H}(z)$ is a *minimum-phase* version of $H(z)$ (that is, a version of $H(z)$ where the zeros outside the unit circle are mirrored inside it) and α is chosen so that the constraint $ntf(0) = 1$ is respected. Then, the ΔΣM input is set to a sequence obtained by repeating $\boldsymbol{w}$ multiple times and the modulator is operated with *dithering* to enhance its *exploration* abilities (i.e., to help it deliver different output sequences for the same input). Finally, the modulator output is scanned for the N-length sub-sequence that produces the lowest value when used as $\boldsymbol{x}$ in the cost function of the C-UDQP problem associated with the P-FA one.

B. Solving C-UDQP problems by ΔΣMs, original proposal

The previous strategy to solve P-FA problems by ΔΣMs can be cascaded to the C-UDQP transformation in Sect. III-A1 to deliver a working approach to solve C-UDQP problems by ΔΣMs, including problems coming from domains other than electronics or signal processing [7].

Nonetheless, the $NTF(z)$ design phase may require manual intervention and expertise. The reason lies in the strongly nonlinear nature of ΔΣMs that prevents the approximated linear modeling in Sect. II-C from providing stability guarantees [2]. The matter of stability is complex, yet a key concept is that a too-large signal at the quantizer input can exacerbate the nonlinear effects hindering the applicability of the approximated linear model. In the structure in Fig. 2, this may easily happen if the modulator input is too large or if the quantization error is fed back with a too large gain. Thus, conventional wisdom is to limit both the input amplitude and the peak gain of $NTF(z)$. The latter requirement is typically expressed via the *Lee constraint* $\|NTF\|_\infty < \gamma$ [15], where γ is a suitable constant depending on the quantizer features (typically, 1.5 to 2 for one-bit quantizers). Aiming at the use of ΔΣMs for C-UDQP problems, these considerations may impose a manual *compression* of the dynamics of $\boldsymbol{w}$ or some *clipping* of its largest values because $\boldsymbol{w}$ determines the modulator input. Similarly, one may need to compress the magnitude response of $NTF(z)$ to avoid too high peaks. Following [6], both operations can be practiced acting on $\boldsymbol{Q}$ at the C-UDQP to P-FA transformation step. Pre-treating $\boldsymbol{Q}$ to adjust its smallest eigenvalues can lead both to a reduction in the peaks of $\alpha/\hat{H}(z)$ and smaller values of $\boldsymbol{w}$ when the latter is obtained as in (5). Still, having two potentially critical items to be manually taken care of can be a nuisance.

C. Optimal ΔΣM design

Right after the publication of [6] a new paradigm for the design of ΔΣMs was introduced in [13], [16], [19]. Rather than designing modulators under the assumption that an *ideal filter* could then be applied for the removal of quantization noise, the proposal is to design modulators so that they work best when some *pre-assigned filter* $H(z)$ is used for the noise removal. Not only this approach is more consistent with the requirements of many applications, it implicitly introduces an FA framework in the picture. Operationally, one wants to minimize the quantity in (8) while guaranteeing the ΔΣM realizability (which needs $ntf(0) = 1$) and compliance with the Lee constraint. Assuming that the $NTF(z)$ being sought is P-th order FIR and that its coefficients $b_0, \ldots, b_{P-1}$ are collected in a vector $\boldsymbol{b}$, the goal can be achieved by defining a Toeplitz symmetric $P \times P$ real matrix $\tilde{\boldsymbol{Q}}$ where the entries of the first row $\tilde{\boldsymbol{Q}}_0$ are given by the first P entries of the inverse discrete-time Fourier transform (IDTFT) of a weighting function $w_H(\omega) = \left|H(\mathrm{e}^{\mathrm{i}\omega})\right|^2$ as in:

$$\tilde{q}_{0,j} = \frac{1}{2\pi} \int_{-\pi}^{\pi} w_H(\omega)\, \mathrm{e}^{\mathrm{i}j\omega}\, \mathrm{d}\omega \ . \tag{9}$$

The remaining rows are then obtained from $\tilde{\boldsymbol{Q}}_0$ thanks to the Toeplitz property. The matrix ends up positive definite and $\boldsymbol{b}^{\mathrm{T}}\tilde{\boldsymbol{Q}}\boldsymbol{b}$ turns out to be proportional to the quantity in (8) and thus usable as a convex cost function for the $NTF(z)$ optimization. The latter must be constrained by fixing $b_0 = 1$ (i.e., $ntf(0) = 1$) and by a further convex matrix inequality to express the Lee constraint as detailed in [6].

D. Solving C-UDQP problems by ΔΣMs, revisited

The novel ΔΣM design paradigm outlined in the previous section can be applied to the solution of C-UDQP problems by ΔΣMs. To this aim, rather than getting the ΔΣM $NTF(z)$ by inverting the $H(z)$ of the intermediate P-FA problem as $\alpha/\hat{H}(z)$, one can optimally design $NTF(z)$ from the power response of $H(z)$, interpreting it as a weighting function. The first expectable advantage is an explicit management of the Lee constraint that can help obtain a stable ΔΣM without manual adjustment. Yet, there is also another benefit. An explicit derivation of $H(z)$ becomes unnecessary, since only its power response is required and the latter can be directly obtained by interpolating the entries in $\boldsymbol{\lambda}$.

Operationally, as in Sect. III-B, one starts by getting vector $\boldsymbol{\lambda}$ as the DFT of the first row of $\boldsymbol{Q}$. However, at this point one does not define an FIR filter $H(z)$ from (7). Conversely, the entries in $\boldsymbol{\lambda}$ are used to define an *interpolating function* $w_H(\omega)$ passing through λ_k for $\omega = 2\pi k/N$. This $w_H(\omega)$ is then employed for the optimization of a P-order $NTF(z)$ according to Sect. III-C and [16], so defining a ΔΣM. From

this point on, one goes back to the procedure in Sect. III-B and [6]. A $\boldsymbol{w}$ vector is computed according to (5) and used to define the modulator input by its repetition in a long sequence. The modulator is operated with some dither and its output is scanned for all the N-length sub-sequences. Among them, one picks as the heuristic optimum the one delivering the smallest value of the C-UDQP cost function once used as $\boldsymbol{x}$.

A few items are worth noticing. First, this novel strategy has a cost since it trades the original P-FA optimization for the SDP required by the ΔΣM design. However, being convex and defined on a continuous support, the latter can be expected to be more efficiently approachable. Secondly, it is not really a single strategy, rather a *strategy collection*, since different interpolations of $\boldsymbol{\lambda}$ can be adopted. Finally, this strategy decouples the ΔΣM order P from the size N of the C-UDQP problem (while the one in Sect. III-B and [6] made the $NTF(z)$ order $N-1$). Taking a low P, reduces the computational cost in the $NTF(z)$ optimization as well as the implementation/simulation cost of the ΔΣM itself, and may also favor its stability. On the other hand, it may make the ΔΣM less accurate as a heuristic optimizer for the original C-UDQP problem.

E. Solving C-UDQP problems by ΔΣMs, revisited again

In addition to the above, the ΔΣM design strategy in Sect. III-C also opens the way to a third option for solving C-UDQP problems via ΔΣMs. This comes from the observation that $\boldsymbol{\lambda}$ is the DFT of $\boldsymbol{Q}_0$, but it must also be a sampling of the discrete-time Fourier transform (DTFT) of a sequence $r(n)$ containing $\tilde{\boldsymbol{Q}}_0$. Namely,

$$\lambda_k = \sum_{j=0}^{N-1} q_{0,j}\, \mathrm{e}^{-\mathrm{i}2\pi\frac{jk}{N}} = \sum_{n=-\infty}^{\infty} r(n)\, \mathrm{e}^{-\mathrm{i}n\omega} \qquad (10)$$

at $\omega = 2\pi k/N$ when $r(n) = \tilde{q}_{0,n}$ for $n \in \{0,\dots,P-1\}$. Restricting to cases where $r(n) = 0$ for $n < 0$ or $n > N-1$, the equality becomes

$$\lambda_k = \sum_{j=0}^{N-1} q_{0,j}\, \mathrm{e}^{-\mathrm{i}2\pi\frac{jk}{N}} = \sum_{n=0}^{N-1} r(n)\, \mathrm{e}^{-\mathrm{i}2\pi\frac{nk}{N}} \qquad (11)$$

which can be satisfied by simply taking $r(n) = q_{0,n}$. Evidently, it is possible to derive $\tilde{\boldsymbol{Q}}$ directly from $\boldsymbol{Q}$ by taking $P \le \lfloor N/2 \rfloor$ and setting $\tilde{q}_{0,j} = q_{0,j}$ for $j = 0,\dots,P$. This approach frees from having to deal with the P-FA problem representation altogether. It also saves having to compute the DFT of $\boldsymbol{Q}_0$ and having to find an interpolating function for it.

Again, a few observations are worth making. First, in relation to the strategy in Sect. III-D, this approach corresponds to taking a specific interpolation for $w_H(\omega)$, namely the one corresponding to the DTFT of $\boldsymbol{Q}_0$. Secondly, in this strategy, P and N are only partially decoupled, because one can only take $P \le \lfloor N/2 \rfloor$. Furthermore, it is evident that to take advantage of all the data in the C-UDQP problem specification, one should take P exactly equal to $\lfloor N/2 \rfloor$.

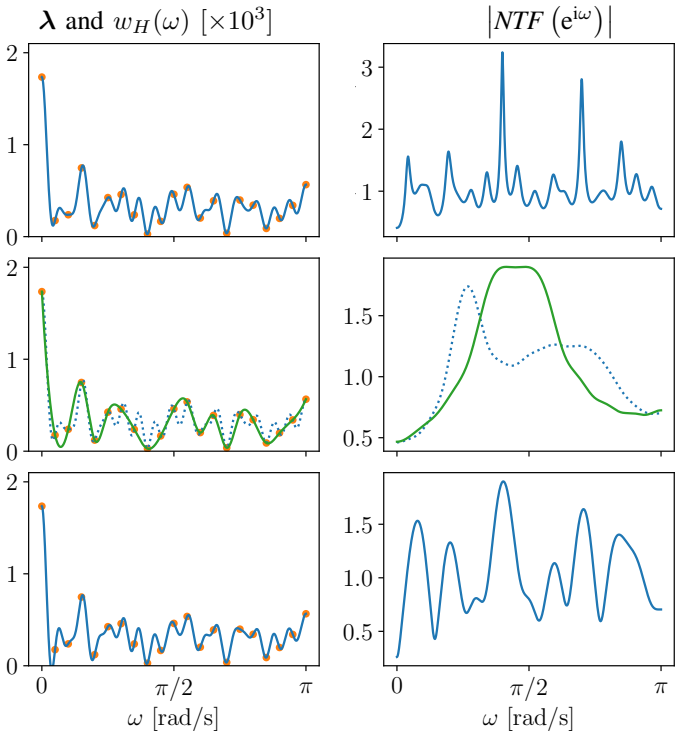

Figure 3. Plots from the validation tests, based on the benchmark problem in [7]. Top row: original ΔΣM design strategy in Sect. III-B; Middle row: revisited strategy in Sect. III-D; Bottom row: revisited strategy in Sect. III-E. In each row, the left plot shows the $\boldsymbol{\lambda}$ values and an interpolant thereof, while the right plot is the magnitude response of the resulting $NTF(z)$.

IV. Examples and Validation

To provide some examples and validation of the proposed concepts, the sample C-UDQP problem employed in [7] is used as a benchmark. The problem comes from a formalization of the requirements to achieve optimal flutter control in compression systems by blade mistuning [20]. It provides an appealing test bench since: (i) it derives from a non-electronic domain; (ii) it is sufficiently small ($N = 40$) to have a known exact solution. For the original definition, refer to the binary case in [21, Example 4], while [7] illustrates how to convert such definition into a proper C-UDQP formulation.

The top plots in Fig. 3 illustrate the application of the original strategy in Sect. III-B. The left pane shows the $\boldsymbol{\lambda}$ entries at their corresponding frequencies as well as the power response of the interpolating FIR $H(z)$. The right pane illustrates the magnitude response of the ΔΣM $NTF(z) \propto 1/\hat{H}(z)$, with the $NTF(z)$ order being $N-1 = 19$. By running the ΔΣM for 12500 periods with dither standard deviation set at $\sigma_d = 0.2$, one can find the exact solution of the flutter problem, at merit factor 9024, in about 20 % of the runs, with the median at 8880 (higher values are better since the original problem is a maximization and 0 is a baseline for no blade mistuning). In one case, ΔΣM instability was detected in the experiments.

The middle row of Fig. 3 illustrates the application of the revisited strategy in Sect. III-D. Again, in the left pane, one has the $\boldsymbol{\lambda}$ entries, now plotted together with two different interpolating functions $w_H(\omega)$: the power response of a FIR $H(z)$ computed as in the previous case (blue, dashed); and a

direct *cubic spline interpolation* of the $\boldsymbol{\lambda}$ values (green solid). On the right, matched by color and line style with the previous curves, one has the magnitude responses of the two $NTF(z)$ designed by SDP from these weightings. The Lee constraint used for the $NTF(z)$ design takes $\gamma = 1.9$ and the modulator order is $P = 19$ for fairness of comparison with the previous case. It turns out that with this design strategy the ΔΣM is a less capable solver for the optimization problem, particularly for the FIR-based $w_H(\omega)$. In this case, it is extremely rare to get the truly optimal solution, and the merit factor median is slightly below 0 (i.e., mistuning worsens flutter). With the spline interpolant, the median merit factor is about 4200. Trying to rise the modulator order P above 19 improves these results only marginally. For $P = 19$ the SDP optimization takes about 2.5 s on a 2013 Intel i7-4750HQ Haswell (4$^{\text{th}}$ gen) mobile CPU, this being the *slow phase* in using the ΔΣM as a C-UDQP optimizer with this approach.

Finally, the third row of Fig. 3 illustrates the application of the revisited strategy in Sect. III-E. In left pane one has, as usual, the $\boldsymbol{\lambda}$ entries and an interpolation thereof, the latter now being implicitly defined by the approach as the DTFT of a sequence $r(n)$ such that $r(n) = q_{0,n}$ for $0 \leq n \leq N-1$ and zero elsewhere. In the right pane, one can find the corresponding magnitude responses of the optimized $NTF(z)$. The order of $NTF(z)$ is now $P = \lfloor N/2 \rfloor = 20$ and its design employs $\gamma = 1.9$ as before. Operated in the usual conditions, the ΔΣM can now find the truly optimal solution of the original problem in ∼20 % of the runs, getting the median merit factor at 8880. The $NTF(z)$ design phase now takes about 2 s.

V. Discussion and Conclusions

The results in the previous section confirm the possibility to heuristically solve C-UDQP problems via ΔΣ modulation and show that revisiting the technique in [6] as proposed can simplify and make more reliable the ΔΣM design phase (reducing the need of manual adjustments to prevent instability) paying an acceptable computational cost for this advantage. They also illustrate that the effectiveness of the revised techniques is sensitive to the specific choice of interpolating function used in the procedure. This is evident not just from the different merit factors being achieved in the benchmark case, but more strikingly from the quite different noise transfer functions obtained for the ΔΣM (e.g., note the difference in the magnitude responses in the right-hand plots in Fig. 3). For an intuitive explanation, observe that the optimal $NTF(z)$ design employed in the approach delivers FIR forms that may have issues in conforming to the constraints set by some interpolations, particularly when they have fine details. The smoothing of $|NTF(\mathrm{e}^{\mathrm{i}\omega})|$ in the middle row of Fig. 3 is eloquent in this sense. Interestingly, the final version of the approach (in Sect. III-E) seems to automatically pick a favorable interpolation. For what concerns computational efficiency, in the benchmark the ∼2 s taken by the ΔΣM design in the proposed approach compare favorably with the > 60 s reported in [21] for a finely tuned specialized exact solver even considering a 5–8-fold performance improvement between the computer used for those tests and the current one.

References

[1] J. M. de la Rosa, "Sigma-Delta modulators: Tutorial overview, design guide, and state-of-the-art survey", *IEEE Trans. Circuits Syst. I*, vol. 58, no. 1, pp. 1–21, Jan. 2011. DOI: 10.1109/TCSI.2010.2097652

[2] S. Pavan, R. Schreier, and G. C. Temes, *Understanding Delta-Sigma data converters*, 2nd ed. Wiley-IEEE Press, 2017.

[3] A. K. Gupta and O. M. Collins, "A new interpretation and extension of ΣΔ modulation", in *Proc. of the IEEE International Symposium on Information Theory (ISIT'01)*, Washington, DC, USA, Jun. 2001, p. 194.

[4] F. Bizzarri, S. Callegari, R. Rovatti, and G. Setti, "On the synthesis of periodic signals by discrete pulse-trains and optimisation techniques", in *Proc. of ECCTD'09*, Antalya (TR), Aug. 2009, pp. 584–587. DOI: 10.1109/ECCTD.2009.5275053

[5] F. Bizzarri, C. Buchheim, S. Callegari, A. Caprara, A. Lodi, R. Rovatti, and G. Setti, "Practical solution of periodic filtered approximation as a convex quadratic integer program", in *Procedings of the First International Conference on Complex Systems Design and Management (CSDM)*. Paris: Springer, Oct. 2010, pp. 149–160. DOI: 10.1007/978-3-642-15654-0_11

[6] S. Callegari, F. Bizzarri, R. Rovatti, and G. Setti, "On the approximate solution of a class of large discrete quadratic programming problems by ΔΣ modulation: the case of circulant quadratic forms", *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6126–6139, Dec. 2010. DOI: 10.1109/TSP.2010.2071866

[7] S. Callegari and F. Bizzarri, "A heuristic solution to the optimisation of flutter control in compression systems (and to some more binary quadratic programming problems) via ΔΣ modulation circuits", in *Proc. of ISCAS'10*, Paris, FR, May 2010, pp. 1815–1818. DOI: 10.1109/ISCAS.2010.5537729

[8] T. Iwasaki and S. Hara, "Generalized KYP lemma: Unified frequency domain inequalities with design applications", *IEEE Trans. Autom. Control*, vol. 50, no. 1, pp. 41–59, Jan. 2005. DOI: 10.1109/TAC.2004.840475

[9] E. De Klerk, *Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications*, ser. Applied Optimization. Kluwer, 2002.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, 7th ed. Cambridge University Press, 2009.

[11] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, ser. SIAM studies in applied mathematics. Philadelphia: SIAM, 1994, vol. 15.

[12] M. Nagahara and Y. Yamamoto, "Frequency domain Min-Max optimization of noise-shaping Delta-Sigma modulators", *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2828–2839, Jun. 2012. DOI: 10.1109/TSP.2012.2188522

[13] S. Callegari and F. Bizzarri, "Output filter aware optimization of the noise shaping properties of ΔΣ modulators via semi-definite programming", *IEEE Trans. Circuits Syst. I*, vol. 60, no. 9, pp. 2352–2365, Sep. 2013. DOI: 10.1109/TCSI.2013.2239091

[14] X. Li, C. B. Yu, and H. Gao, "Design of delta–sigma modulators via generalized Kalman-Yakubovich-Popov lemma", *Automatica*, vol. 50, no. 10, pp. 2700–2708, Oct. 2014. DOI: 10.1016/j.automatica.2014.09.002

[15] W. L. Lee, "A novel high order interpolative modulator topology for high resolution oversampling A/D converters", Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 1987.

[16] S. Callegari and F. Bizzarri, "Noise weighting in the design of ΔΣ modulators (with a psychoacoustic coder as an example)", *IEEE Trans. Circuits Syst. II*, vol. 60, no. 11, pp. 756–760, Nov. 2013. DOI: 10.1109/TCSII.2013.2281892

[17] R. M. Gray, "Toeplitz and circulant matrices: A review", *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006. [Online]. Available: http://ee.stanford.edu/~gray/toeplitz.pdf

[18] L. Lennart, *System Identification: Theory for the User*, 2nd ed. Prentice Hall, 1999.

[19] S. Callegari and F. Bizzarri, "Optimal design of the noise transfer function of ΔΣ modulators: IIR strategies, FIR strategies, FIR strategies with preassigned poles", *Signal Processing*, vol. 114, pp. 117–130, Sep. 2015. DOI: 10.1016/j.sigpro.2015.02.001

[20] B. Shapiro, "Symmetry approach to extension of flutter boundaries via mistuning", *Journal of Propulsion and Power*, vol. 14, no. 3, pp. 354–356, May-Jun. 1998.

[21] N. T. Phuong, H. Tuy, and F. Al-Khayyal, "Optimization of a quadratic function with a circulant matrix", *Computational Optimization and Applications*, vol. 35, no. 2, pp. 135–159, Oct. 2006.