IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 5, 2018, accepted August 23, 2018, date of publication September 3, 2018, date of current version September 21, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2868266

On the Design of Channel Shortening Demodulators for Iterative Receivers in Linear Vector Channels

SHA HU, (Member, IEEE), AND FREDRIK RUSEK
Department of Electrical and Information Technology, Lund University, SE-22100 Lund, Sweden
Corresponding author: Sha Hu (sha.hu@eit.lth.se)

ABSTRACT We consider designing demodulators for linear vector channels that use reduced-size trellis descriptions for the received signal. We assume an iterative receiver and use interference cancellation (IC) with soft-information provided by an outer decoder to mitigate the signal part that is not covered by a reduced-size trellis description. In order to reach a trellis description, a linear filter is applied as a front end to compress the signal structure into a small trellis. This process requires three parameters to be designed: 1) the front-end filter; 2) the feedback filter through which the IC is done; and 3) a target response which specifies the trellis. Demodulators of this form have been studied before under the name *channel shortening* (CS), but the interplay between CS, IC, and the trellis-search processes has not been adequately addressed in the literature. In this paper, we analyze two types of CS demodulators that are based on the Forney and Ungerboeck detection models, respectively. The parameters are jointly optimized with a generalized mutual information (GMI) function. We also introduce a third type of CS demodulator that is, in general, suboptimal, which has closed-form solutions. Furthermore, signal-to-noise ratio asymptotic properties are analyzed, and we show that the third CS demodulator asymptotically converges to the optimal CS demodulator in the sense of GMI maximization.

INDEX TERMS Channel shortening (CS), intersymbol interference (ISI), multi-input multi-output (MIMO), front-end filter, feedback filter, target response, generalized mutual information (GMI), Forney model, Ungerboeck model, turbo equalization, demodulator, linear minimum mean square error (LMMSE), interference cancellation (IC), extrinsic information transfer (EXIT), block-error-rate (BLER), BCJR.

I. INTRODUCTION

Channel shortening (CS) demodulators have a long and rich history, see [3]–[16]. For intersymbol interference (ISI) channels, Forney, Jr., [17] showed that Viterbi Algorithm (VA) [19] implements maximum likelihood (ML) detection. However, the complexity of VA is exponential in the memory of the channel which prohibits its use in many cases of interest. As a remedy, Falconer and Magee [4] proposed in 1973 the concept of CS. The concept is to filter the received signal with a prefilter such that the effective channel has a much shorter duration than the original one, and then apply VA to the shortened channel.

Traditionally, CS demodulators have been optimized from various criteria such as minimum mean square error (MMSE) and signal-to-interference-plus-noise ratio (SINR) [5]–[13]. Venkataramani and Erden [14] attempted to minimize the error probability of an uncoded system, which leads to a new notion of posterior equivalence between the target response and the filtered channel. But as [14] works with uncoded error probabilities, the analysis does not adequately address the case of coded systems and Shannon capacity properties. The first paper that works with capacity-related cost measures is [15] where the authors considered the achievable rate, in the form of generalized mutual information (GMI) [18], [20]–[23], that the transceiver system can achieve if a CS demodulator is adopted. However, [15] is limited to ISI channels only, and the design method in [15] of the CS demodulator is in fact not always possible to execute. The limitations of [15] were first dealt with in [18], which extended the CS concept to any linear vector channel such as multi-input multi-output (MIMO) and ISI channels, and resulted in a closed-form optimization procedure.

On the other hand, iterative receivers such as turbo equalization [24]–[28] followed as a natural extension to

turbo codes for designs of iterative detection and decoding receivers. When it comes to turbo equalization, common settings of the demodulator are [24] the maximum *a posterior* (MAP) demodulator [43] and its suboptimal variants such as dimension-reduction and subspace based detections [29], [30], and linear MMSE (LMMSE) [31], [32]. The suboptimal demodulators replace the MAP with a linear equalizer or a decision feedback equalizer (DFE) to reduce the prohibitive complexity. One open problem in the area of turbo equalization is the development of other non-trellis-based detection methods that provide performance between MAP and LMMSE [24], [32].

Instead of fully removing the trellis-based detection, another approach is to reduce the memory-size of the original linear vector channel through an interference cancellation (IC) based prefiltering. To the best of our knowledge, there is limited literature [26], [33] on such a demodulator design that combines both IC based prefiltering and a memory-size shortened BCJR in an iterative receiver. A closely related concept is delayed-decision-feedback-sequence-estimation (DDFSE) [34], [35], which also reduces the number of states in the BCJR. However, in DDFSE the IC is done within a single iteration, and not between the iterations.

In this paper, we generalize the GMI-maximization based CS demodulator in [18] to cooperate with iterative receivers. With iterative receivers, it is reasonable to expect that better detection-performance can be attained by allowing the parameters of CS demodulator to change over each iteration. However, the CS demodulator in [18] does not take the prior information into account, rendering a static design in all iterations. We aim at constructing a CS demodulator that takes soft-information provided by the outer-decoder into account such that the parameters of CS demodulator are designed for a particular level of prior knowledge. This procedure includes an IC mechanism to deal with the signal part that are not handled by the trellis-search, i.e., the BCJR. Preliminary results for CS demodulators in iterative receivers are available in [2] and [36], but this paper non-trivially advances the state-of-the-art.

Although the trellis-search based detection is still utilized in CS demodulator, the memory-size ν of the linear vector channel has been reduced which results in significant complexity-reduction compared to MAP. Meanwhile, with different values of ν, CS demodulator provides trade-offs between the performance of LMMSE and MAP. As what will become clear later that CS demodulator is closely related to the concept of LMMSE with parallel interference cancellation (LMMSE-PIC) [37]–[39], which cooperates the soft-information into its filter-coefficients and IC process. By setting $\nu = 0$ in CS demodulator, it is identical to LMMSE-PIC whose trellis-search process is trivial as different layers are assumed to be independent after the front-end filtering. However, CS demodulator advances the LMMSE-PIC by including a trellis-search, where the parameters of the front-end filter, IC, and the trellis-search are jointly optimized. On the other hand, by setting ν to the original memory-size of the linear vector channel, CS demodulator is identical to MAP. Therefore, CS demodulator provides a generalized framework that includes LMMSE-PIC and MAP as two extreme cases for designing an iterative receiver.

The rest of the paper is organized as follows: The linear vector channel model and the iterative receiver structure are introduced in Section II, while the general form of CS demodulator and the GMI are described in Section III. In Section IV we analyze three types of CS demodulators for MIMO channels. In Section V we deal with ISI channels as asymptotic versions of the results established in Section IV. The signal-to-noise ratio (SNR) asymptotic of the CS demodulators are discussed in Section VI. Numerical results are then shown in Section VII, and Section VIII summarizes the paper.

NOTATION

Throughout the paper, a capital bold letter such as $\boldsymbol{A}$ represents a matrix, a lower case bold letter $\boldsymbol{a}$ represents a vector, and a capital letter A represents a number. The expression $\boldsymbol{A} \prec \boldsymbol{0}$ means matrix $\boldsymbol{A}$ is negative-definite, while $\boldsymbol{A} \succ \boldsymbol{0}$ means $\boldsymbol{A}$ is positive-definite. We let $\boldsymbol{I}_K$ represent a $K \times K$ identity matrix and the dimension is omitted when it can be understood from the context. The superscripts have the following meanings: $(\cdot)^*$ is complex conjugate, $(\cdot)^{\mathrm{T}}$ is matrix transpose, $(\cdot)^\dagger$ denotes the conjugate transpose of a matrix, $(\cdot)^{-1}$ is matrix inverse. In addition, $\propto$ means proportional to, $\mathbb{E}[\cdot]$ is the expectation operator, $\mathrm{Tr}(\cdot)$ takes the trace of a matrix, $\mathcal{R}\{\cdot\}$ returns the real part of a variable, $\otimes$ is the Kronecker multiplication operator, $\mathrm{vec}(\boldsymbol{A})$ is a column vector containing the columns of matrix $\boldsymbol{A}$ stacked on top of each other, and $[A, B]$ is the set of integers $\{k : A \le k \le B\}$.

Further, we say that a matrix $\boldsymbol{A}$ is banded within diagonals $[-\nu_1, \nu_2]$ $(\nu_1, \nu_2 \ge 0)$, if the (k, ℓ)th element $A(k, \ell)$ satisfies[1]

$$A(k, \ell) = 0, \ \ell - k > \nu_1 \ \text{ or } \ k - \ell > \nu_2.$$

Moreover, we define two matrix notations $[\,]_\nu$ and $[\,]_{\setminus\nu}$ such that $\boldsymbol{A} = [\boldsymbol{A}]_\nu + [\boldsymbol{A}]_{\setminus\nu}$, and $[\boldsymbol{A}]_\nu$ is banded within diagonals $[-\nu, \nu]$ where $[\boldsymbol{A}]_{\setminus\nu}$ are constrained to zeros.

II. SYSTEM MODEL

We consider linear vector channels according to the received signal model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{y}$ is an $N \times 1$ vector of the received signal; $\boldsymbol{x}$ is a $K \times 1$ vector comprising unit-energy coded symbols that belong to a constellation $\mathcal{X}$; $\boldsymbol{H}$ is an $N \times K$ matrix representing the communication channel which is perfectly known to the receiver; and $\boldsymbol{n}$ is zero-mean complex-valued Gaussian noise vector with a covariance matrix $N_0\boldsymbol{I}$.

The model (1) may represent many different communication systems and we consider two typical cases, i.e., when

[1]Note that ν_1 refers to the number of upper diagonals of $\boldsymbol{A}$ that are non-zero. We have this convention in order to subsequently follow standard notation for Toeplitz matrices [40].

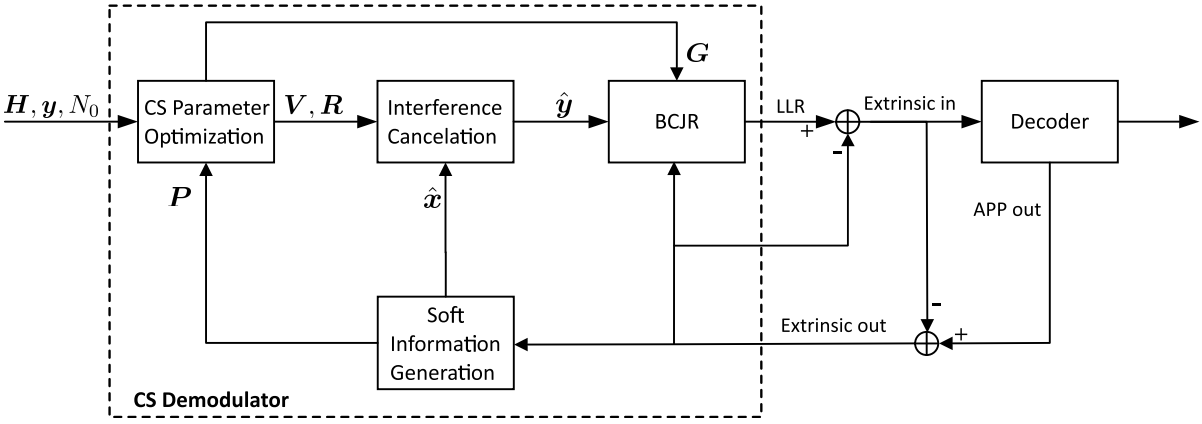


FIGURE 1. Iterative receiver structure with CS demodulator and outer-decoder. The target of the CS demodulator is to maximize the GMI through jointly optimizing the parameters V, R, and G, which are referred to as the front-end filter, IC matrix, and trellis-representation matrix, respectively.

$\boldsymbol{H}$ represents a multi-input multi-output (MIMO) or an ISI channel. In the MIMO case, the variables N and K are finite while they grow without bounds in the ISI case. For the MIMO case, a block-fading model is assumed to perform an analysis for a whole transmitted data-block.

In an iterative receiver, the feedbacks from the outer-decoder can be utilized to improve the performance. As the outer-decoder provides the demodulator with *a posteriori* probability (APP) and extrinsic information (in terms of bit log-likelihood ratio (LLR)) [41], [42], side-information is present about the symbols $\boldsymbol{x}$ and we represent this by the probability mass function $p_k(s) = \mathrm{P}(x_k = s), (0 \le k \le K-1)$. Note that the side-information does not consider the dependency among the symbols, but are symbol-wise marginal probabilities. This reflects the situation encountered in iterative receivers with perfect interleaving. In those cases, the prior probabilities provided from previous iterations are assumed independent, i.e., $\mathrm{P}(\boldsymbol{x} = \boldsymbol{s}) = \prod p_k(s)$, and the demodulator can compute $\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}] = [\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_{K-1}]^T$ in a per-entry fashion as

$$\hat{x}_k = \sum_{s \in \mathcal{X}} s p_k(s),$$

where the expectations are computed with respect to the prior distribution $p_k(s)$.

With soft-information $\hat{\boldsymbol{x}}$, we define a $K \times K$ diagonal matrix $\boldsymbol{P}$ reflecting the accuracy of the side-information as

$$\boldsymbol{P} = \mathbb{E}_{\boldsymbol{y}}[\boldsymbol{x}\hat{\boldsymbol{x}}^\dagger] = \mathbb{E}_{\boldsymbol{y}}[\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\dagger], \tag{2}$$

where $\boldsymbol{x}$ is the transmitted symbol for the received signal $\boldsymbol{y}$, and the exception "$\mathbb{E}_{\boldsymbol{y}}$" is taken by averaging $\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}^\dagger$ through the whole transmitted data-block. Note that $\boldsymbol{0} \preceq \boldsymbol{P} \preceq \boldsymbol{I}$, and when there is no soft-information available we have $\boldsymbol{P} = \boldsymbol{0}$, while with perfect feedback we get $\boldsymbol{P} = \boldsymbol{I}$.

For the ISI case, as there is only a single transmit-layer and K is the length of data-block, $\boldsymbol{P}$ can be simplified as

$$\boldsymbol{P} = \alpha \boldsymbol{I}, \tag{3}$$

where

$$\alpha = \frac{1}{K} \sum_{k=0}^{K-1} |\hat{x}_k|^2.$$

The task of the demodulator is to generate soft-information about $\boldsymbol{x}$ given the observable $\boldsymbol{y}$ and the side-information. The optimal demodulator is the MAP [43], [44] which evaluates the posterior probabilities $\mathrm{P}(x_k = s|\boldsymbol{y})$. However, the number of leaves of the search-tree in MAP is in general $|\mathcal{X}|^K$ which is prohibitive for practical applications. With CS demodulator we force the signal model to be an lower-triangular matrix with only $\nu+1$ $(0 \le \nu < K-1)$ non-zero diagonals, by means of a linear filter,[2] where ν is referred to as the memory-size of the CS demodulator. Then, the BCJR [45] can be applied over a trellis with $|\mathcal{X}|^\nu$ states. Further, as there is side-information present about $\boldsymbol{x}$, the parts of $\boldsymbol{H}$ that are outside the memory of BCJR can be partly eliminated by means of IC through the prior mean $\hat{\boldsymbol{x}}$.

The structure of an iterative receiver utilizing a CS demodulator is depicted in Fig. 1. The extrinsic information from an outer-decoder is used to compute an estimate $\hat{\boldsymbol{x}}$ and a matrix $\boldsymbol{P}$ that reflects the feedback quality. Based on the updated $\boldsymbol{P}$ in each iteration, the optimal CS parameters are solved by maximizing the GMI. A prefiltering and IC process are then implemented on $\boldsymbol{y}$ with optimal $\boldsymbol{V}$ and $\boldsymbol{R}$ to obtain the signal $\hat{\boldsymbol{y}}$, which is sent to a memory-ν BCJR module specified by an optimal $\boldsymbol{G}$. Further, the extrinsic information iteratively exchanged between the BCJR and the outer-decoder is also used as *a priori* information for decoding the transmitted symbols. Note that if we set $\nu = K-1$, the search-space of the CS demodulator is no longer a trellis but corresponds to the original tree and is equivalent to the MAP, while the LMMSE-PIC is a special case of the CS demodulation with $\nu = 0$.

[2]For finite length linear vector channels such as MIMO channel, "filtering" means matrix multiplication.

III. THE GENERAL FORM OF THE CS DEMODULATOR

We state two lemmas that are useful later, and Lemma 2 can be verified straightforwardly.

Lemma 1: Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K \times K$ matrices, where $\boldsymbol{A}_1$ is invertible and banded within diagonals $[-\nu, \nu]$. If $[\boldsymbol{A}_1^{-1}]_\nu = [\boldsymbol{A}_2]_\nu$, then

$$\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big) = \mathrm{Tr}\big(\boldsymbol{I}\big).$$

Proof: Let $\boldsymbol{A}_3 = \boldsymbol{A}_2 - \boldsymbol{A}_1^{-1}$, then it holds that $[\boldsymbol{A}_3]_\nu = \boldsymbol{0}$ and $\boldsymbol{A}_3 = [\boldsymbol{A}_3]_{\setminus\nu}$. As $\boldsymbol{A}_1 = [\boldsymbol{A}_1]_\nu$, the elements along the main diagonal of $\boldsymbol{A}_1\boldsymbol{A}_3$ are zeros. Hence, we have $\mathrm{Tr}\big(\boldsymbol{A}_1\boldsymbol{A}_2\big) = \mathrm{Tr}\big(\boldsymbol{A}_1(\boldsymbol{A}_1^{-1} + \boldsymbol{A}_3)\big) = \mathrm{Tr}(\boldsymbol{I})$. ∎

Lemma 2: Let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be two $K \times K$ matrices that are banded within diagonals $[-\nu_1, \nu_2]$ and $[-\nu_3, \nu_4]$, respectively. Then the product $\boldsymbol{A}_1\boldsymbol{A}_2$ is banded within diagonals $[\max(-(\nu_1+\nu_3), 1-K), \min(\nu_2+\nu_4, K-1)]$.

A. SYSTEM MODEL OF THE CS DEMODULATOR

The CS demodulators that we investigate operate on the basis of a mismatched[3] function

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}})\} - \boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big) \tag{4}$$

for given feedbacks $\hat{\boldsymbol{x}}$, instead of the true conditional probability distribution function (pdf)

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(\pi N_0)^N} \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2}{N_0}\right). \tag{5}$$

Note that $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ may not be a valid pdf, but this is irrelevant for demodulation, see [46]. The matrices $\boldsymbol{V}$, $\boldsymbol{R}$, and $\boldsymbol{G}$ are denoted as the front-end filter, IC matrix, and trellis-representation matrix as mentioned earlier, respectively. Without loss of generality, we have absorbed the noise power N_0 into them. Detection models (4) and (5) are equivalent if we set $\boldsymbol{V} = \boldsymbol{H}^\dagger/N_0$, $\boldsymbol{R} = \boldsymbol{0}$, and $\boldsymbol{G} = \boldsymbol{H}^\dagger\boldsymbol{H}/N_0$, in which case the CS demodulator represents the MAP.

The detection model (4) has its roots in Falconer and Magee's paper [4] with an additional IC process, in which case it is described as

$$\tilde{T}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(-\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{T}\hat{\boldsymbol{x}} - \boldsymbol{F}\boldsymbol{x}\|^2\big) \tag{6}$$

By setting $\boldsymbol{T} = \boldsymbol{0}$, we obtain the same model in [4]. If identifying $\boldsymbol{V} = \boldsymbol{F}^\dagger\boldsymbol{W}$, $\boldsymbol{R} = \boldsymbol{F}^\dagger\boldsymbol{T}$, and $\boldsymbol{G} = \boldsymbol{F}^\dagger\boldsymbol{F}$, the model (6) is equivalent to (4) since

$$\begin{aligned}\tilde{T}(\boldsymbol{y}|\boldsymbol{x}) &\propto \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{F}^\dagger\boldsymbol{W}\boldsymbol{y} - \boldsymbol{F}^\dagger\boldsymbol{T}\hat{\boldsymbol{x}})\} - \boldsymbol{x}^\dagger\boldsymbol{F}^\dagger\boldsymbol{F}\boldsymbol{x}\big)\\ &= \exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger(\boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}})\} - \boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big).\end{aligned}$$

Detection model (6) is usually denoted as "Forney" model due to its Euclidean-distance form, while the general model (4) is called "Ungerboeck" model [47]–[49]. An advantage of the Ungerboeck model is that the parameter optimization through GMI-maximization is simpler [18] than the Forney model. However, as both models can be viewed as "natural" CS demodulators, we investigate both in CS demodulator design for iterative receivers in this paper.

In order to optimize $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$, we choose to work with the GMI which is an achievable rate for a receiver that operates on the basis of a mismatched version of the channel law. The GMI in nats per channel-use is defined as

$$I_{\mathrm{GMI}} = -\mathbb{E}_{\boldsymbol{y}}\left[\log\tilde{p}(\boldsymbol{y})\right] + \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\log\tilde{p}(\boldsymbol{y}|\boldsymbol{x})\right] \tag{7}$$

where

$$\tilde{p}(\boldsymbol{y}) = \frac{1}{\pi^K}\int \tilde{p}(\boldsymbol{y}|\boldsymbol{x})\exp(-\|\boldsymbol{x}\|^2)\mathrm{d}\boldsymbol{x}$$

and the expectation is taken over the true pdfs $p(\boldsymbol{y})$ and $p(\boldsymbol{x}, \boldsymbol{y})$. Although finite constellations $\mathcal{X}$ are used in practice, they are hard to analyze. In order to obtain a mathematically tractable problem, here we use a zero-mean complex-valued Gaussian constellation with an unit-variance for each entry in $\boldsymbol{x}$. With Gaussian inputs, the trellis discussed earlier has no proper meaning as the number of states is infinite. However, the Gaussian assumption is only made to design the CS parameters $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$ that can also be used for finite constellations.

We first state Theorem 1 that states the GMI for model (4).

Theorem 1: The GMI for the detection model (4) equals

$$\begin{aligned}&I_{\mathrm{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})\\ &\quad= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}(\boldsymbol{G}) + 2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H} - \boldsymbol{R}\boldsymbol{P})\big\}\\ &\qquad - \mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}(N_0\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\dagger)\boldsymbol{V}^\dagger\\ &\qquad - 2\mathcal{R}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^\dagger\big\} + \boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^\dagger\big)\big).\end{aligned} \tag{8}$$

The proof of Theorem 1 is given in Appendix A. Here we use the fact $(\boldsymbol{I}+\boldsymbol{G}) \succ \boldsymbol{0}$ shown in [18], otherwise the GMI is not well-defined. With any parameters $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$, the GMI can be calculated via (8), however, they may not be optimal in the sense of GMI-maximization.

We next illustrate Theorem 1 with two examples.

Example 1: Extended Zero-Forcing filter (EZF). We extend the zero-Forcing filter [50] to only partly invert the channel so that a trellis-search is necessary after the EZF. In view of the CS demodulator, we select the parameters in (4) as:

$$\boldsymbol{V} = (\boldsymbol{I}+\boldsymbol{G})(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\boldsymbol{H}^\dagger, \quad \text{and } \boldsymbol{R} = \boldsymbol{0},$$

and then optimize (8) over $\boldsymbol{G}$. To satisfy the constraint of having a trellis with $|\mathcal{X}|^\nu$ states, it must hold $\boldsymbol{G} = [\boldsymbol{G}]_\nu$. The optimal $\boldsymbol{G}$, in the sense of maximizing (8), is shown (later in Theorem 2) to satisfy

$$[(\boldsymbol{I}+\boldsymbol{G})^{-1}]_\nu = N_0[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu.$$

Utilizing Lemma 1, the GMI in (8) for the optimal $\boldsymbol{G}$ equals

$$\begin{aligned}I_{\mathrm{GMI}} &= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\boldsymbol{I} - N_0(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}(\boldsymbol{I}+\boldsymbol{G})\big)\\ &= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big).\end{aligned}$$

Example 2: Truncated Matched filter (TMF). As previously mentioned, the MAP demodulator (5) can be written in the form (4) by setting $\boldsymbol{V} = \boldsymbol{H}^\dagger/N_0$, $\boldsymbol{R} = \boldsymbol{0}$, and $\boldsymbol{G} = \boldsymbol{H}^\dagger\boldsymbol{H}/N_0$.

[3] By "mismatched" we mean that $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ may not be a valid pdf and in general differs from the true pdf $p(\boldsymbol{y}|\boldsymbol{x})$ even with $\hat{\boldsymbol{x}} = \boldsymbol{0}$, but such a "mismatched" property is for the purpose of reducing the size of trellis-description in the BCJR.

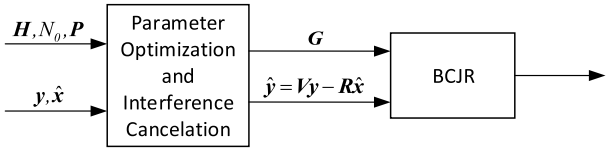


FIGURE 2. CS demodulator that maximizes the GMI with the tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$.

The front-end in this case is a matched filter [51] and the BCJR is implemented over the Ungerboeck model. To reach a trellis with $|\mathcal{X}|^{\nu}$ states, we truncate $\boldsymbol{G}$ to its center $2\nu+1$ diagonals, i.e., we use the following parameters in (4):

$$\boldsymbol{V}=\boldsymbol{H}^{\dagger}/N_0,\ \boldsymbol{R}=\boldsymbol{0},\quad \text{and } \boldsymbol{G}=[\boldsymbol{H}^{\dagger}\boldsymbol{H}/N_0]_{\nu}.$$

With these settings in TMF, the GMI in (8) equals

$$I_{\mathrm{GMI}} = \log\big(\det(\boldsymbol{I}+[\boldsymbol{H}^{\dagger}\boldsymbol{H}/N_0]_{\nu})\big) - \mathrm{Tr}\big(\boldsymbol{H}^{\dagger}\boldsymbol{H}\big(N_0\boldsymbol{I}+[\boldsymbol{H}^{\dagger}\boldsymbol{H}]_{\nu}\big)^{-1}[\boldsymbol{H}^{\dagger}\boldsymbol{H}]_{\setminus\nu}\big).$$

B. CONSTRAINTS ON THE PARAMETER R FOR CS DEMODULATORS

As illustrated in Fig. 2, our approach to design a CS demodulator consists of two steps:

- Construction of a signal $\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}}$ based on the received signal $\boldsymbol{y}$ and a prior mean $\hat{\boldsymbol{x}}$;
- BCJR demodulation of $\hat{\boldsymbol{y}}$ operating on a reduced number of states $|\mathcal{X}|^{\nu}$.

This procedure is analogous to LMMSE-PIC which first subtracts the interference followed by applying a Wiener filter, and then concludes by the BCJR that operates on a diagonal $\boldsymbol{G}$. The statistical behavior of $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ can be superior to that of the original $(\boldsymbol{x}, \boldsymbol{y})$, as the former tuple corresponds to a statistically different channel. As shown in Example 3 below, the GMI obtained with tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ with a perfect feedback $\hat{\boldsymbol{x}}$ can be infinitely large, which exceeds the channel capacity with the original tuple $(\boldsymbol{x}, \boldsymbol{y})$. Therefore, the computed value of GMI may have little relevance for the performance of the transceiver system. In order for GMI to have bearing on performance, it is critical to put constraints on IC matrix $\boldsymbol{R}$.

Example 3: Let the system model be

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}.$$

and assume a perfect feedback, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{x}$. The demodulator parameters are taken as $\boldsymbol{V} = \boldsymbol{0}$, $\boldsymbol{R} = -(1+\beta)\boldsymbol{I}$, and $\boldsymbol{G} = \beta\boldsymbol{I}$, where β is an arbitrary value. Then, we have

$$\hat{\boldsymbol{y}} = \boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}} = (1+\beta)\boldsymbol{x},$$

and the GMI in (8) for the tuple $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ equals

$$I_{\mathrm{GMI}}(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G}) = K\big(1+\log(1+\beta)\big).$$

In order to maximize the GMI, the demodulator will choose $\beta \to \infty$ to make I_{GMI} infinite. This is because, except for using the feedback information for IC, the demodulator uses the prior mean $\hat{\boldsymbol{x}}$ as a signal energy via $\boldsymbol{R}$. A demodulator equipped with these parameters will have significant error propagation and does not have much operational meaning for an iterative receiver. Thus, we conclude that unless constraints are put on $\boldsymbol{R}$, the GMI value is not relevant. Three typical shapes of $\boldsymbol{R}$ are specified in Fig. 3. All three have in common that rather than adding signal energy, the rationale of $\boldsymbol{R}$ should be to remove interference. Therefore, at the very minimum the diagonal elements of $\boldsymbol{R}$ should be constrained to zeros, so that the demodulation of each symbol in $\boldsymbol{x}$ does not rely on its own prior mean $\hat{\boldsymbol{x}}$. Such a constraint is perfectly aligned with the operations of LMMSE-PIC, where $\hat{x}_{\ell}$ is not used for demodulation of x_{ℓ}. Further, the rationale of the constraints we impose on $\boldsymbol{R}$ is to follow the principle of extrinsic information: The BCJR module should not rely on the prior information $\hat{x}_{\ell}$ when demodulating x_{ℓ} (this requires more than just the diagonal of $\boldsymbol{R}$ to be zero).

We point out the fact that the GMI can exceed the channel capacity is a consequence of our choice not to include the side-information as a prior distribution on $\boldsymbol{x}$ when evaluating the GMI. If we did, then the GMI is decaying with increasing quality of the side-information (due to the mutual information $I(\boldsymbol{x}, \boldsymbol{y}|\hat{\boldsymbol{x}})$ goes to 0 as $\hat{\boldsymbol{x}}$ becomes perfect). At last, we acknowledge the fact that a permutation of the columns of $\boldsymbol{H}$ can boost the performance of the CS demodulator whenever $0 < \nu < K-1$ for MIMO channels. However, minimum-phase conversions of ISI channels are not beneficial as we will solve for the optimal front-end filter.

IV. PARAMETER OPTIMIZATION FOR THE MIMO CASE

In this section, we elaborate the parameter optimization for MIMO channels. We introduce three different methods, namely, Method I, II, and III. We start with the classical Forney model (6) based demodulator, i.e., Method I, and then extend the model to the Ungerboeck model (4), i.e., Method II. As both Method I and Method II need gradient-based approaches for optimizations over target response, by carefully examining the properties of the CS demodulator with Ungerboeck model, we propose a suboptimal Method III which has an explicit construction and all CS parameters are in closed-forms.

A. METHOD I

In Method I, the CS demodulator is based on detection model (6) and the following structures of the CS parameters $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$ are imposed:

- The $K \times N$ matrix $\boldsymbol{W}$ has no constraints.
- The $K \times K$ matrix $\boldsymbol{F}$ is lower-triangular where only the main diagonal and the first ν (the memory-size of $\boldsymbol{F}$) lower diagonals are non-zero, i.e., $\boldsymbol{F}$ is banded within diagonals $[0, \nu]$ $(0 \le \nu < K-1)$. Moreover, the main diagonal of $\boldsymbol{F}$ is constrained to have positive values.
- The $K \times K$ matrix $\boldsymbol{T}$ is constrained to be zero wherever $\boldsymbol{F}$ can take non-zero values.

The constraint on $\boldsymbol{F}$ is to shorten the memory of the trellis in BCJR, while the purpose of the constraint on $\boldsymbol{T}$ is to cancel the signal part that $\boldsymbol{F}$ cannot handle. From Theorem 1 and by

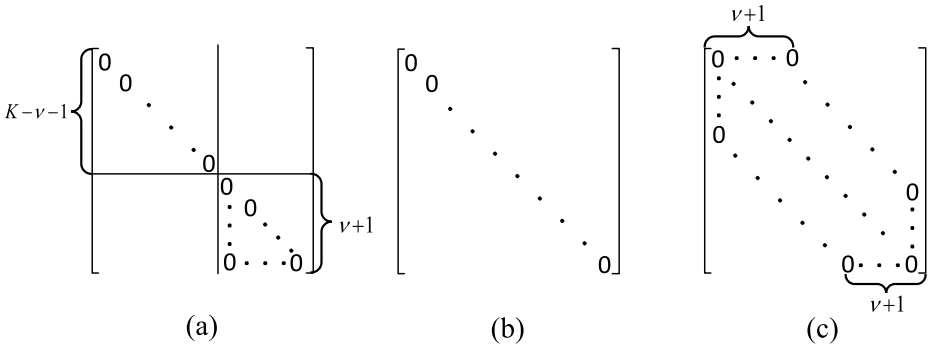


FIGURE 3. Three different types of shape of matrix R, where ν is the memory-size of F or G, i.e., the memory-size of the BCJR.

identifying $\boldsymbol{V} = \boldsymbol{F}^\dagger \boldsymbol{W}$, $\boldsymbol{R} = \boldsymbol{F}^\dagger \boldsymbol{T}$, and $\boldsymbol{G} = \boldsymbol{F}^\dagger \boldsymbol{F}$, the GMI in (8) of Method I equals

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F}) = {} & \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) - \mathrm{Tr}(\boldsymbol{F}^\dagger\boldsymbol{F}) \\ & + 2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{F}^\dagger(\boldsymbol{W}\boldsymbol{H}-\boldsymbol{T}\boldsymbol{P})\big)\big\} \\ & - \mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{L}_1\big) \end{aligned} \tag{9}$$

where

$$\begin{aligned} \boldsymbol{L}_1 = {} & \boldsymbol{F}^\dagger\boldsymbol{W}(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^\dagger)\boldsymbol{W}^\dagger\boldsymbol{F} - 2\mathcal{R}\big\{\boldsymbol{F}^\dagger\boldsymbol{W}\boldsymbol{H}\boldsymbol{P}\boldsymbol{T}^\dagger\boldsymbol{F}\big\} \\ & + \boldsymbol{F}^\dagger\boldsymbol{T}\boldsymbol{P}\boldsymbol{T}^\dagger\boldsymbol{F}. \end{aligned}$$

With the aforementioned constraints on $\boldsymbol{F}$ and $\boldsymbol{T}$, the matrix $\boldsymbol{R}=\boldsymbol{F}^\dagger\boldsymbol{T}$ has a form of shape (a) in Fig. 3. That is, all diagonal elements are zero as well as the lower-triangular part of the $(\nu+1)\times(\nu+1)$ small matrix at the right-bottom corner.

In order to optimize (9) over $(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$, we first introduce an $S\times K^2$ indication matrix $\boldsymbol{\Omega}$ only consisting of ones and zeros,[4] having a single 1 in each row, and S equals the number of elements in $\boldsymbol{T}$ that are allowed to be non-zero. Let $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ be a vector that contains the positions where the vector $\mathrm{vec}(\boldsymbol{T})$ is allowed to be non-zero. Then, the value of the kth entry in $\mathbb{I}(\mathrm{vec}(\boldsymbol{T}))$ gives the column where row k of $\boldsymbol{\Omega}$ is 1. That is, the $S\times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T})$ stacks the columns of $\boldsymbol{T}$ on top of each other but with all elements that are constrained to zero removed.

With such a definition of $\boldsymbol{\Omega}$ and defining two $K\times K$ matrices

$$\boldsymbol{M} = \boldsymbol{H}^\dagger(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^\dagger)^{-1}\boldsymbol{H}-\boldsymbol{I}, \tag{10}$$

$$\tilde{\boldsymbol{M}} = \boldsymbol{P}(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P}-\boldsymbol{P}, \tag{11}$$

the GMI for the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$ is given in Proposition 1 with the proof in Appendix B.

Proposition 1: Defining an $S\times K^2$ matrix

$$\boldsymbol{D} = \boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*)\otimes\boldsymbol{I}_K\big),$$

the optimal $\boldsymbol{W}$ maximizing the GMI in (9) is

$$\boldsymbol{W}_{\mathrm{opt}} = \boldsymbol{F}^{-\dagger}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F}+\boldsymbol{F}^\dagger\boldsymbol{T}\boldsymbol{P})\boldsymbol{H}^\dagger(N_0\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^\dagger)^{-1}, \tag{12}$$

and when $\boldsymbol{P}\neq\boldsymbol{0}$ the optimal $\boldsymbol{T}$ is given by

$$\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) = -\boldsymbol{\Omega}^T\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^T\big)^{-1}\mathrm{vec}(\boldsymbol{F}). \tag{13}$$

With the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the GMI reads,

$$I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F}) = \begin{cases} I_1(\boldsymbol{F}), & \boldsymbol{P}=\boldsymbol{0}, \\ I_1(\boldsymbol{F})+\delta_1(\boldsymbol{F}), & \boldsymbol{P}\neq\boldsymbol{0}, \end{cases} \tag{14}$$

where the functions $I_1(\boldsymbol{F})$ and $\delta_1(\boldsymbol{F})$ are defined as

$$I_1(\boldsymbol{F}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big), \tag{15}$$

$$\begin{aligned} \delta_1(\boldsymbol{F}) = {} & -\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big) \\ & \boldsymbol{\Omega}^T\Big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}). \end{aligned} \tag{16}$$

Remark 1: From the definitions in (10) and (11), $\boldsymbol{M}$ is the negative of MSE matrix and consequently, it holds that $\tilde{\boldsymbol{M}}\preceq 0$. Hence, $\delta_1(\boldsymbol{F})\geq 0$ represents the GMI increments from the soft-feedback.

Before maximization the GMI, we state Theorem 2 that deals with a general maximization problem.

Theorem 2: Define a scalar function I with respect to a $K\times K$ matrix $\boldsymbol{G}$ as

$$I(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big) \tag{17}$$

where $\boldsymbol{G}$ satisfies $\boldsymbol{G}=[\boldsymbol{G}]_\nu$. Then, the optimal $\boldsymbol{G}$ maximizing I is the unique solution that satisfies

$$[(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})^{-1}]_\nu = -[\boldsymbol{M}]_\nu. \tag{18}$$

With $\boldsymbol{G}_{\mathrm{opt}}$, the maximal I equals

$$I(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big). \tag{19}$$

Proof: Taking the first order differential of I with respect to $\boldsymbol{G}$ and noticing that $\boldsymbol{G}$ is banded within diagonals $[-\nu,\nu]$, yields (18) after some manipulations. The existence and uniqueness of such an optimal solution for (18) is proved in [52, Th. 2] and also illustrated in [18, Proposition 2]. By Lemma 1 and (18), it holds that $\mathrm{Tr}\big([\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}}]^{-1}\boldsymbol{M}\big)=-K$, and then (19) follows. ■

Optimizing over $\boldsymbol{F}$ in (14) when $\boldsymbol{P}\neq\boldsymbol{0}$ is difficult and cannot be carried out in closed-form. In Appendix C we show

[4] For instance, assuming $\boldsymbol{T}=\begin{bmatrix}0&1\\2&0\end{bmatrix}$, then the indication matrix $\boldsymbol{\Omega}=\begin{bmatrix}0&1&0&0\\0&0&1&0\end{bmatrix}$, and the vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T})=\begin{bmatrix}2\\1\end{bmatrix}$.

by an example that (14) is in general non-concave. Therefore, a gradient based numerical optimization procedure is utilized to search for the optimal $\boldsymbol{F}$. In the ith iteration, we construct

$$\boldsymbol{F}^{(i)} = \boldsymbol{F}^{(i-1)} + \nabla_{\boldsymbol{F}^*} I_{\mathrm{GMI}}\big(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F}^{(i-1)}\big)$$

where $\nabla_{\boldsymbol{F}^*} I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}}, \boldsymbol{T}_{\mathrm{opt}}, \boldsymbol{F})$ is the conjugate gradient of the GMI with respect to (the non-zero part of) $\boldsymbol{F}$, which is derived in Appendix D.

If replacing $\boldsymbol{F}^\dagger\boldsymbol{F}$ by $\boldsymbol{G}$, (15) has the same form as (17) when $\boldsymbol{P}=\boldsymbol{0}$, and $\boldsymbol{G}_{\mathrm{opt}}$ is in closed-form shown in Theorem 2. If $\boldsymbol{G}_{\mathrm{opt}} \succ \boldsymbol{0}$, then the optimal $\boldsymbol{F}$ is equal to its Cholesky decomposition. Whenever it is not, a gradient based numerical optimization is utilized to optimize (15), and $\boldsymbol{G}_{\mathrm{opt}}$ from Theorem 2 is used to initialize the starting point of $\boldsymbol{F}$, which has been observed to be reliable.

Next we establish a connection between the front-end filter $\boldsymbol{W}$ and IC matrix $\boldsymbol{T}$ in Method I.

Proposition 2: For $\boldsymbol{P} \neq \boldsymbol{0}$, and with the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$, the matrix $\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}})$ is banded within diagonals $[-\nu, K-1]$.

Proof: Noting that $\boldsymbol{\Omega}^T\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) = \mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}})$ and using $\boldsymbol{\Omega}\boldsymbol{\Omega}^T = \boldsymbol{I}$, from (13) and (76) it holds that

$$\begin{aligned}
&\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^* \otimes \big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^T\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}_{\mathrm{opt}}) \\
&\quad = \boldsymbol{\Omega}\mathrm{vec}\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}}\big) \\
&\qquad - \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{FMP}),
\end{aligned} \tag{20}$$

which shows that the elements of the matrix

$$\Delta = \boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}_{\mathrm{opt}}\tilde{\boldsymbol{M}} + \boldsymbol{FMP}$$

are zeros wherever $\boldsymbol{T}$ can be non-zero. Hence Δ is banded within diagonals $[0, \nu]$. On the other hand, with the optimal $\boldsymbol{W}$ in (12), and $\boldsymbol{M}$, $\tilde{\boldsymbol{M}}$ defined in (10) and (11), respectively, we have

$$\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}}) - (\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F}) = (\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\boldsymbol{F}^{-1}\Delta\boldsymbol{P}^{-1}. \tag{21}$$

By noting that $\boldsymbol{F}^{-1}$ is lower-triangular, $\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F}$ is banded within diagonals $[-\nu, \nu]$ and $\boldsymbol{P}$ is diagonal. Utilizing Lemma 2, the r.h.s of (21) is banded within diagonals $[-\nu, K-1]$. Therefore, $\boldsymbol{F}^\dagger(\boldsymbol{W}_{\mathrm{opt}}\boldsymbol{H} - \boldsymbol{T}_{\mathrm{opt}})$ is banded within diagonals $[-\nu, K-1]$. ∎

Proposition 2 reveals an interesting and somewhat surprising fact that although the BCJR only has a memory-size ν, the interference outside the memory-size shall not be perfectly canceled with the optimal CS demodulator in Method I. As will be shown later, such a property also holds for the other two designs of CS demodulator, namely, Method II and III.

B. METHOD II

Method II origins from Ungerboeck's 1974 paper [47]. Different from Method I, an Ungerboeck detection model (4) instead of the Forney model (6) is applied. The Ungerboeck model has been extensively discussed in [48], [49], and [53]. The system model (4) has the following constraints:

- The $K \times N$ matrix $\boldsymbol{V}$ has no constraints.
- The $K \times K$ matrix $\boldsymbol{G}$ is Hermitian and satisfies $\boldsymbol{G} = [\boldsymbol{G}]_\nu$ and $\boldsymbol{I}+\boldsymbol{G} \succ \boldsymbol{0}$, where ν is the memory-size of $\boldsymbol{G}$.
- The $K \times K$ matrix $\boldsymbol{R}$ can have specified shapes, but at a minimum the main diagonal is constrained to zeros.

Instead of $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$, in Method II we optimize $(\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{G})$ for (8). The same definition of an indication matrix $\boldsymbol{\Omega}$ is used as in Method I, but now $\boldsymbol{\Omega}$ corresponds to $\boldsymbol{R}$ instead of $\boldsymbol{T}$. We continue to let S denote the number of elements that are allowed to be non-zero in $\boldsymbol{R}$. That is, the $S \times 1$ vector $\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R})$ stacks the columns of $\boldsymbol{R}$ on top of each other and with all elements that are constrained to zeros removed.

In Method II, we have Proposition 3 showing the GMI with optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, with the proof given in Appendix E.

Proposition 3: Define an $S \times 1$ vector $\boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{MP})$, the optimal $\boldsymbol{V}$ for the GMI in (8) is

$$\boldsymbol{V}_{\mathrm{opt}} = (\boldsymbol{I}+\boldsymbol{G}+\boldsymbol{RP})\boldsymbol{H}^\dagger(\boldsymbol{HH}^\dagger + N_0\boldsymbol{I})^{-1}, \tag{22}$$

and when $\boldsymbol{P} \neq \boldsymbol{0}$ the optimal $\boldsymbol{R}$ is given by

$$\mathrm{vec}(\boldsymbol{R}_{\mathrm{opt}}) = -\boldsymbol{\Omega}^T\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^* \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T\big)^{-1}\boldsymbol{d}. \tag{23}$$

With the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, the GMI in (8) equals

$$I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G}) = \begin{cases} I_2(\boldsymbol{G}), & \boldsymbol{P}=\boldsymbol{0}, \\ I_2(\boldsymbol{G}) + \delta_2(\boldsymbol{G}), & \boldsymbol{P}\neq\boldsymbol{0}, \end{cases} \tag{24}$$

where the functions $I_2(\boldsymbol{G})$ and $\delta_2(\boldsymbol{G})$ are defined as

$$I_2(\boldsymbol{G}) = K + \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) + \mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big), \tag{25}$$

$$\delta_2(\boldsymbol{G}) = -\boldsymbol{d}^\dagger\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^* \otimes (\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T\big)^{-1}\boldsymbol{d}. \tag{26}$$

Similar to $\delta_1(\boldsymbol{F})$ in Method I, $\delta_2(\boldsymbol{G}) \geq 0$ represents the GMI increments from the feedback. Further, when $\boldsymbol{P} \neq \boldsymbol{0}$, the optimization over $\boldsymbol{G}$ in (24) also uses a gradient based numerical optimization, and the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}}, \boldsymbol{R}_{\mathrm{opt}}, \boldsymbol{G})$ with respect to (the non-zero part of) $\boldsymbol{G}$ is provided in Appendix F. The closed-from $\boldsymbol{G}$ from Theorem 2 can be used as the starting point. However, different from Method I, the optimization procedure in this case is concave whose proof is in Appendix G.

Although the optimal $\boldsymbol{R}$ is solved in closed-form in (23), we shall specify the constraint (reflected by $\boldsymbol{\Omega}$) on it. We consider two types of $\boldsymbol{R}$ in Method II. Firstly, as we are interested in the comparison between Method I and II, we consider the shape (a) in Fig. 3, which has the same shape as for $\boldsymbol{R} = \boldsymbol{F}^\dagger\boldsymbol{T}$ in Method I. Secondly, we consider a band-shaped $\boldsymbol{R}$ with memory-size ν_{R}, where shape (b) and (c) in Fig. 3 are typical cases with $\nu_{\mathrm{R}} = 0$ and $\nu_{\mathrm{R}} = \nu$, respectively. With shape (b), we only limit the diagonal elements of $\boldsymbol{R}$ to be zero and intend to eliminate the interference as much as possible. With shape (c), we limit $\boldsymbol{R}$ to have the opposite form of $\boldsymbol{G}$, that is, the elements of $\boldsymbol{R}$ are constrained to be zero wherever $\boldsymbol{G}$ is non-zero. The intention is to only cancel the interference that the BCJR represented by $\boldsymbol{G}$ cannot handle. Shape (c) is based on the same idea as Method I, but now operates on the Ungerboeck model.

The connection between the optimal front-end filter $\boldsymbol{V}$ and IC matrix $\boldsymbol{R}$ in Method II is established in Proposition 4.

Proposition 4: For $\boldsymbol{P}\neq\boldsymbol{0}$ and the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$,

$$[\boldsymbol{V}_{\rm opt}\boldsymbol{H}]_{\setminus(\nu+\nu_{\rm R})}=[\boldsymbol{R}_{\rm opt}]_{\setminus(\nu+\nu_{\rm R})}. \tag{27}$$

That is, the elements of $\boldsymbol{V}_{\rm opt}\boldsymbol{H}$ and $\boldsymbol{R}_{\rm opt}$ are equal outside the center $2(\nu+\nu_{\rm R})+1$ diagonals for any $\boldsymbol{G}$ that is banded within diagonals $[-\nu,\nu]$, where $\nu_{\rm R}=0$ for $\boldsymbol{R}$ with both shapes (a) and (b), and $\nu_{\rm R}=\nu$ for $\boldsymbol{R}$ with shape (c), respectively.

Proof: Following the similar steps in the proof of Proposition 2, (23) can be rewritten as

$$\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}_{\rm opt}\tilde{\boldsymbol{M}}\big)=-\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}). \tag{28}$$

It shows that the elements of

$$\Delta=(\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}_{\rm opt}\tilde{\boldsymbol{M}}\boldsymbol{P}^{-1}+\boldsymbol{M}$$

are zeros wherever $\boldsymbol{R}$ can be non-zero. On the other hand, with the optimal $\boldsymbol{V}$ in (22) we have

$$\boldsymbol{V}_{\rm opt}\boldsymbol{H}-\boldsymbol{R}_{\rm opt}-(\boldsymbol{I}+\boldsymbol{G})=(\boldsymbol{I}+\boldsymbol{G})\Delta. \tag{29}$$

As $\boldsymbol{I}+\boldsymbol{G}$ is banded within diagonals $[-\nu,\nu]$, using Lemma 2 and with the three shapes[5] of $\boldsymbol{R}$ in Fig. 3, it can be shown that the r.h.s of (29) is banded within diagonals $[-(\nu+\nu_{\rm R}),\nu+\nu_{\rm R}]$, where $\nu_{\rm R}=0$ for the shape (a) and (b), and $\nu_{\rm R}=\nu$ for the shape (c). Therefore, $\boldsymbol{V}_{\rm opt}\boldsymbol{H}-\boldsymbol{R}_{\rm opt}$ on the l.h.s of (29) is banded within diagonals $[-(\nu+\nu_{\rm R}),\nu+\nu_{\rm R}]$. ■

The same as Proposition 2 for Method I, Proposition 4 shows that the signal part not considered in $\boldsymbol{G}$ (the BCJR) shall not be perfectly canceled inside the center $2(\nu+\nu_{\rm R})+1$ diagonals, instead of the center $2\nu+1$ diagonals where $\boldsymbol{G}$ is constrained to be non-zero. With LMMSE-PIC, we have $\nu=\nu_{\rm R}=0$ and Proposition 4 is natural and frequently used. However, when $\nu_{\rm R}>0$, a more general property is now revealed that $\boldsymbol{V}_{\rm opt}\boldsymbol{H}$ and $\boldsymbol{R}$ are only equal outside the center $2(\nu+\nu_{\rm R})+1$ diagonals.

C. METHOD III

So far we have discussed two types of CS demodulators based on Forney and Ungerboeck detection models. One disadvantage is that both methods need numerical optimizations to obtain the optimal target responses. Next we construct a third CS demodulator that has closed-form solutions for all CS parameters, whose GMI is slightly suboptimal.

Method III relies on the same operations as Method II with $\boldsymbol{P}=\boldsymbol{0}$. By inserting $\boldsymbol{V}_{\rm opt}$ in (22) back into (4) and setting $\boldsymbol{R}=\boldsymbol{0}$, the CS demodulator actually operates on the mismatched function

$$\begin{aligned}\tilde{p}(\boldsymbol{y}|\boldsymbol{x})&=\exp\big(2\mathcal{R}\{\boldsymbol{x}^\dagger\boldsymbol{V}_{\rm opt}\boldsymbol{y}\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big)\\&=\exp\big(2\mathcal{R}\big\{\boldsymbol{x}^\dagger(\boldsymbol{I}+\boldsymbol{G})\check{\boldsymbol{x}}\big\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big)\end{aligned} \tag{30}$$

where

$$\check{\boldsymbol{x}}=\boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{H}^\dagger+N_0\boldsymbol{I})^{-1}\boldsymbol{y}$$

is the LMMSE estimate of $\boldsymbol{x}$. As can be seen from (30), the BCJR is based on $\check{\boldsymbol{x}}$. With soft-feedback we can therefore replace $\check{\boldsymbol{x}}$ by LMMSE-PIC estimates $\tilde{\boldsymbol{x}}$. That is, instead of (30) we now operate on

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})=\exp\big(2\mathcal{R}\big\{\boldsymbol{x}^\dagger(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{x}}\big\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big) \tag{31}$$

where $\boldsymbol{G}$ has the same banded-shape as the first two methods, but optimized according to $\tilde{\boldsymbol{x}}$.

The estimate $\tilde{\boldsymbol{x}}$ is constructed as follows. As we prefer to handle the interference through the trellis-search process, the IC should not be present within the memory-size ν. In other words, the signal vector after the IC that is used to form the kth symbol of $\tilde{\boldsymbol{x}}$ is denoted as

$$\tilde{\boldsymbol{y}}_k=\boldsymbol{y}-\sum_{n\in\mathcal{A}_k}\boldsymbol{h}_n\hat{x}_n \tag{32}$$

where

$$\mathcal{A}_k=\big\{0\le n\le K-1:n\notin[\max(0,k-\nu),\min(k+\nu,K-1)]\big\}.$$

Denoting p_n as the nth diagonal element of $\boldsymbol{P}$, the Wiener-filter coefficients [54] for the kth symbol are calculated as

$$\hat{\boldsymbol{w}}_k=\boldsymbol{h}_k^\dagger(\boldsymbol{H}^\dagger\boldsymbol{C}_k\boldsymbol{H}+N_0\boldsymbol{I}\big)^{-1} \tag{33}$$

where $\boldsymbol{C}_k$ is a diagonal whose nth diagonal element is

$$C_k(n)=\begin{cases}1-p_n, & k\in\mathcal{A}_k,\\ 1, & \text{otherwise.}\end{cases} \tag{34}$$

The estimate $\tilde{\boldsymbol{x}}$ is then obtained through

$$\begin{aligned}\tilde{\boldsymbol{x}}&=\big[\hat{\boldsymbol{w}}_1\tilde{\boldsymbol{y}}_1,\ \hat{\boldsymbol{w}}_2\tilde{\boldsymbol{y}}_2,\ \cdots,\ \hat{\boldsymbol{w}}_K\tilde{\boldsymbol{y}}_K\big]^T\\&=\hat{\boldsymbol{W}}\boldsymbol{y}-\hat{\boldsymbol{C}}\hat{\boldsymbol{x}}\end{aligned} \tag{35}$$

where the coefficient matrix $\hat{\boldsymbol{W}}$ and IC matrix $\hat{\boldsymbol{C}}$ defined as

$$\hat{\boldsymbol{W}}=[\,\hat{\boldsymbol{w}}_1^T,\ \hat{\boldsymbol{w}}_2^T,\ \cdots,\ \hat{\boldsymbol{w}}_K^T\,]^T, \tag{36}$$

$$\hat{\boldsymbol{C}}=[\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus\nu}. \tag{37}$$

Inserting $\tilde{\boldsymbol{x}}$ in (35) back into (31), the detection model we operate on is

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x})=\exp\big(2\mathcal{R}\big\{\boldsymbol{x}^\dagger\big((\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{y}-(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}\hat{\boldsymbol{x}}\big)\big\}-\boldsymbol{x}^\dagger\boldsymbol{G}\boldsymbol{x}\big). \tag{38}$$

Note that (38) is a also special case of (4), by identifying

$$\begin{aligned}\boldsymbol{V}&=(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{W}},\\ \boldsymbol{R}&=(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}.\end{aligned}$$

The GMI in (8) in this case reads, after some manipulations,

$$I_{\rm GMI}(\boldsymbol{G})=K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\hat{\boldsymbol{M}}(\boldsymbol{I}+\boldsymbol{G})\big), \tag{39}$$

with $\hat{\boldsymbol{M}}$ (the updated $\boldsymbol{M}$ in Method II) defined as[6]

$$\begin{aligned}\hat{\boldsymbol{M}}&=\hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger+\hat{\boldsymbol{W}}\boldsymbol{H}-\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger\\&\quad\big(\hat{\boldsymbol{W}}\boldsymbol{H}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger+\hat{\boldsymbol{W}}\boldsymbol{H}-\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger\big)^\dagger\\&\quad-\hat{\boldsymbol{W}}(\boldsymbol{H}\boldsymbol{H}^\dagger+N_0\boldsymbol{I})\hat{\boldsymbol{W}}^\dagger-\hat{\boldsymbol{C}}\boldsymbol{P}\hat{\boldsymbol{C}}^\dagger-\boldsymbol{I},\end{aligned} \tag{40}$$

[5]The case that $\boldsymbol{R}$ with shape (a) is slightly different, but it can be verified straightforwardly.

[6]It can also be shown that $\hat{\boldsymbol{M}}$ is the negative of the updated MSE matrix, i.e., $\hat{\boldsymbol{M}}=-\mathbb{E}\big[(\boldsymbol{x}-\tilde{\boldsymbol{x}})(\boldsymbol{x}-\tilde{\boldsymbol{x}})^\dagger\big]=-\mathbb{E}\big[(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})(\boldsymbol{x}-\hat{\boldsymbol{W}}\boldsymbol{y}+\hat{\boldsymbol{C}}\hat{\boldsymbol{x}})^\dagger\big]$.

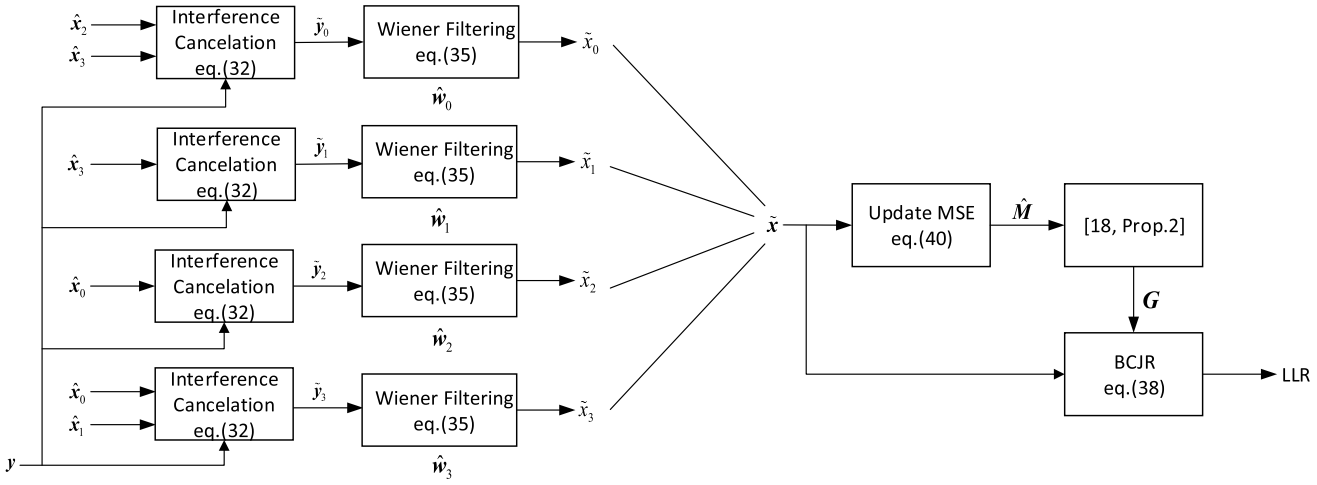


FIGURE 4. An graphical overview of Method III with $K=4$ and $\nu=1$.

The optimal $\boldsymbol{G}$ for (39) is then obtained from Theorem 2, and the optimal GMI is

$$I_{\mathrm{GMI}}(\boldsymbol{G}_{\mathrm{opt}}) = \log\big(\det(\boldsymbol{I}+\boldsymbol{G}_{\mathrm{opt}})\big).$$

An graphical overview of Method III for $K=4$ and $\nu=1$ is illustrated in Fig. 4. For any $\boldsymbol{G}$ with memory-size ν, the IC matrix $(\boldsymbol{I}+\boldsymbol{G})\tilde{\boldsymbol{C}}$ is zero along the main diagonal, which guarantees that the extrinsic information will not be used for current symbols in the IC process. In GMI sense, Method III will not outperform Method II with $\boldsymbol{R}$ in shape (b), but it can outperform the GMI of Method II with $\boldsymbol{R}$ in shape (c), as it can be verified that $\boldsymbol{R}$ in shape (c) has zeros at positions where $(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}$ are also zeros.

Remark 2: Since $\hat{\boldsymbol{W}}\boldsymbol{H}-\hat{\boldsymbol{C}}=[\hat{\boldsymbol{W}}\boldsymbol{H}]_{\nu}$ and by Lemma 2, $(\boldsymbol{I}+\boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H}-\hat{\boldsymbol{C}})$ is banded within diagonals $[-2\nu, 2\nu]$. Therefore, it holds that $[(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus 2\nu}=[(\boldsymbol{I}+\boldsymbol{G})\hat{\boldsymbol{C}}]_{\setminus 2\nu}$, and Proposition 4 is also true for Method III where $\nu_{\mathrm{R}}=\nu$.

V. PARAMETER OPTIMIZATION FOR THE ISI CASE

In this section, we extend the CS demodulators to the ISI case with the block length K being sufficiently large and $\boldsymbol{P}=\alpha\boldsymbol{I}$. The formulas for the GMI in (8), (9), and (39) can be directly applied to (1), but as the rate I_{GMI} depends on K, we consider an asymptotic rate

$$\bar{I} = \lim_{K\to\infty}\frac{1}{K}I_{\mathrm{GMI}}.$$

Ideally, in the ISI case the front-end matrix $\boldsymbol{V}$ and IC matrix $\boldsymbol{R}$ corresponding to linear filtering operations with infinite taps, but in practice they have finite tap-lengths. Hence, we analyze the properties of $\boldsymbol{V}$, $\boldsymbol{R}$ (and also $\boldsymbol{W}$, $\boldsymbol{T}$) with finite numbers of taps and approximate them by band-shaped Toeplitz matrices. Further, the trellis-representation matrix $\boldsymbol{G}$ (and $\boldsymbol{F}$), and channel matrix $\boldsymbol{H}$ are also band-shaped Toeplitz matrices. Therefore, in the ISI case all matrices we consider are assumed to be band-shaped Toeplitz matrices, whose dimensions are sufficiently large such that the asymptotic properties can be analyzed. In [55] a complete theoretic machinery for ISI channels is derived and a result is that as $K\to\infty$ the linear convolution in (1) can be replaced by a circular convolution.

In the following, we denote the Fourier series introduced by a band-shaped Toeplitz matrix $\boldsymbol{E}$ with infinitely large dimensions as $E(\omega)$, where $\boldsymbol{E}$ is constrained to zero outside the middle $2N_{\mathrm{E}}+1$ diagonals, and N_{E} is referred to as the tap-length of $E(\omega)$. The Fourier series $E(\omega)$ is defined as

$$E(\omega)=\sum_{k=-N_{\mathrm{E}}}^{N_{\mathrm{E}}} e_k\exp(jk\omega)$$

and specified by a vector

$$\boldsymbol{e}=[e_{-N_{\mathrm{E}}},\ \ldots,\ e_{-1},\ e_0,\ e_1,\ \ldots,\ e_{N_{\mathrm{E}}}]$$

where e_0 is the element on the main diagonal and e_k is the element on kth lower ($k>0$) or upper ($k<0$) diagonal. As all quantities are evaluated as K grows large, $E(\omega)$ approaches the eigenvalue distribution of $\boldsymbol{E}$ (see [40], [56] for a precise statement of this result).

We first state Theorem 3, which is an asymptotic version of Theorem 2 for ISI channels.

Theorem 3: Assume two band-shaped Toeplitz matrices $\boldsymbol{G}$ and $\boldsymbol{M}$ with infinitely large dimensions satisfying the conditions: $[\boldsymbol{G}]_{\setminus\nu}=\boldsymbol{0}$, $\boldsymbol{I}+\boldsymbol{G}\succ\boldsymbol{0}$, and $\boldsymbol{M}\prec 0$, and define a scalar function

$$\bar{I}=1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(\log(1+G(\omega))+M(\omega)(1+G(\omega))\big)\mathrm{d}\omega. \quad (41)$$

Then, the optimal $G(\omega)$ that maximizes $\bar{I}$ is

$$G_{\mathrm{opt}}(\omega)=|u_0+\hat{\boldsymbol{u}}\varphi(\omega)|^2-1,$$

where the $1\times\nu$ vector

$$\varphi(\omega)=[\exp(j\omega),\ \exp(j2\omega),\ \ldots,\ \exp(j\nu\omega)]^T,$$

and

$$u_0 = \frac{1}{\sqrt{\boldsymbol{\tau}_1^\dagger \boldsymbol{\tau}_2^{-1} \boldsymbol{\tau}_1 - \tau_0}},$$
$$\hat{\boldsymbol{u}} = -u_0 \boldsymbol{\tau}_1^\dagger \boldsymbol{\tau}_2^{-1}. \quad (42)$$

The real scalar τ_0, $\nu \times 1$ vector $\boldsymbol{\tau}_1$, and $\nu \times \nu$ matrix $\boldsymbol{\tau}_2$ are defined as

$$\tau_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\mathrm{d}\omega,$$
$$\boldsymbol{\tau}_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\varphi(\omega)\mathrm{d}\omega,$$
$$\boldsymbol{\tau}_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\varphi(\omega)\varphi(\omega)^\dagger \mathrm{d}\omega,$$

respectively. Further, with $G_{\mathrm{opt}}(\omega)$ the optimal $\bar{I}$ is

$$\bar{I} = 2\log(u_0). \quad (43)$$

Proof: As $\boldsymbol{I}+\boldsymbol{G} \succ \boldsymbol{0}$, we have $1+G(\omega)=|U(\omega)|^2$ where $U(\omega)=u_0+\hat{\boldsymbol{u}}\varphi(\omega)$ and $\hat{\boldsymbol{u}}=[u_1,\ u_2,\ \ldots,\ u_\nu]$. Then, $\bar{I}$ in (41) can be written as

$$\bar{I} = 1+2\log(u_0)+\frac{1}{2\pi}\int_{-\pi}^{\pi} M(\omega)\big(u_0^2+2\mathcal{R}\{u_0\hat{\boldsymbol{u}}\varphi(\omega)\} + \hat{\boldsymbol{u}}\varphi(\omega)\varphi^\dagger(\omega)\hat{\boldsymbol{u}}^\dagger\big)\mathrm{d}\omega. \quad (44)$$

Taking the first order derivatives with respect to both u_0 and $\hat{\boldsymbol{u}}$ and optimizing them directly yields the optimal solutions in (42). Inserting (42) back into (44) and after some manipulations, the optimal asymptotic rate is in (43). ∎

A. METHOD I

The structures of $(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F})$ are the same as in Sec. IV-A, except that now the matrices have infinite dimensions. Applying Szegö's eigenvalue distribution theorem [56] to (9), the asymptotic rate equals

$$\begin{aligned}\bar{I}\big(W(\omega), T(\omega), F(\omega)\big) &= \lim_{K\to\infty}\frac{1}{K} I_{\mathrm{GMI}}(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{F}) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log\big(1+|F(\omega)|^2\big)-|F(\omega)|^2-\frac{L_1(\omega)}{1+|F(\omega)|^2}\right)\mathrm{d}\omega \\ &\quad +\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)\big(W(\omega)H(\omega)-\alpha T(\omega)\big)\big\}\mathrm{d}\omega \quad (45)\end{aligned}$$

where

$$L_1(\omega) = |F(\omega)W(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|F(\omega)T(\omega)|^2 - 2\alpha|F(\omega)|^2\mathcal{R}\big\{H(\omega)W(\omega)T^*(\omega)\big\}.$$

Note that, the Fourier series associated to $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ in (10) and (11) are

$$M(\omega) = \frac{|H(\omega)|^2}{N_0+|H(\omega)|^2}-1, \quad (46)$$
$$\tilde{M}(\omega) = \alpha^2(M(\omega)+1)-\alpha. \quad (47)$$

Further, define a $(2N_{\mathrm{T}}-\nu)\times 1$ vector

$$\phi(\omega) = \big[\exp\big(-jN_{\mathrm{T}}\omega\big),\ \ldots,\ \exp\big(-j(\nu+1)\omega\big), \exp\big(j(\nu+1)\omega\big),\ \ldots,\ \exp\big(jN_{\mathrm{T}}\omega\big)\big]^T, \quad (48)$$

a $(2N_{\mathrm{T}}-\nu)\times 1$ vector $\boldsymbol{\varepsilon}_1$, and a $(2N_{\mathrm{T}}-\nu)\times(2N_{\mathrm{T}}-\nu)$ Hermitian matrix $\boldsymbol{\varepsilon}_2$ as

$$\boldsymbol{\varepsilon}_1 = \frac{\alpha}{2\pi}\int_{-\pi}^{\pi} M(\omega)F^*(\omega)\phi(\omega)\mathrm{d}\omega,$$
$$\boldsymbol{\varepsilon}_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2\phi(\omega)\phi(\omega)^\dagger}{1+|F(\omega)|^2}\mathrm{d}\omega, \quad (49)$$

respectively, where N_{T} is the tap-length of $T(\omega)$, and $\nu+1$ is the band-size that matrix $\boldsymbol{T}$ is constrained to zero. Then, we have Proposition 5 with its proof[7] given in Appendix H.

Proposition 5: The optimal $W(\omega)$ for the asymptotic rate in (45) is

$$W_{\mathrm{opt}}(\omega) = \frac{H^*(\omega)}{F^*(\omega)(N_0+|H(\omega)|^2)}\big(1+|F(\omega)|^2 + \alpha F^*(\omega)T_{\mathrm{opt}}(\omega)\big), \quad (50)$$

and when $0<\alpha\le 1$, the optimal $T(\omega)$ reads

$$T_{\mathrm{opt}}(\omega) = -\boldsymbol{\varepsilon}_1^\dagger\boldsymbol{\varepsilon}_2^{-1}\phi(\omega). \quad (51)$$

With the optimal $W(\omega)$ and $T(\omega)$, the asymptotic rate equals

$$\bar{I}\big(W_{\mathrm{opt}}(\omega), T_{\mathrm{opt}}(\omega), F(\omega)\big) = \begin{cases}\bar{I}_1(F(\omega)), & \alpha = 0,\\ \bar{I}_1(F(\omega))+\bar{\delta}_1(F(\omega)), & 0<\alpha\le 1.\end{cases} \quad (52)$$

The functions $\bar{I}_1(F(\omega))$ and $\bar{\delta}_1(F(\omega))$ are defined as,[8]

$$\bar{I}_1(F(\omega)) = 1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big) + M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega, \quad (53)$$

$$\bar{\delta}_1(F(\omega)) = -\boldsymbol{\varepsilon}_1^\dagger\boldsymbol{\varepsilon}_2^{-1}\boldsymbol{\varepsilon}_1. \quad (54)$$

In ISI case, Method I is not concave and an example is also provided in Appendix C. Hence, a gradient based optimization is used to optimize $F(\omega)$ with the optimal $G_{\mathrm{opt}}(\omega)$ from Theorem 3 utilized for initialization. The connection between the optimal front-end filter $W(\omega)$ and the IC filter $T(\omega)$ in Proposition 2 also holds for ISI channels. An asymptotic version of Proposition 2 is stated in Proposition 6.

Proposition 6: When $0<\alpha\le 1$, $a_k = b_k$ holds for $k<-(\nu+1)$, where

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^*(\omega)W_{\mathrm{opt}}(\omega)H(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega,$$
$$b_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F^*(\omega)T_{\mathrm{opt}}(\omega)\exp\big(-jk\omega\big)\mathrm{d}\omega.$$

[7] Note that, Proposition 5 is the same as [33, Th. 1] that has been derived for hard feedback symbols.

[8] Similar to MIMO channels, $\bar{\delta}_1(F(\omega))$ in (54) is only defined for $\alpha\neq 0$ which represents the rate increment with soft-information. The same holds for Method II.

Proof: As shown in Appendix H, the optimal $\tilde{\boldsymbol{t}}$ in (90) satisfies

$$\tilde{\boldsymbol{t}}_{\rm opt}\boldsymbol{\varepsilon}_2 = -\boldsymbol{\varepsilon}_1^{\dagger}.$$

With the definitions of $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ in (49), this is equivalent to

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|F(\omega)|^2T_{\rm opt}(\omega)\phi(\omega)^{\dagger}}{1+|F(\omega)|^2}{\rm d}\omega = -\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}F(\omega)M(\omega)\phi(\omega)^{\dagger}{\rm d}\omega. \quad (55)$$

On the other hand, with $W_{\rm opt}$ in (50) and $M(\omega)$, $\tilde{M}(\omega)$ defined in (46) and (47), we have

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\pi}^{\pi}(F^*(\omega)W_{\rm opt}(\omega)H(\omega)-F^*(\omega)T_{\rm opt}\\&\quad-\big(1+|F(\omega)|^2\big)\big)\exp\big(-jk\omega\big){\rm d}\omega\\&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\frac{\tilde{M}(\omega)F^*(\omega)T_{\rm opt}(\omega)}{\alpha}\\&\quad+\big(1+|F(\omega)|^2\big)M(\omega)\Big)\exp\big(-jk\omega\big){\rm d}\omega. \quad (56)\end{aligned}$$

Transforming (55) and (56) back into matrix forms, we have that (20) and (21) hold. Following the same arguments as in the proof of Proposition 2, $\boldsymbol{F}^{\dagger}(\boldsymbol{W}_{\rm opt}\boldsymbol{H}-\boldsymbol{R}_{\rm opt})$ is banded within diagonals $[-\nu, K-1]$. Therefore, we have

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(F^*(\omega)W_{\rm opt}(\omega)H(\omega)-F^*(\omega)T_{\rm opt}(\omega)\big)\exp\big(-jk\omega\big){\rm d}\omega=0$$

whenever $k<-(\nu+1)$, which proves Proposition 6. ■

B. METHOD II

The matrices $(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ have the same constraints as in Sec. IV-B while the dimensions of these matrices are infinitely large. However, as the shape (a) of $\boldsymbol{R}$ in Fig. 3 is not meaningful when $N, K\to\infty$, it is not considered for ISI case. Applying Szegö's eigenvalue distribution theorem to (8), the asymptotic rate of Method II reads

$$\begin{aligned}&\bar{I}(V(\omega),R(\omega),G(\omega))\\&=\lim_{K\to\infty}\frac{1}{K}I_{\rm GMI}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})\\&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big)-G(\omega)-\frac{L_2(\omega)}{1+G(\omega)}\Big){\rm d}\omega\\&\quad+\frac{1}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{\big(V(\omega)H(\omega)-\alpha R(\omega)\big)\big\}{\rm d}\omega \quad (57)\end{aligned}$$

where

$$L_2(\omega)=|V(\omega)|^2\big(N_0+|H(\omega)|^2\big)+\alpha|R(\omega)|^2-2\alpha\mathcal{R}\big\{H(\omega)V(\omega)R^*(\omega)\big\}. \quad (58)$$

Define a $2(N_{\rm R}-\nu_{\rm R})\times 1$ vector

$$\psi(\omega)=\big[\exp\big(-jN_{\rm R}\omega\big),\ \ldots,\ \exp\big(-j(\nu_{\rm R}+1)\omega\big),\ \exp\big(j(\nu_{\rm R}+1)\omega\big),\ \ldots,\ \exp\big(jN_{\rm R}\omega\big)\big]^T, \quad (59)$$

a $2(N_{\rm R}-\nu_{\rm R})\times 1$ vector $\boldsymbol{\zeta}_1$, and a $2(N_{\rm R}-\nu_{\rm R})\times 2(N_{\rm R}-\nu_{\rm R})$ Hermitian matrix $\boldsymbol{\zeta}_2$ as

$$\begin{aligned}\boldsymbol{\zeta}_1&=\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}M(\omega)\psi(\omega){\rm d}\omega,\\\boldsymbol{\zeta}_2&=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)\psi(\omega)\psi(\omega)^{\dagger}}{1+G(\omega)}{\rm d}\omega, \quad (60)\end{aligned}$$

where $N_{\rm R}$ denotes the tap-length of $R_{\rm opt}(\omega)$, and $2\nu_{\rm R}+1$ is the band-size that $\boldsymbol{R}$ is constrained to zero, and $M(\omega)$ and $\tilde{M}(\omega)$ are in (46) and (47), respectively. Then, we have Proposition 7 with the proof in Appendix I, where we also show that $R(\omega)$ is real and correspondingly, matrix $\boldsymbol{R}$ has Hermitian symmetry.

Proposition 7: The optimal $V(\omega)$ for (57) is

$$V_{\rm opt}(\omega)=\frac{H^*(\omega)}{N_0+|H(\omega)|^2}\big(1+G(\omega)+\alpha R_{\rm opt}(\omega)\big), \quad (61)$$

and when $0<\alpha\le 1$ the optimal $R(\omega)$ reads

$$R_{\rm opt}(\omega)=-\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}\psi(\omega). \quad (62)$$

With the optimal $V(\omega)$ and $R(\omega)$, the asymptotic rate equals

$$\bar{I}\big(V_{\rm opt}(\omega),R_{\rm opt}(\omega),G(\omega)\big)=\begin{cases}\bar{I}_2(G(\omega)), & \alpha=0,\\ \bar{I}_2(G(\omega))+\bar{\delta}_2(G(\omega)), & 0<\alpha\le 1.\end{cases} \quad (63)$$

The functions $\bar{I}_1(G(\omega))$ and $\bar{\delta}_2(G(\omega))$ are defined as

$$\bar{I}_2(G(\omega))=1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big)+M(\omega)\big(1+G(\omega)\big)\Big){\rm d}\omega, \quad (64)$$

$$\bar{\delta}_2(G(\omega))=-\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}\boldsymbol{\zeta}_1. \quad (65)$$

When $0<\alpha\le 1$, it still needs a gradient based optimization to find the optimal $G(\omega)$ for (64), and the closed-form solution in Theorem 3 is utilized as the starting point. Similar to the MIMO case, the asymptotic rate $\bar{I}(V_{\rm opt}(\omega),R_{\rm opt}(\omega),G(\omega))$ is concave with respect to $G(\omega)$, which is shown in Appendix J.

Proposition 8: When $0<\alpha\le 1$ it holds that $a_k=b_k$ for $|k|>\nu+\nu_{\rm R}$, where

$$a_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}V_{\rm opt}(\omega)H(\omega)\exp\big(-jk\omega\big){\rm d}\omega, \quad (66)$$

$$b_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}R_{\rm opt}(\omega)\exp\big(-jk\omega\big){\rm d}\omega. \quad (67)$$

The connection of the optimal $V(\omega)$ and $R(\omega)$ stated in Proposition 8 is an asymptotic version of Proposition 4, and the proof follows the similar approach as Proposition 6 which is omitted. We show an example in Fig. 5 to illustrate Proposition 8 with Method II and $\nu=\nu_{\rm R}=1$. The Proakis-C [57] channel is tested at an SNR of 10 dB and α equals 0.1, 0.4 and 0.8, respectively. As $\nu_{\rm R}=1$, b_k as defined in (67) is constrained to zero for $0\le k\le 1$. As can be seen, a_k as defined in (66) equals b_k only for $|k|>2$, and when $|k|=2$, a_k and b_k are not identical. This shows that with the optimal $V(\omega)$ and $R(\omega)$, the signal part along the second upper and lower diagonals that is not considered in $G(\omega)$ shall not be

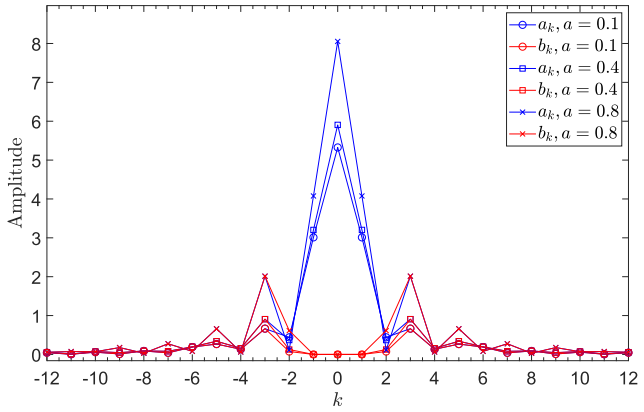


FIGURE 5. Comparison between a_k and b_k for Method II under Proakis-C channel $h=[0.227\ 0.46\ 0.688\ 0.46\ 0.227]$.

perfectly canceled out. This behavior cannot be seen in [32] which treats LMMSE-PIC for ISI channels due to $\nu=\nu_{\rm R}=0$.

C. METHOD III

In Method III, from (39) the asymptotic rate reads

$$\begin{aligned}\bar{I}(G(\omega)) &= \lim_{K\to\infty}\frac{1}{K}I_{\rm GMI}(\boldsymbol{G})\\ &= 1+\frac{1}{2\pi}\int_{-\pi}^{\pi}\big(\log\big(1+G(\omega)\big)+\hat{M}(\omega)(1+G(\omega))\big)\mathrm{d}\omega\end{aligned} \tag{68}$$

where according to (40) it holds that

$$\begin{aligned}\hat{M}(\omega) = 2\mathcal{R}\big\{&\alpha\hat{W}(\omega)H(\omega)\hat{C}^*(\omega)+\hat{W}(\omega)H(\omega)-\alpha\hat{C}^*(\omega)\big\}\\ &-\frac{|\hat{W}(\omega)|^2}{N_0+|H(\omega)|^2}-\alpha|\hat{C}(\omega)|^2-1.\end{aligned}$$

Replacing $M(\omega)$ by $\hat{M}(\omega)$, the optimal $G(\omega)$ and asymptotic rate $\bar{I}$ follow from Theorem 3.

Remark 3: Proposition 8 also holds for Method III with $\nu_{\rm R}=\nu$, due to the fact that $[(\boldsymbol{I}+\boldsymbol{G})(\hat{\boldsymbol{W}}\boldsymbol{H}-\hat{\boldsymbol{C}})]_{\backslash 2\nu}=\boldsymbol{0}$.

VI. SNR ASYMPTOTICS

In this section, we analyze asymptotic properties of the elaborated CS demodulators, and show that as N_0 goes to 0 and ∞, Method II and III are asymptotically equivalent. As Method I is inferior to Method II in GMI sense, we limit the analysis to Method II and III, and start with the analysis for the MIMO case.

The following limits can be verified straightforwardly:

$$\begin{aligned}\lim_{N_0\to 0}\boldsymbol{M}/N_0 &= -(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1},\\ \lim_{N_0\to\infty}N_0(\boldsymbol{I}+\boldsymbol{M}) &= \boldsymbol{H}^\dagger\boldsymbol{H}.\end{aligned} \tag{69}$$

Further, it also holds that

$$\begin{aligned}\lim_{N_0\to 0}\tilde{\boldsymbol{M}} &= \boldsymbol{P}^2-\boldsymbol{P},\\ \lim_{N_0\to\infty}\tilde{\boldsymbol{M}} &= -\boldsymbol{P}.\end{aligned} \tag{70}$$

As $\tilde{\boldsymbol{M}}$ should be invertible from the definition of $\delta_2(\boldsymbol{G})$ in (26), we restrict that $\boldsymbol{P}\prec\boldsymbol{I}$.

Lemma 3: When $N_0\to 0$ and ∞, the optimal $\boldsymbol{G}$ for (24) in Method II satisfies (18), and the following limits hold,

$$\lim_{N_0\to 0}\Big[(N_0(\boldsymbol{I}+\boldsymbol{G}_{\rm opt}))^{-1}\Big]_\nu = [(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu, \tag{71}$$

$$\lim_{N_0\to\infty}\big[N_0\boldsymbol{G}_{\rm opt}\big]_\nu = [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu. \tag{72}$$

Proof: When $\boldsymbol{P}=\boldsymbol{0}$, from Theorem 2 the optimal $\boldsymbol{G}$ for (24) satisfies (18). From (69), when $N_0\to 0$ it holds that $\boldsymbol{M}\to\boldsymbol{0}$, and when $N_0\to\infty$ we have $\boldsymbol{M}\to-\boldsymbol{I}$. Therefore, by the definition of $\boldsymbol{\Omega}$ it holds that

$$\lim_{N_0\to 0,\infty}\boldsymbol{d} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{MP}) = \boldsymbol{0}.$$

This implies that the gradient $d_{\boldsymbol{G}}(\delta_2)$ (showing in (85) in Appendix F) converges to zero. Hence the differentials of $I_{\rm GMI}(\boldsymbol{V}_{\rm opt},\boldsymbol{R}_{\rm opt},\boldsymbol{G})$ in (24) with $\boldsymbol{P}\neq\boldsymbol{0}$ converges to the those with $\boldsymbol{P}=\boldsymbol{0}$. From (18) and (69), the limit (71) follows.

Next, since

$$\begin{aligned}\lim_{N_0\to\infty}\Big[N_0\Big(\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}\Big)\Big]_\nu &= \lim_{N_0\to\infty}[N_0(\boldsymbol{I}+\boldsymbol{M})]_\nu\\ &= [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu,\end{aligned} \tag{73}$$

which shows that $\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}\to\boldsymbol{0}$[9] as $N_0\to\infty$. By the matrix inversion lemma [58], $\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}\to\boldsymbol{G}_{\rm opt}$ as $N_0\to\infty$, and combining it with (73) proves (72). ■

Lemma 4: In Method II, with the optimal $\boldsymbol{G}$, when $N_0\to 0$ the GMI increment $\delta_2(\boldsymbol{G})$ in (26) converges to zero with speed $\mathcal{O}(1/N_0)$,[10] and when $N_0\to\infty$ it converges to zero with speed $\mathcal{O}(N_0^2)$, respectively.

Proof: As $N_0\to 0$, from (69) we have

$$\begin{aligned}\lim_{N_0\to 0}\boldsymbol{d}/N_0 &= \lim_{N_0\to 0}\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{MP}/N_0)\\ &= -\boldsymbol{\Omega}\mathrm{vec}\big((\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\boldsymbol{P}\big).\end{aligned}$$

Based on (70) and Lemma 3, the below equalities hold,

$$\begin{aligned}\delta_2(\boldsymbol{G}_{\rm opt}) &= N_0\frac{\boldsymbol{d}^\dagger}{N_0}\Big(\boldsymbol{\Omega}\Big(\tilde{\boldsymbol{M}}^*\otimes\frac{(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}}{N_0}\Big)\boldsymbol{\Omega}^T\Big)^{-1}\frac{\boldsymbol{d}}{N_0}\\ &= \mathcal{O}(N_0).\end{aligned}$$

On the other hand, as $N_0\to\infty$, and by the definition of $\boldsymbol{\Omega}$, from (69) we also have

$$\begin{aligned}\lim_{N_0\to\infty}N_0\boldsymbol{d} &= \lim_{N_0\to\infty}\boldsymbol{\Omega}\mathrm{vec}(N_0\boldsymbol{MP})\\ &= \lim_{N_0\to\infty}\boldsymbol{\Omega}\mathrm{vec}\big(N_0(\boldsymbol{I}+\boldsymbol{M})\boldsymbol{P}\big)\\ &= \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{H}^\dagger\boldsymbol{HP}).\end{aligned}$$

[9] A matrix $\boldsymbol{A}\to\boldsymbol{B}$ or a vector $\boldsymbol{a}\to\boldsymbol{b}$ means the non-zero elements of $\boldsymbol{A}-\boldsymbol{B}$ or $\boldsymbol{a}-\boldsymbol{b}$ converges to zero.

[10] Two scalars A and B as functions of a variable n converging to each other with speed $\mathcal{O}(n)$ means that there exists a constant C such that $\lim_{n\to\infty}n|A-B|<C$.

Again utilizing (70) and Lemma 3, the below equalities hold,

$$\delta_2(\boldsymbol{G}_{\text{opt}}) = \frac{1}{N_0^2}(N_0\boldsymbol{d}^\dagger)\left(\boldsymbol{\Omega}\left(\tilde{\boldsymbol{M}}^* \otimes (\boldsymbol{I}+\boldsymbol{G}_{\text{opt}})^{-1}\right)\boldsymbol{\Omega}^T\right)^{-1}(N_0\boldsymbol{d}) = \mathcal{O}(1/N_0^2).$$

Therefore, Lemma 4 holds. ∎

Lemma 5: When both $N_0 \to 0$ and ∞, the optimal GMI in Method III is independent of $\boldsymbol{P}$ and converges to the optimal GMI with $\boldsymbol{P} = \boldsymbol{0}$. Further, (71) and (72) also hold for Method III.

The proof is given in Appendix K. Combining Lemmas 3-5, and using the fact that Method III and Method II are equivalent with $\boldsymbol{P}=\boldsymbol{0}$, we have the following Theorem 4.

Theorem 4: Assume that $\boldsymbol{P} \prec \boldsymbol{I}$, when $N_0 \to 0$ and ∞, the optimal GMI in Method III converges to the optimal GMI in Method III with $\boldsymbol{P} = \boldsymbol{0}$. Further, the optimal GMI in Method II also converges to the same, with speed $\mathcal{O}(1/N_0)$ as SNR increase, and $\mathcal{O}(N_0^2)$ as SNR decreases, respectively. The optimal $\boldsymbol{G}$ for both methods has the limits (71) and (72).

From Theorem 4 we know that, except for the case where one of the elements in the diagonal matrix $\boldsymbol{P}$ is 1, the soft-feedback information becomes asymptotically insignificant for the design of the CS parameters. The reason is that, when $N_0 \to 0$, $\hat{\boldsymbol{x}}$ is overwhelmed by the noise, while when $N_0 \to \infty$, the optimal front-end filter will null out $\hat{\boldsymbol{x}}$ since the receiver can perfectly reconstruct the transmitted symbols without the side-information.

Remark 4: When $N_0 \to 0$ and ∞, the optimal CS demodulator is the EZF demodulator defined in Example 1, and the TMF defined in Example 2, respectively.

With ISI channels, as the same constraint $\boldsymbol{P} = \alpha\boldsymbol{I} \prec \boldsymbol{I}$ shall hold, we make the restriction that $\alpha < 1$. The asymptotic properties for ISI channels are presented in Corollary 1, which is an asymptotic version of Theorem 4 when the channel matrix $\boldsymbol{H}$ and CS parameters are band-shaped Toeplitz matrices with infinite dimensions. The detailed proof follows the same analysis as for the MIMO case and is omitted.

Corollary 1: Assume that $0 \le \alpha < 1$, when $N_0 \to 0$ and ∞, the optimal GMI in Method III converges to the optimal GMI in Method III with $\alpha = 0$. Further, the optimal GMI in Method II also converges to the same, with speed $\mathcal{O}(1/N_0)$ as SNR increase, and $\mathcal{O}(N_0^2)$ as SNR decreases, respectively. The optimal $\boldsymbol{G}$ for both methods has the following asymptotic properties hold for $|k| \le \nu$:

$$\lim_{N_0\to 0}\int_{-\pi}^{\pi}\frac{1}{N_0(1+G_{\text{opt}}(\omega))}\exp(-jk\omega)\mathrm{d}\omega = \int_{-\pi}^{\pi}\frac{1}{|H(\omega)|^2}\exp(-jk\omega)\mathrm{d}\omega,$$

$$\lim_{N_0\to \infty}\int_{-\pi}^{\pi}N_0G_{\text{opt}}(\omega)\exp(-jk\omega)\mathrm{d}\omega = \int_{-\pi}^{\pi}|H(\omega)|^2\exp(-jk\omega)\mathrm{d}\omega.$$

VII. NUMERICAL RESULTS

In this section, we provide numerical results to elaborate the behaviors of CS demodulators in an iterative detection and decoding receiver designs. In the considered MIMO channels, all channel elements are assumed to be independent and identically distributed (i.i.d.) complex-valued Gaussian variables with zero-means and unit-variance. Further, the transmit-power at each transmit-antenna is normalized to unity. For the ISI case, we use the typical Proakis-C channel as in Fig. 5.

A. GMI EVALUATION

We first evaluate the GMI under 5×5 MIMO channels with memory-size $\nu = 1$ for all CS demodulators. We simulate 10000 channel realizations for each SNR point. The GMIs are compared with those of the static CS demodulator [18], which is equivalent to Method II with $\boldsymbol{P}=\boldsymbol{0}$. The channel capacity is also presented for comparisons. The results of GMI are shown in Fig. 6. As the quality of soft-information improves beyond $\boldsymbol{P}=\boldsymbol{0}$, Method II with $\nu_{\text{R}}=0$ performs the best among all CS demodulators, as it has the most degrees of freedom (DoF) in designing $\boldsymbol{R}$. Method II with $\nu_{\text{R}} = \nu$ is the worst among Method I and II, while Method I is slightly worse than Method II with $\boldsymbol{R}$ of shape (a) in Fig. 3. This is because although the IC matrix $\boldsymbol{R}$ is of shape (a) in both cases, $\boldsymbol{R}$ in Method II is more general than in Method I, since the latter one is constrained to $\boldsymbol{R} = \boldsymbol{F}^\dagger\boldsymbol{T}$. Further, the GMI of Method III is inferior to Method II as expected.

The results show consistent GMI increments for all CS demodulators when the feedback quality improves. When $\boldsymbol{P}$ increases from $\boldsymbol{P} = \boldsymbol{0}$ to the ideal case $\boldsymbol{P} = \boldsymbol{I}$, the channel capacity becomes inferior to the GMI as the pair $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ is superior to $(\boldsymbol{x}, \boldsymbol{y})$ for information-transfer.

B. SNR ASYMPTOTIC OF THE GMI

Next, we evaluate the asymptotic properties described in Theorem 4 under 5×5 MIMO channels. As shown in Fig. 7, the GMIs of both Method II and III converge to that of Method III with $\boldsymbol{P}=\boldsymbol{0}$. In addition, the GMI converges to the EZF in Example 1 at high SNR, and the TMF in Example 2 at low SNR, respectively, which are aligned with the limits in Theorem 4.

C. EXIT CHARTS OF CS DEMODULATORS

In order to predict the dynamics of iterative receivers, we extrinsic information transfer (EXIT) charts [41], [59] to analyse iterative receivers. The demodulators and the decoder measure the output extrinsic information I_E based on a sequence of observations $\boldsymbol{y}$ and *a priori* mutual information I_A.

In Fig. 8, we evaluate the EXIT charts for CS demodulators under 4×6 MIMO channels with $\nu = 2$ for $\boldsymbol{F}$ and $\boldsymbol{G}$ at an SNR of 10 dB. With Method II, we test different values of ν_{R}. As can be seen, when $\nu_{\text{R}} > \nu$, the demodulation-performance is inferior to $\nu_{\text{R}} \le \nu$. This is because, the

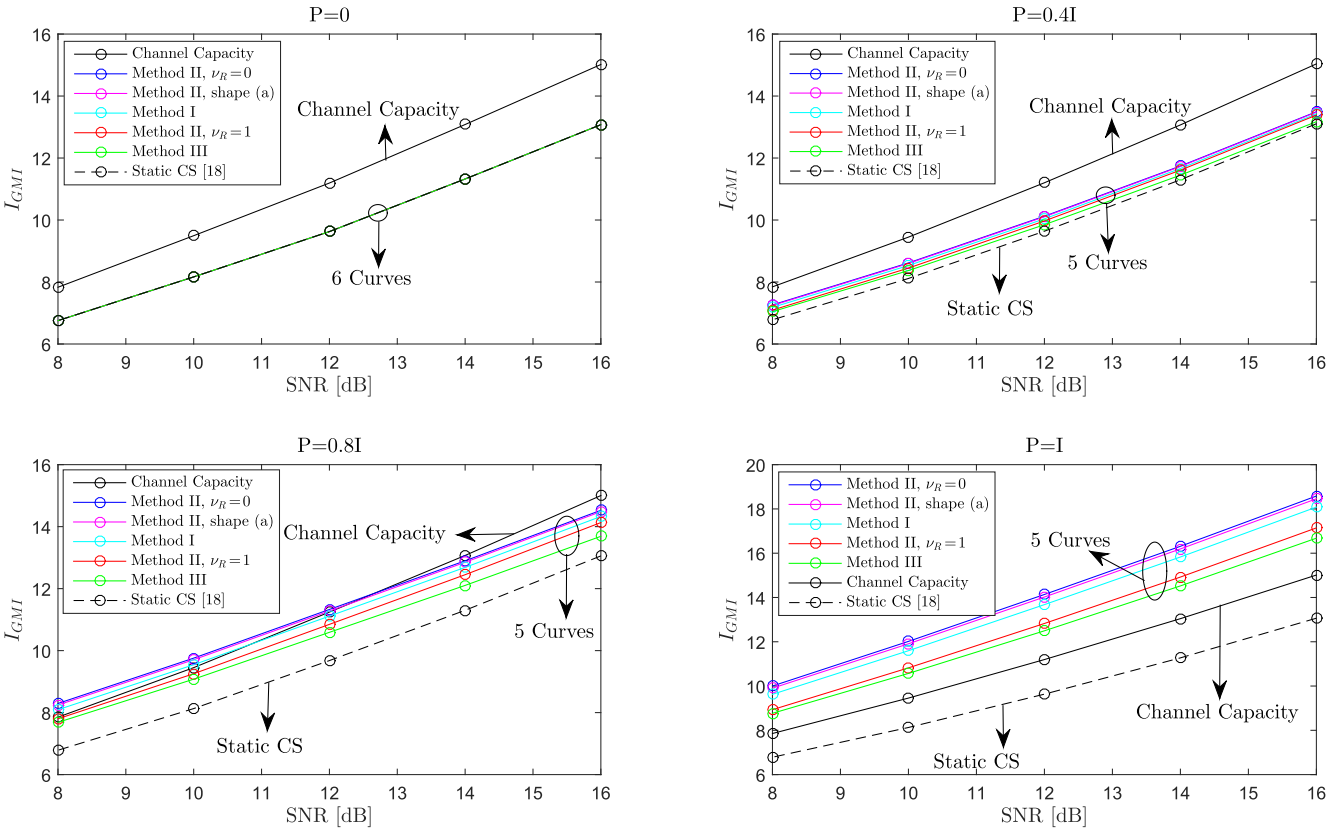


FIGURE 6. GMI of the CS demodulators under 5 × 5 MIMO i.i.d. complex-valued Gaussian channels with $\nu = 1$.

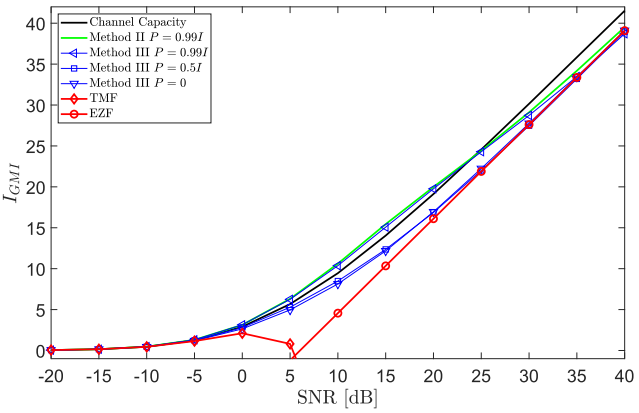


FIGURE 7. SNR asymptotic under 5 × 5 MIMO i.i.d. complex-valued Gaussian channels with $\nu = 1$.

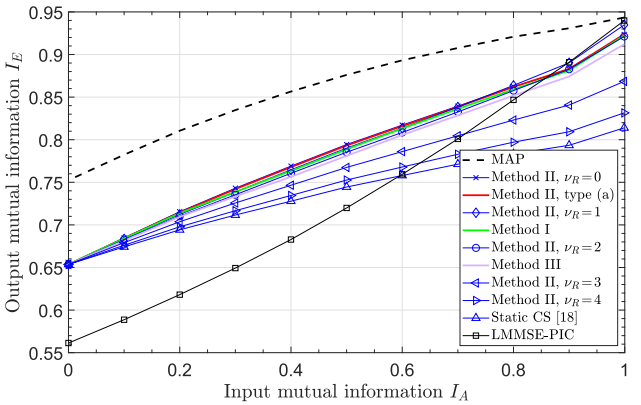


FIGURE 8. EXIT charts under 4 × 6 i.i.d. complex-valued Gaussian MIMO channels with $\nu = 2$ and different values of ν_R.

interference outside memory-size ν and inside memory-size ν_R is neither considered in the IC nor in the BCJR. Moreover, with $\nu_R \leq \nu$ the CS demodulators with Method II performs quite close to each other as well as Method I and III. For Method II with $\nu_R < \nu$, the interference inside memory-size ν and outside ν_R are considered both in the IC and BCJR processes. However, an interesting observation is that, with a larger input I_A, Method II with $\nu_R = 0$ is inferior to $\nu_R = 1$ and $\nu_R = 2$. Therefore, a conservative and robust approach with Method II is to set $\nu_R = \nu$ such that the interference is either removed in IC or dealt with in BCJR, to get rid of potential error-propagation caused by redundant processing of the same signal part.

In Fig. 9, we show the iterative detection and decoding trajectories for CS demodulators under Proakis-C channel with $\nu = 2$ and at an SNR of 10 dB. We use a (7, 5)-convolutional code [24] with a coded block-length $K = 2004$, and a random permutation is applied to the coded-bits. As can

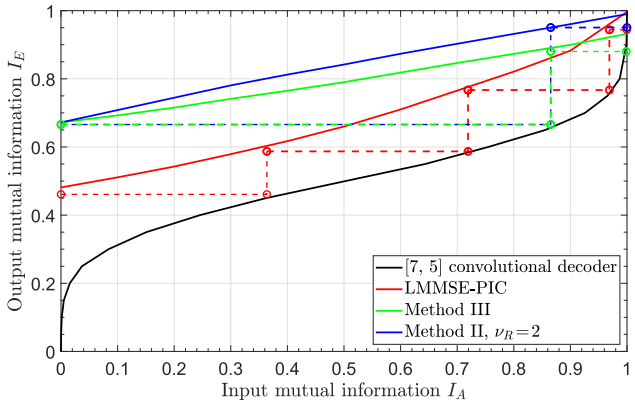


FIGURE 9. Iterative detection and decoding trajectories under Proakis-C channel at an SNR of 10 dB. The outer code is a (7, 5)-convolutional code with generator polynomials $g_0(D)=1+D^2$ and $g_1(D)=1+D+D^2$. The black curve represents the decoder and the dashed lines are the iterative detection and decoding trajectories for LMMSE-PIC, Method III and II, respectively.

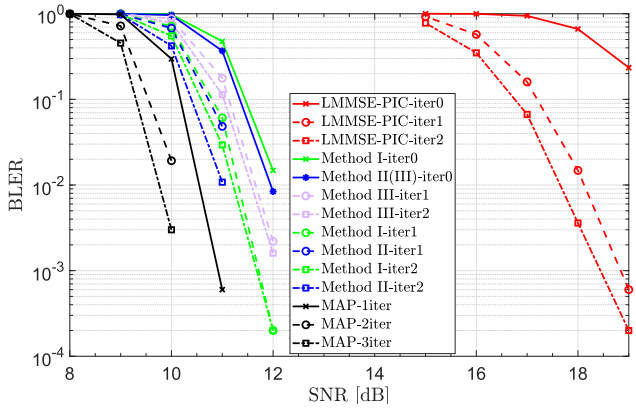


FIGURE 10. BLER performance of the LMMSE-PIC, Method I, II, and III, and MAP under Proakis-C channel with QPSK modulation.

be seen, the CS demodulators with Method II and III are superior to the LMMSE-PIC demodulator, and the iterative detection and decoding trajectories are well aligned with the EXIT charts.

D. LINK PERFORMANCE

We next turn to link-level simulations with turbo codes [60] where the outer-decoder uses 8 internal iterations. A single code-block over all transmit symbols is used. At each SNR point 20000 data-blocks are simulated and the block-error-ratio (BLER) is measured. In all simulations, maximal three global iterations are used between the demodulators. The tap-length of both the front-end fitler and IC filter are set to $8L$, and $\nu_R=\nu$ for Method II.

In Fig. 10, we evaluate the BLER under Proakis-C channel with QPSK symbols and $\nu=2$ for all CS demodulators. A (1064, 1600) turbo code is used. Note that, at the first iteration when there is no soft-information, Method II and III are overlapped with each other. With CS demodulators, the gap to the MAP demodulator is less than 1 dB,

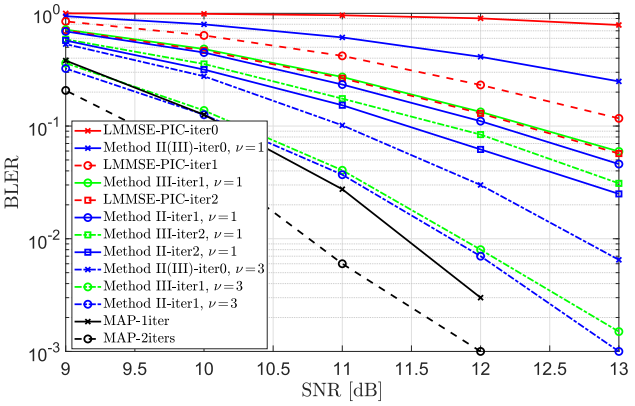


FIGURE 11. BLER performance of the LMMSE-PIC, Method II, and III, and MAP under 4×6 MIMO channels with QPSK modulation.

while the LMMSE-PIC has a gap to the MAP that is up to 10 dB. Moreover, Method II performs slightly better than Method I, and Method III is slightly inferior to both methods. However, Method III has the advantage of less computational complexity than the others since all parameters are in closed-forms.

In Fig. 11, we evaluate the BLER under 4×6 MIMO channels with QPSK symbols and $\nu=3$ for all CS demodulators. A (1064,1800)-turbo code is used. As $N<K$, the LMMSE-PIC fails [61] at the first iteration due to the lack of sufficient receive diversity. However, the CS demodulators with $\nu=3$ significantly improve the performance and with less than 1 dB gap to the MAP at 10% BLER. Further, the CS demodulators with $\nu=1$ after three iterations is less than 2 dB away from the MAP. More interestingly, with less complexity, Method III performs close to Method II.

Finally we remark that for the sake of complexity savings, the CS parameters do not need to be updated through all iterations. Once the feedback information quality is good enough and the changes in $\boldsymbol{P}$ or α are small, the CS parameters can be kept unchanged in successive iterations.

VIII. SUMMARY

In this paper we have considered the design of CS demodulators for linear vector channels that use a trellis-representation of the received signal, in combination with interference cancellation (IC) of the signal part that is not appropriately modeled by the trellis. In order to reach a trellis-representation, a linear filter is applied as front-end. It is an extension of the well studied CS demodulators to iterative receivers and a generalization of the LMMSE-PIC demodulator to cooperate with trellis-search in turbo equalization.

We have analyzed the properties of three different methods for designing optimal CS demodulators, as all of them may come across as natural CS demodulators. In the used framework, there are three parameters that need to be optimized: the front-end filter, the IC filter, and the target response. Based on maximizing the GMI, the first two are solved in closed-forms, while the third one needs numerical optimizations and simple

gradient based approaches have been shown to perform well. In particular, in Method III we have constructed all three CS parameters explicitly at a cost of small GMI losses.

Numerical results have been provided to illustrate the behaviors of the proposed CS demodulators. In general, Method II based on the Ungerboeck model is superior to Method I based on the Forney model. Method II has the advantage over Method I that the optimization procedure is concave. Further, the suboptimal Method III performs close to Method I and II. An interesting result is that the IC of the CS demodulators should not perfectly cancel the effective channel outside the memory-size ν, a property that cannot be seen in the LMMSE-PIC demodulator as $\nu = 0$. Moreover, we have also analyzed asymptotic properties of the CS demodulators and showed that Method III converges to Method II asymptotically when the noise power goes either zero or infinity.

APPENDIX A
DERIVATION OF THE GMI

By making the eigenvalue decomposition $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\dagger = \boldsymbol{G}$ and letting $\boldsymbol{s} = \boldsymbol{Q}^\dagger\boldsymbol{x}$. As $\boldsymbol{x}$ is assumed to be zero mean complex Gaussian random vector with covariance matrix $\boldsymbol{I}$, we can write $\tilde{p}(\boldsymbol{y}|\boldsymbol{x})$ in (4) as

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{x}) = \exp\big(2\mathcal{R}\{\boldsymbol{s}^\dagger\boldsymbol{d}\} - \boldsymbol{s}^\dagger\boldsymbol{\Lambda}\boldsymbol{s}\big), \tag{74}$$

where $\boldsymbol{d} = \boldsymbol{Q}^\dagger(\boldsymbol{V}\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{x}})$. We can now evaluate

$$\begin{aligned}\tilde{p}(\boldsymbol{y}) &= \int \tilde{p}(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{\pi^K}\int \exp\big(2\mathcal{R}\{\boldsymbol{s}^\dagger\boldsymbol{d}\} - \boldsymbol{s}^\dagger\boldsymbol{\Lambda}\boldsymbol{s}\big)\exp\big(-\boldsymbol{s}^\dagger\boldsymbol{s}\big)\mathrm{d}\boldsymbol{s} \\ &= \prod_{k=1}^{N}\frac{1}{1+\lambda_k}\exp\left(\frac{|d_k|^2}{1+\lambda_k}\right),\end{aligned}$$

where λ_k is the kth diagonal element of $\boldsymbol{\Lambda}$ and d_k is the kth entry of $\boldsymbol{d}$. Taking the average over $\boldsymbol{y}$ gives

$$-\mathbb{E}_{p(\boldsymbol{y})}[\log(\tilde{p}(\boldsymbol{y}))] = \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)$$

where the matrix $\boldsymbol{L} = \mathbb{E}[\boldsymbol{Q}\boldsymbol{d}\boldsymbol{d}^\dagger\boldsymbol{Q}^\dagger]$ is given by

$$\boldsymbol{L} = \boldsymbol{V}(N_0\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^\dagger)\boldsymbol{V}^\dagger - \boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^\dagger - \boldsymbol{R}\boldsymbol{P}\boldsymbol{H}^\dagger\boldsymbol{V}^\dagger + \boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^\dagger.$$

On the other hand, we have

$$-\mathbb{E}_{p(\boldsymbol{y},\boldsymbol{x})}\big[\log(\tilde{p}(\boldsymbol{y}|\boldsymbol{x}))\big] = \mathrm{Tr}(\boldsymbol{G}) - 2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H} - \boldsymbol{R}\boldsymbol{P})\big\}.$$

Combining the two expectations, the GMI reads

$$\begin{aligned}&I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G}) \\ &\quad= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}\big(\boldsymbol{L}(\boldsymbol{I}+\boldsymbol{G})^{-1}\big) - \mathrm{Tr}(\boldsymbol{G}) \\ &\qquad+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H} - \boldsymbol{R}\boldsymbol{P})\big\} \\ &\quad= \log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big) - \mathrm{Tr}(\boldsymbol{G}) + 2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{V}\boldsymbol{H} - \boldsymbol{R}\boldsymbol{P})\big\} \\ &\qquad-\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big(\boldsymbol{V}[\boldsymbol{H}\boldsymbol{H}^\dagger + N_0\boldsymbol{I}]\boldsymbol{V}^\dagger \\ &\qquad-2\mathcal{R}\big\{\boldsymbol{V}\boldsymbol{H}\boldsymbol{P}\boldsymbol{R}^\dagger\big\} + \boldsymbol{R}\boldsymbol{P}\boldsymbol{R}^\dagger\big)\big).\end{aligned}$$

APPENDIX B
THE PROOF OF PROPOSITION 1

As the formula of GMI in (9) is quadratic in $\boldsymbol{W}$ and no constraints apply to $\boldsymbol{W}$, taking the gradient of $I_{\mathrm{GMI}}(\boldsymbol{W},\boldsymbol{T},\boldsymbol{F})$ with respect to $\boldsymbol{W}$ and setting it to zero, the optimal $\boldsymbol{W}$ is then in (12). Inserting $\boldsymbol{W}_{\mathrm{opt}}$ back into (9) yields, after some manipulations,

$$\begin{aligned}&I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T},\boldsymbol{F}) \\ &\quad= K + \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) + \mathrm{Tr}\big(\boldsymbol{T}^\dagger\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}\tilde{\boldsymbol{M}}\big) \\ &\qquad+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) + 2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^\dagger\boldsymbol{T}\big)\big\}.\end{aligned} \tag{75}$$

where $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$ are defined in (10) and (11), respectively. When $\boldsymbol{P} = \boldsymbol{0}$, (75) equals

$$I_1(\boldsymbol{F}) = K + \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) + \mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big).$$

In this case, there is no soft-information available and the matrix $\boldsymbol{T}$ is not included in the formula. When $\boldsymbol{P} \neq \boldsymbol{0}$, the terms of I_{GMI} in (75) related to $\boldsymbol{T}$ are

$$f(\boldsymbol{T}) = \mathrm{Tr}\big(\boldsymbol{T}^\dagger\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\boldsymbol{T}\tilde{\boldsymbol{M}}\big) + 2\mathcal{R}\big\{\mathrm{Tr}\big(\boldsymbol{P}\boldsymbol{M}\boldsymbol{F}^\dagger\boldsymbol{T}\big)\big\}.$$

Let $\boldsymbol{t}_k$ denote the kth column of $\boldsymbol{T}$, but with all elements in rows $[k, \min(k+\nu, K-1)]$ removed, and define the column vector $\boldsymbol{t} = [\boldsymbol{t}_0^T,\ \boldsymbol{t}_1^T,\ \ldots,\ \boldsymbol{t}_{K-1}^T]^T$. Then, by the definition of indication matrix $\boldsymbol{\Omega}$, we have

$$\boldsymbol{t} = \boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{T}).$$

Similarly, let z_k denote the kth column of the matrix $\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}$ with all elements in rows $[k, \min(k+\nu, K-1)]$ removed, and define a row vector $\boldsymbol{z} = [z_0^T,\ z_1^T,\ \ldots,\ z_{K-1}^T]^T$. Then, we have

$$\boldsymbol{z} = \Omega\mathrm{vec}(\boldsymbol{F}\boldsymbol{M}\boldsymbol{P}) = \boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*)\otimes\boldsymbol{I}_K\big)\mathrm{vec}(\boldsymbol{F}).$$

Finally, defining a Hermitian matrix

$$\hat{\boldsymbol{B}}_1 = \boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^T,$$

and with that we can rewrite $f(\boldsymbol{T})$ as

$$f(\boldsymbol{T}) = \boldsymbol{t}^\dagger\hat{\boldsymbol{B}}_1\boldsymbol{t} + 2\mathcal{R}\{\boldsymbol{z}^\dagger\boldsymbol{t}\}.$$

Taking the gradient of $f(\boldsymbol{T})$ with respect to $\boldsymbol{t}$ and setting it to zero yields

$$\boldsymbol{t}_{\mathrm{opt}} = -\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}. \tag{76}$$

Transferring $\boldsymbol{t}_{\mathrm{opt}}$ back into $\boldsymbol{T}_{\mathrm{opt}}$ gives the optimal $\boldsymbol{T}$ in (13), and inserting it back into $f(\boldsymbol{T})$ gives

$$f(\boldsymbol{T}_{\mathrm{opt}}) = -\boldsymbol{z}^\dagger\hat{\boldsymbol{B}}_1^{-1}\boldsymbol{z}.$$

Thus, with the optimal $\boldsymbol{W}$ and $\boldsymbol{T}$ and when $\boldsymbol{P} \neq \boldsymbol{0}$ the GMI equals

$$\begin{aligned}&I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F}) \\ &= K + \log\big(\det(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) + \mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})\big) \\ &\quad-\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\Big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes\big(\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger\big)\big)\boldsymbol{\Omega}^T\Big)^{-1}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}).\end{aligned}$$

where

$$\boldsymbol{D} = \boldsymbol{\Omega}\big((\boldsymbol{P}\boldsymbol{M}^*)\otimes\boldsymbol{I}_K\big).$$

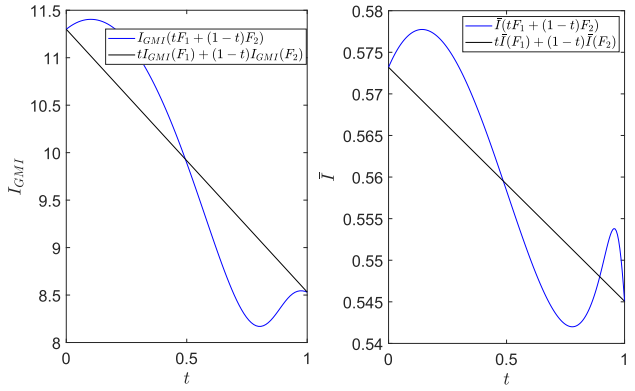


FIGURE 12. Non-concaveness of Method I under 5×5 MIMO channel (left figure) and Proakis-C ISI channel (right figure).

APPENDIX C
NON-CONCAVITY EXAMPLES OF METHOD I

We give two examples to demonstrate the non-concavity of Method I for both MIMO and ISI cases by assuming that $\boldsymbol{P}=\boldsymbol{I}$ and $\alpha=1$, respectively. The memory-size is $\nu=1$ and the noise power N_0 equals 1 in both cases. A 5×5 MIMO channel and the Proakis-C ISI channel are used.

Example 4: MIMO case:

$$\boldsymbol{H}=\begin{bmatrix} 2 & 0 & -3 & 5 & 4\\ -5 & 2 & -1 & 0 & 2\\ 2 & -4 & 3 & 3 & 3\\ -1 & -5 & -4 & 1 & 2\\ 0 & -2 & 0 & 5 & 5 \end{bmatrix},$$

$$\boldsymbol{F}_1=\begin{bmatrix} 4.94 & 4.45 & 0 & 0 & 0\\ 0 & 0.21 & 3.85 & 0 & 0\\ 0 & 0 & 5.56 & 1.76 & 0\\ 0 & 0 & 0 & 0.61 & 7.10\\ 0 & 0 & 0 & 0 & 2.79 \end{bmatrix},$$

$$\boldsymbol{F}_2=\begin{bmatrix} 2.03 & 6.17 & 0 & 0 & 0\\ 0 & 5.22 & 3.56 & 0 & 0\\ 0 & 0 & 7.43 & 0.73 & 0\\ 0 & 0 & 0 & 4.98 & 4.32\\ 0 & 0 & 0 & 0 & 10.11 \end{bmatrix}.$$

Example 5: ISI case:

$$\boldsymbol{h}=\begin{bmatrix}0.227 & 0.460 & 0.688 & 0.460 & 0.227\end{bmatrix},$$

$$\boldsymbol{f}_1=\begin{bmatrix}0.1606 & 0.9009\end{bmatrix},\ \boldsymbol{f}_2=\begin{bmatrix}0.2230 & 0.2035\end{bmatrix}.$$

The $I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})$ in (14) as a function $\boldsymbol{F}$ is plotted on the left of Fig. 12, while the $\bar{I}(W_{\mathrm{opt}}(\omega),T_{\mathrm{opt}}(\omega),F(\omega))$ in (52) as a function of $F(\omega)$ is plotted on the right. If $I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{\mathrm{opt}},\boldsymbol{F})$ and $\bar{I}(W_{\mathrm{opt}}(\omega),T_{\mathrm{opt}}(\omega),F(\omega))$ are concave or convex, the blue curves lie above or below the black curves, which clearly does not hold in the examples.

APPENDIX D
THE GRADIENT IN METHOD I FOR THE MIMO CASE

In this section we derive the first order differential of the GMI given in (14) with respect to $\boldsymbol{F}$. In order to utilize the differential with respect to a matrix, we use the α-differential as defined in [62]. Assume a matrix $\boldsymbol{Y}_{N,K}$ with dimension $N\times K$ and a matrix $\boldsymbol{X}_{M,S}$ with dimension $M\times S$, define $d_{\boldsymbol{X}}\boldsymbol{Y}$ as the α-differential of $\boldsymbol{Y}$ with respect to $\boldsymbol{X}$. Further, defining y_ℓ and x_ℓ as $[y_1,\ y_2,\ \cdots,\ y_{NK}]=\mathrm{vec}(\boldsymbol{Y})^T$ and $[x_1,\ x_2,\ \cdots,\ x_{MS}]=\mathrm{vec}(\boldsymbol{X})^T$, the α-differential $d_{\boldsymbol{X}}\boldsymbol{Y}$ is

$$d_{\boldsymbol{X}}\boldsymbol{Y}=\frac{\partial\mathrm{vec}(\boldsymbol{Y})}{\partial\mathrm{vec}(\boldsymbol{X})^T}=\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_{MS}}\\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_{MS}}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial y_{NK}}{\partial x_1} & \frac{\partial y_{NK}}{\partial x_2} & \cdots & \frac{\partial y_{NK}}{\partial x_{MS}} \end{bmatrix}.$$

The reason for adopting the α-differential is because it keeps both the chain rule and the product rule. We introduce an $NK\times NK$ permutation matrix $\boldsymbol{Z}_{N,K}$, which satisfies the condition $\mathrm{vec}(\boldsymbol{Y}^T)=\boldsymbol{Z}_{N,K}\mathrm{vec}(\boldsymbol{Y})$. It is easy to verify that $\boldsymbol{Z}_{N,K}^{-1}=\boldsymbol{Z}_{K,N}$. In addition, when $N=1$ or $K=1$, $\boldsymbol{Y}$ is a vector and it holds that $\mathrm{vec}(\boldsymbol{Y}^T)=\mathrm{vec}(\boldsymbol{Y})$, hence, we have $\boldsymbol{Z}_{N,1}=\boldsymbol{I}_N$ and $\boldsymbol{Z}_{1,K}=\boldsymbol{I}_K$. Furthermore, by definition it holds that $d_{\boldsymbol{F}}(\boldsymbol{F})=d_{\boldsymbol{F}}(\mathrm{vec}(\boldsymbol{F}))=\boldsymbol{I}$, and $d_{\boldsymbol{F}}(\boldsymbol{F}^\dagger)=d_{\boldsymbol{F}}\left(\mathrm{vec}(\boldsymbol{F}^\dagger)\right)=\boldsymbol{0}$.

We start by reviewing a few properties [62], [63] of α-differential below that will be used later, where both matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ are functions of $\boldsymbol{F}$ and the dimensions are specified by subscripts associated to them:

$$\begin{aligned} d_{\boldsymbol{F}}(\boldsymbol{X}_{K,K}^{-1}) &= -(\boldsymbol{X}_{K,K}^{-T}\otimes\boldsymbol{X}_{K,K}^{-1})d_{\boldsymbol{F}}\boldsymbol{X}_{K,K}\\ d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K}\boldsymbol{X}_{K,S}) &= (\boldsymbol{X}_{K,S}^T\otimes\boldsymbol{I}_N)d_{\boldsymbol{F}}\boldsymbol{Y}_{N,K}\\ &\quad+(\boldsymbol{I}_S\otimes\boldsymbol{Y}_{N,K})d_{\boldsymbol{F}}\boldsymbol{X}_{K,S}\\ d_{\boldsymbol{F}}(\log(\det(\boldsymbol{X}_{K,K}))) &= \mathrm{vec}(\boldsymbol{X}_{K,K}^{-T})^T d_{\boldsymbol{F}}\boldsymbol{X}_{K,K}\\ d_{\boldsymbol{F}}(\boldsymbol{Y}_{N,K}\otimes\boldsymbol{X}_{M,S}) &= (\boldsymbol{I}_K\otimes\boldsymbol{Z}_{S,N}\otimes\boldsymbol{I}_M)\\ &\quad\times(\boldsymbol{I}_{NK}\otimes\mathrm{vec}(\boldsymbol{X})d_{\boldsymbol{F}}\boldsymbol{Y}_{N,K}\\ &\quad+(\boldsymbol{I}_K\otimes\boldsymbol{Z}_{S,N}\otimes\boldsymbol{I}_M)\\ &\quad\times(\mathrm{vec}(\boldsymbol{Y})\otimes\boldsymbol{I}_{MS})d_{\boldsymbol{F}}\boldsymbol{X}_{M,S}. \end{aligned}$$

The α-differential of $I_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$\begin{aligned} d_{\boldsymbol{F}}(I_1) &= \mathrm{vec}((\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-T})^T(\boldsymbol{I}_K\otimes\boldsymbol{F}^\dagger)+\mathrm{vec}(\boldsymbol{F}^*\boldsymbol{M}^T)^T\\ &= \mathrm{vec}(\boldsymbol{F}\boldsymbol{M}+\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-\dagger})^\dagger. \end{aligned}\tag{77}$$

Defining a $K\times K$ matrix $\boldsymbol{B}=\boldsymbol{F}(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger$ and an $S\times S$ matrix $\boldsymbol{\Pi}=\left(\boldsymbol{\Omega}(\tilde{\boldsymbol{M}}^*\otimes\boldsymbol{B})\boldsymbol{\Omega}^T\right)^{-1}$, the α-differential of $\delta_1(\boldsymbol{F})$ with respect to $\boldsymbol{F}$ is

$$\begin{aligned} d_{\boldsymbol{F}}(\delta_1) = -\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\left((\mathrm{vec}(\boldsymbol{F})^T\boldsymbol{D}^T)\otimes\boldsymbol{I}_S\right)d_{\boldsymbol{F}}(\boldsymbol{\Pi})\\ -\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\boldsymbol{\Pi}\boldsymbol{D}, \end{aligned}\tag{78}$$

where

$$\begin{aligned} d_{\boldsymbol{F}}(\boldsymbol{\Pi}) &= -(\boldsymbol{\Pi}^T\otimes\boldsymbol{\Pi})d_{\boldsymbol{F}}(\boldsymbol{\Omega}(\tilde{\boldsymbol{M}}^*\otimes\boldsymbol{B})\boldsymbol{\Omega}^T)\\ &= -\left((\boldsymbol{\Pi}^T\boldsymbol{\Omega})\otimes(\boldsymbol{\Pi}\boldsymbol{\Omega})\right)(\boldsymbol{I}_K\otimes\boldsymbol{Z}_{K,K}\otimes\boldsymbol{I}_K)\\ &\quad\cdot(\mathrm{vec}(\tilde{\boldsymbol{M}}^*)\otimes\boldsymbol{I}_{K^2})d_{\boldsymbol{F}}\boldsymbol{B} \end{aligned}\tag{79}$$

and

$$\begin{aligned} d_{\boldsymbol{F}}(\boldsymbol{B}) &= d_{\boldsymbol{F}}\big(\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-1}\big) \\ &= \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-T}\big)\otimes\big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-1}\big)d_{\boldsymbol{F}}(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger) \\ &= \big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-T}\big)\otimes\big((\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-1}\big)(\boldsymbol{F}^*\otimes\boldsymbol{I}_K) \\ &= \big(\boldsymbol{F}^*(\boldsymbol{I}+\boldsymbol{F}\boldsymbol{F}^\dagger)^{-T}\big)\otimes(\boldsymbol{I}-\boldsymbol{B}). \end{aligned} \tag{80}$$

Then, defining a $K\times K$ matrix

$$\tilde{\boldsymbol{F}}=(\boldsymbol{I}+\boldsymbol{F}^\dagger\boldsymbol{F})^{-1}\boldsymbol{F}^\dagger$$

and a $K^4\times K^2$ matrix

$$\boldsymbol{\Psi} = (\boldsymbol{I}_K\otimes\boldsymbol{Z}_{K,K}\otimes\boldsymbol{I}_K)\big(\mathrm{vec}(\tilde{\boldsymbol{M}}^*)\otimes\boldsymbol{I}_{K^2}\big), \tag{81}$$

and by combing (77)-(81), we finally have when $\boldsymbol{P}\neq\boldsymbol{0}$ that

$$\begin{aligned} &d_{\boldsymbol{F}}\big(I_{\mathrm{GMI}}(\boldsymbol{W}_{\mathrm{opt}},\boldsymbol{T}_{2\mathrm{pt}},\boldsymbol{F})\big) \\ &= d_{\boldsymbol{F}}(I_1)+d_{\boldsymbol{F}}(\delta_1) \\ &= \mathrm{vec}\big(\boldsymbol{F}\boldsymbol{M}+\tilde{\boldsymbol{F}}^\dagger\big)^\dagger-\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\boldsymbol{\Pi}\boldsymbol{D} \\ &\quad+\mathrm{vec}(\boldsymbol{F})^\dagger\boldsymbol{D}^\dagger\big((\boldsymbol{\Omega}^T\boldsymbol{\Pi}\boldsymbol{D}\mathrm{vec}(\boldsymbol{F}))^T\otimes\big(\boldsymbol{\Pi}\boldsymbol{\Omega}\big)\big)\boldsymbol{\Psi}\big(\tilde{\boldsymbol{F}}^T\otimes(\boldsymbol{I}-\boldsymbol{B})\big). \end{aligned}$$

APPENDIX E
THE PROOF OF PROPOSITION 3

As the formula of GMI in (8) is quadratic in $\boldsymbol{V}$ and no constraints apply to $\boldsymbol{V}$, taking the gradient of $I_{\mathrm{GMI}}(\boldsymbol{V},\boldsymbol{R},\boldsymbol{G})$ with respect to $\boldsymbol{V}$ and setting it to zero, yields the optimal $\boldsymbol{V}$ in (22). Inserting $\boldsymbol{V}_{\mathrm{opt}}$ back into (8) gives, after some manipulations,

$$\begin{aligned} &I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G}) \\ &\quad= K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\} \\ &\qquad+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^\dagger\big). \end{aligned} \tag{82}$$

When $\boldsymbol{P}=\boldsymbol{0}$, (82) equals

$$I_2(\boldsymbol{G}) = K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big).$$

Further, when $\boldsymbol{P}\neq\boldsymbol{0}$ the terms of $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R},\boldsymbol{G})$ in (82) related to $\boldsymbol{R}$ are

$$g(\boldsymbol{R})=2\mathcal{R}\big\{\mathrm{Tr}(\boldsymbol{P}\boldsymbol{M}\boldsymbol{R})\big\}+\mathrm{Tr}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\boldsymbol{R}\tilde{\boldsymbol{M}}\boldsymbol{R}^\dagger\big).$$

Let $\boldsymbol{r}_k$ denote the kth column of $\boldsymbol{R}$, with all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ removed, and define the column vector $\boldsymbol{r}=[\,\boldsymbol{r}_0^T,\ \boldsymbol{r}_1^T,\ \ldots,\ \boldsymbol{r}_{K-1}^T\,]^T$. Then, we have

$$\boldsymbol{r}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{R}).$$

Further, let $\boldsymbol{d}_k$ denote the kth column of the matrix $\boldsymbol{M}\boldsymbol{P}$ with all elements in rows $[\max(0,k-\nu_{\mathrm{R}}),\min(k+\nu_{\mathrm{R}},K-1)]$ removed, and define the vector $\boldsymbol{d}=[\,\boldsymbol{d}_0^T,\ \boldsymbol{d}_1^T,\ \ldots,\ \boldsymbol{d}_{K-1}^T\,]^T$. From the definition of $\boldsymbol{d}$, we have

$$\boldsymbol{d}=\boldsymbol{\Omega}\mathrm{vec}(\boldsymbol{M}\boldsymbol{P}).$$

Defining a Hermitian matrix $\hat{\boldsymbol{B}}_2$ as

$$\hat{\boldsymbol{B}}_2=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T,$$

we can write $f(\boldsymbol{R})$ as

$$g(\boldsymbol{R})=\boldsymbol{r}^\dagger\hat{\boldsymbol{B}}_2\boldsymbol{r}+2\mathcal{R}\{\boldsymbol{d}^\dagger\boldsymbol{r}\}.$$

Therefore, the optimal $\boldsymbol{r}$ is

$$\boldsymbol{r}_{\mathrm{opt}}=-\hat{\boldsymbol{B}}_2^{-1}\boldsymbol{d}. \tag{83}$$

Transferring $\boldsymbol{r}_{\mathrm{opt}}$ back into $\boldsymbol{R}_{\mathrm{opt}}$ gives the optimal $\boldsymbol{R}$ in (23), and inserting it back into $g(\boldsymbol{R})$ gives

$$g(\boldsymbol{R}_{\mathrm{opt}})=-\boldsymbol{d}^\dagger\hat{\boldsymbol{B}}_2^{-1}\boldsymbol{d}.$$

Thus, with the optimal $\boldsymbol{V}$ and $\boldsymbol{R}$, when $\boldsymbol{P}\neq\boldsymbol{0}$ the GMI equals

$$\begin{aligned} I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G}) &= K+\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)+\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big) \\ &\quad-\boldsymbol{d}^\dagger\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T\big)^{-1}\boldsymbol{d}. \end{aligned}$$

APPENDIX F
THE GRADIENT IN METHOD II FOR THE MIMO CASE

Taking the α-differential of $I_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields

$$d_{\boldsymbol{G}}(I_2) = \mathrm{vec}((\boldsymbol{I}+\boldsymbol{G})^{-1}+\boldsymbol{M})^\dagger. \tag{84}$$

Defining an $S\times S$ Hermitian matrix

$$\boldsymbol{\Phi}=\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T\big)^{-1}$$

and taking the α-differential of $\delta_2(\boldsymbol{G})$ with respect to $\boldsymbol{G}$ yields

$$\begin{aligned} d_{\boldsymbol{G}}(\delta_2) &= -(\boldsymbol{d}^T\otimes\boldsymbol{d}^\dagger)d_{\boldsymbol{G}}(\boldsymbol{\Phi}) \\ &= (\boldsymbol{d}^T\otimes\boldsymbol{d}^\dagger)(\boldsymbol{\Phi}^T\otimes\boldsymbol{\Phi})d_{\boldsymbol{G}}\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T\big) \\ &= \big((\boldsymbol{d}^T\boldsymbol{\Phi}^T)\otimes(\boldsymbol{d}^\dagger\boldsymbol{\Phi})\big)(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})d_{\boldsymbol{G}}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= \big((\boldsymbol{d}^T\boldsymbol{\Phi}^T\boldsymbol{\Omega})\otimes(\boldsymbol{d}^\dagger\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}d_{\boldsymbol{G}}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \\ &= -\big((\boldsymbol{d}^T\boldsymbol{\Phi}^T\boldsymbol{\Omega})\otimes(\boldsymbol{d}^\dagger\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-T}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big) \end{aligned} \tag{85}$$

where $\boldsymbol{\Psi}$ is defined in (81). Combining (84) and (85), we obtain

$$\begin{aligned} &d_{\boldsymbol{G}}\big(I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G})\big) \\ &\quad= d_{\boldsymbol{G}}(I_2)+d_{\boldsymbol{G}}(\delta_2) \\ &\quad= \mathrm{vec}\big((\boldsymbol{I}+\boldsymbol{G})^{-1}+\boldsymbol{M}\big)^\dagger \\ &\qquad-\big((\boldsymbol{d}^T\boldsymbol{\Phi}^T\boldsymbol{\Omega})\otimes(\boldsymbol{d}^\dagger\boldsymbol{\Phi}\boldsymbol{\Omega})\big)\boldsymbol{\Psi}\big((\boldsymbol{I}+\boldsymbol{G})^{-T}\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big). \end{aligned}$$

APPENDIX G
THE CONCAVITY PROOF OF METHOD II FOR THE MIMO CASE

When $\boldsymbol{P}=\boldsymbol{0}$, as $\log\big(\det(\boldsymbol{I}+\boldsymbol{G})\big)$ is concave [64] and $\mathrm{Tr}\big(\boldsymbol{M}(\boldsymbol{I}+\boldsymbol{G})\big)$ is linear in $\boldsymbol{G}$, the function $I_2(\boldsymbol{G})$ in (25) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}\succ\boldsymbol{0}$.

The concavity when $\boldsymbol{P}\neq\boldsymbol{0}$ can be deduced from the composition theorem in [64, Ch. 3.6]. For a positive-definite matrix $\boldsymbol{X}$, $\boldsymbol{d}^\dagger\boldsymbol{X}^{-1}\boldsymbol{d}$ is convex and non-increasing (with respect to the generalized inequality for positive-definite Hermitian matrices, see [64], [65]) for any column vector $\boldsymbol{d}$. Furthermore, since $\boldsymbol{I}+\boldsymbol{G}\succ\boldsymbol{0}$, $(\boldsymbol{I}+\boldsymbol{G})^{-1}$ is convex. Since $\tilde{\boldsymbol{M}}\prec\boldsymbol{0}$ $\boldsymbol{X}=\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes(\boldsymbol{I}+\boldsymbol{G})^{-1}\big)\boldsymbol{\Omega}^T$ is concave in $\boldsymbol{G}$. By the composition theorem, $\boldsymbol{d}^\dagger\big(\boldsymbol{\Omega}\big(\tilde{\boldsymbol{M}}^*\otimes[\boldsymbol{I}+\boldsymbol{G}]^{-1}\big)\boldsymbol{\Omega}^T\big)^{-1}\boldsymbol{d}$ is convex, and $\delta_2(\boldsymbol{G})$ is then concave. Therefore, the function $I_{\mathrm{GMI}}(\boldsymbol{V}_{\mathrm{opt}},\boldsymbol{R}_{\mathrm{opt}},\boldsymbol{G})$ in (24) is concave with respect to $\boldsymbol{G}$ whenever $\boldsymbol{I}+\boldsymbol{G}\succ\boldsymbol{0}$.

APPENDIX H
THE PROOF OF PROPOSITION 5

The Fourier series associated to the Toeplitz matrix $\boldsymbol{W}$ is

$$W(\omega) = \sum_{k=-\infty}^{\infty} w_k \exp(jk\omega)\,,$$

and the differential of $\bar{I}(W(\omega), T(\omega), F(\omega))$ in (45) with respect to w_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial \bar{I}}{\partial w_k} = &-\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|F(\omega)|^2\big(N_0+|H(\omega)|^2\big)W^*(\omega)}{1+|F(\omega)|^2}\exp(jk\omega)\,\mathrm{d}\omega\\ &+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(F^*(\omega)H(\omega)+\frac{\alpha|F(\omega)|^2H(\omega)T^*(\omega)}{1+|F(\omega)|^2}\Big)\\ &\times \exp\big(jk\omega\big)\mathrm{d}\omega.\end{aligned} \tag{86}$$

Since (86) should equal zero for all k, the optimal $W(\omega)$ is given in (50). Inserting $W_{\mathrm{opt}}(\omega)$ back into (45) yields

$$\begin{aligned}&\bar{I}(W_{\mathrm{opt}}(\omega), T(\omega), F(\omega))\\ &= 1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega\\ &\quad+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+|F(\omega)|^2\big)\\ &\qquad+\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}+M(\omega)\big(1+|F(\omega)|^2\big)\Big)\mathrm{d}\omega.\end{aligned} \tag{87}$$

where $M(\omega)$ and $\tilde{M}(\omega)$ are defined in (46) and (47), respectively.

When $\alpha=0$, the GMI in (87) equals (53) while when $0<\alpha\le 1$, the terms related to $T(\omega)$ in (87) are

$$\begin{aligned}f(T(\omega)) = &\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{F^*(\omega)T(\omega)M(\omega)\big\}\mathrm{d}\omega\\ &+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|T(\omega)F(\omega)|^2}{1+|F(\omega)|^2}\mathrm{d}\omega.\end{aligned} \tag{88}$$

As the elements of the main diagonal and the first ν lower diagonals of matrix $\boldsymbol{T}$ are constrained to zeros, we define the vector $\tilde{\boldsymbol{t}}$ that specifies the Toeplitz matrix $\boldsymbol{T}$ as

$$\tilde{\boldsymbol{t}}=[\,t_{-N_{\mathrm{T}}},\ \ldots\ t_{-1},\ t_{\nu+1},\ \ldots,\ t_{N_{\mathrm{T}}}\,],$$

and with $\phi(\omega)$ defined in (48), the Fourier series $T(\omega)$ with a finite tap-length N_{T} is

$$T(\omega) = \sum_{-N_{\mathrm{T}}\le k\le N_{\mathrm{T}},k\notin[0,\nu]} t_k\exp\big(jk\omega\big) = \tilde{\boldsymbol{t}}\phi(\omega). \tag{89}$$

Further, with $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ defined in (49), (88) can be rewritten as

$$f(T(\omega))=\tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_2\tilde{\boldsymbol{t}}^{\dagger}+2\mathcal{R}\big\{\tilde{\boldsymbol{t}}\boldsymbol{\varepsilon}_1\big\}.$$

Therefore, the optimal $\tilde{\boldsymbol{t}}$ is

$$\tilde{\boldsymbol{t}}_{\mathrm{opt}} = -\boldsymbol{\varepsilon}_1^{\dagger}\boldsymbol{\varepsilon}_2^{-1}. \tag{90}$$

Inserting $\tilde{\boldsymbol{t}}_{\mathrm{opt}}$ back into (87)-(89), the optimal $T(\omega)$ is then in (51), and $\bar{I}(W(\omega), T(\omega), F(\omega))$ with the optimal $W(\omega)$ and $T(\omega)$ is given in (52) after some manipulations.

APPENDIX I
THE PROOF OF PROPOSITION 7

The Fourier series associated to the Toeplitz matrix $\boldsymbol{V}$ is

$$V(\omega)=\sum_{k=-\infty}^{\infty} v_k\exp(jk\omega)$$

and the differential of $\bar{I}(V(\omega), R(\omega), G(\omega))$ in (57) with respect to v_k (where ω is fixed) is

$$\begin{aligned}\frac{\partial \bar{I}}{\partial v_k} = &-\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\big(N_0+|H(\omega)|^2\big)V^*(\omega)}{1+G(\omega)}\exp(jk\omega)\,\mathrm{d}\omega\\ &+\frac{1}{\pi}\int_{-\pi}^{\pi}\Big(H(\omega)+\frac{\alpha H(\omega)R^*(\omega)}{1+G(\omega)}\Big)\exp\big(jk\omega\big)\mathrm{d}\omega.\end{aligned} \tag{91}$$

Since (91) equals zero for all k, the optimal $V(\omega)$ is given in (61). Putting $V_{\mathrm{opt}}(\omega)$ in (61) back into (57) yields

$$\begin{aligned}&\bar{I}(V_{\mathrm{opt}}(\omega), R(\omega), G(\omega))\\ &= 1+\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega+\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\log\big(1+G(\omega)\big)\\ &\quad+\frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}+M(\omega)\big(1+G(\omega)\big)\Big)\mathrm{d}\omega.\end{aligned} \tag{92}$$

When $\alpha=0$, the GMI in (92) equals (64), and when $0<\alpha\le 1$, the terms of $\bar{I}(V_{\mathrm{opt}}(\omega), R(\omega), G(\omega))$ related to $R(\omega)$ in (92) are

$$\begin{aligned}g(R(\omega)) = &\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\mathcal{R}\big\{M(\omega)R(\omega)\big\}\mathrm{d}\omega\\ &+\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\tilde{M}(\omega)|R(\omega)|^2}{1+G(\omega)}\mathrm{d}\omega.\end{aligned} \tag{93}$$

Defining the vector $\tilde{\boldsymbol{r}}$ that specifies the Toeplitz matrix $\boldsymbol{R}$ as

$$\tilde{\boldsymbol{r}}=[\,r_{-N_{\mathrm{R}}},\ \ldots,\ r_{-\nu_{\mathrm{R}}-1},\ r_{\nu_{\mathrm{R}}+1},\ \ldots,\ r_{N_{\mathrm{R}}}\,],$$

and with $\psi(\omega)$ defined in (59), the Fourier series $R(\omega)$ with a finite tap-length N_{R} is

$$R(\omega) = \sum_{-N_{\mathrm{R}}\le k\le N_{\mathrm{R}},k\notin[-\nu_{\mathrm{R}},\nu_{\mathrm{R}}]} r_k\exp\big(jk\omega\big) = \tilde{\boldsymbol{r}}\psi(\omega) \tag{94}$$

where $2\nu_{\mathrm{R}}+1$ is the band-size that $\boldsymbol{R}$ is constrained to zero. With $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ defined in (60), (93) can be written as

$$g(R(\omega))=\tilde{\boldsymbol{r}}\boldsymbol{\zeta}_2\tilde{\boldsymbol{r}}^{\dagger}+2\mathcal{R}\left\{\tilde{\boldsymbol{r}}\boldsymbol{\zeta}_1\right\}.$$

Therefore, the optimal $\tilde{\boldsymbol{r}}$ is

$$\tilde{\boldsymbol{r}}_{\mathrm{opt}}=-\boldsymbol{\zeta}_1^{\dagger}\boldsymbol{\zeta}_2^{-1}. \tag{95}$$

This shows that $\tilde{\boldsymbol{r}}_{\mathrm{opt}}$ has Hermitian symmetry, since $G(\omega)$, $M(\omega)$, and $\tilde{M}(\omega)$ are all real-valued and $R_{\mathrm{opt}}(\omega)$ is thusly real-valued. Putting $\tilde{\boldsymbol{r}}_{\mathrm{opt}}$ back into (92)-(94), the optimal $R(\omega)$ is in (62) and $\bar{I}(V(\omega), R(\omega), G(\omega))$ with the optimal $V(\omega)$ and $R(\omega)$ is in (63), respectively.

APPENDIX J
THE CONCAVITY PROOF OF METHOD II WITH ISI CHANNELS

In order to prove that $\bar{I}(V_{\rm opt}(\omega), R_{\rm opt}(\omega), G(\omega))$ in (63) is concave with respect to $G(\omega)$, it is sufficient to prove that $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex with respect to $G(\omega)$. For a positive-definite matrix $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex and non-increasing (with respect to a generalized inequality for positive-definite Hermitian matrices) in $G(\omega)$ for any vector $\boldsymbol{\zeta}_1$ and with arbitrary finite tap-length $N_{\rm R}$. As matrix $\tilde{\boldsymbol{M}} \prec \boldsymbol{0}$, $\boldsymbol{\zeta}_2$ in (60) is concave with respect to $G(\omega)$ under the constraint that $\boldsymbol{I}+\boldsymbol{G} \succ \boldsymbol{0}$ is positive-definite. Hence, $\boldsymbol{\zeta}_1^\dagger \boldsymbol{\zeta}_2^{-1} \boldsymbol{\zeta}_1$ is convex in $G(\omega)$ by the composition theorem [64].

APPENDIX K
THE PROOF OF LEMMA 5

From Theorem 2, the optimal $\boldsymbol{G}$ in Method III satisfies

$$[(\boldsymbol{I}+\boldsymbol{G}_{\rm opt})^{-1}]_\nu = -[\hat{\boldsymbol{M}}]_\nu.$$

Note that, when $\boldsymbol{P} = \boldsymbol{0}$ Method II and III are equivalent as $\hat{\boldsymbol{M}} = \boldsymbol{M}$. Hence, in order to prove Lemma 4, it is sufficient to show that $[\hat{\boldsymbol{M}}]_\nu$ converges to $[\boldsymbol{M}]_\nu$ as $N_0 \to 0$ and ∞ in Method III. When $\boldsymbol{P} \prec \boldsymbol{I}$, $\boldsymbol{C}_k \succ \boldsymbol{0}$ in (34), and when $N_0 \to 0$,

$$\begin{aligned}
&\boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^\dagger + N_0\boldsymbol{I})^{-1}\boldsymbol{H} \\
&\quad = \boldsymbol{C}_k^{-1}(\boldsymbol{H}^\dagger\boldsymbol{H} + N_0\boldsymbol{C}_k^{-1})^{-1}\boldsymbol{H}^\dagger\boldsymbol{H} \\
&\quad = \boldsymbol{C}_k^{-1}\left(\boldsymbol{I} - N_0\boldsymbol{C}_k^{-1}(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\right) + \mathcal{O}(N_0^2). \qquad (96)
\end{aligned}$$

Therefore, with $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{C}}$ defined through (33)-(37) and using (96), it holds that

$$\begin{aligned}
\hat{\boldsymbol{W}}\boldsymbol{H} &= \boldsymbol{I} - N_0(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1} + \mathcal{O}(N_0^2), \\
\hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus\nu} = -N_0[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_{\setminus\nu} + \mathcal{O}(N_0^2). \qquad (97)
\end{aligned}$$

With (97) and $\hat{\boldsymbol{M}}$ in (40), it can be verified that

$$\lim_{N_0\to 0}[\hat{\boldsymbol{M}}/N_0]_\nu = -[(\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}]_\nu.$$

On the other hand, when $N_0 \to \infty$, from (33)-(40) we have

$$\begin{aligned}
N_0\hat{\boldsymbol{W}} &= \boldsymbol{H}^\dagger(\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}^\dagger/N_0 + \boldsymbol{I})^{-1} = \boldsymbol{H}^\dagger + \mathcal{O}(1/N_0), \\
N_0\hat{\boldsymbol{C}} &= [\hat{\boldsymbol{W}}\boldsymbol{H}]_{\setminus\nu} = [\boldsymbol{H}^\dagger\boldsymbol{H}]_{\setminus\nu} + \mathcal{O}(1/N_0). \qquad (98)
\end{aligned}$$

With (98), it can be verified that

$$\lim_{N_0\to\infty}[N_0(\boldsymbol{I}+\hat{\boldsymbol{M}})]_\nu = [\boldsymbol{H}^\dagger\boldsymbol{H}]_\nu.$$

Therefore, from (69) it shows that $[\hat{\boldsymbol{M}}]_\nu$ converges to $[\boldsymbol{M}]_\nu$ when $N_0 \to 0$ and ∞, which completes the proof.

ACKNOWLEDGMENT

This work has been included my Ph.D. dissertation [1]. This paper was presented in part at the IEEE 26th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Hong Kong, September 2015 [2].

REFERENCES

[1] S. Hu, "Channel shortening in wireless communication," Ph.D. dissertation, Dept. Elect. Inf. Technol., Lund Univ., Lund, Sweden, Dec. 2017.
[2] S. Hu and F. Rusek, "On the design of reduced state demodulators with interference cancellation for iterative receivers," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2015, pp. 981–985.
[3] N. Al-Dhahir, "FIR channel-shortening equalizers for MIMO ISI channels," *IEEE Trans. Commun.*, vol. 49, no. 2, pp. 213–218, Feb. 2001.
[4] D. D. Falconer and F. R. Magee, Jr., "Adaptive channel memory truncation for maximum likelihood sequence estimation," *Bell Syst. Tech. J.*, vol. 51, no. 9, pp. 1541–1562, Nov. 1973.
[5] S. A. Fredricsson, "Joint optimization of transmitter and receiver filters in digital PAM systems with a Viterbi detector," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 200–210, Mar. 1976.
[6] C. T. Beare, "The choice of the desired impulse response in combined linear-Viterbi algorithm equalizers," *IEEE Trans. Commun.*, vol. COM-26, pp. 1301–1307, Aug. 1978.
[7] N. Sundström, O. Edfors, P. Ödling, H. Eriksson, T. Koski, and P. O. Börjesson, "Combined linear-Viterbi equalizers—A comparative study and a minimax design," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Stockholm, Sweden, vol. 2, Jun. 1994, pp. 1263–1267.
[8] N. Al-Dhahir and J. M. Cioffi, "Efficiently computed reduced-parameter input-aided MMSE equalizers for ML detection: A unified approach," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 903–915, May 1996.
[9] M. Lagunas, A. I. Perez-Neia, and J. Vidal, "Joint beamforming and Viterbi equalizer in wireless communications," in *Proc. Conf. Rec. 31st Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 1997, pp. 915–919.
[10] S. A. Aldosari, S. A. Alshebeili, and A. M. Al-Sanie, "A new MSE approach for combined linear-Viterbi equalizers," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Tokyo, Japan, vol. 3, May 2000, pp. 1707–1711.
[11] R. Venkataramani and S. Sankaranarayanan, "Optimal channel shortening equalization for MIMO ISI channels," in *Proc. IEEE Global Telecom. (GLOBECOM)*, New Orleans, LO, USA, Dec. 2008, pp. 1–5.
[12] A. Shaheem, "Iterative detection for wireless communications," Ph.D. dissertation, School Elect., Electron. Comput. Eng., Univ. Western Australia, Perth, WA, USA, 2008.
[13] U. L. Dang, W. H. Gerstacker, and D. T. M. Slock, "Maximum SINR prefiltering for reduced-state trellis-based equalization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kyoto, Japan, Jun. 2011, pp. 1–6.
[14] R. Venkataramani and M. F. Erden. (2007). "A posteriori equivalence: A new perspective for design of optimal channel shortening equalizers." [Online]. Available: https://arxiv.org/abs/0710.3802
[15] I. Aboy-Faycal and A. Lapidoth, "On the capacity of reduced-complexity receivers for intersymbol interference channels," in *Proc. Conf. Inf. Sci. Syst. (CISS)*. Princeton, NJ, USA: Princeton Univ., Mar. 2000, pp. 263–266.
[16] D. Darsena and F. Verde, "Minimum-mean-output-energy blind adaptive channel shortening for multicarrier SIMO transceivers," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5755–5771, Dec. 2007.
[17] G. D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 3, pp. 363–378, May 1972.
[18] F. Rusek and A. Prlja, "Optimal channel shortening for MIMO and ISI channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 810–818, Feb. 2012.
[19] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.
[20] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.
[21] A. Ganti, A. Lapidoth, and İ. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.
[22] M. R. McKay, I. B. Collings, and A. M. Tulino, "Achievable sum rate of MIMO MMSE receivers: A general analytic framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 396–410, Jan. 2010.
[23] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "Gaussian codes and weighted nearest neighbor decoding in fading multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1665–1686, Aug. 2004.
[24] M. Tüchler and A. C. Singer, "Turbo equalization: An overview," *IEEE Trans. Inf. Theory.*, vol. 57, no. 2, pp. 920–952, Feb. 2011.

[25] S.-J. Lee, A. C. Singer, and N. R. Shanbhag, "Linear turbo equalization analysis via BER transfer and EXIT charts," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2883–2897, Aug. 2005.

[26] A. Shaheem, H.-J. Zepernick, and M. Caldera, "Enhanced channel shortened turbo equalization," in *Proc. Int. Conf. Adv. Technol. Commun. (ATC)*, Oct. 2008, pp. 8–11.

[27] A. Glavieux, C. Laot, and J. Labat, "Turbo equalization over a frequency selective channel," in *Proc. Int. Symp. Turbo Codes Rel. Topics*, Brest, France, Sep. 1997, pp. 96–102.

[28] R. R. Lopes and J. R. Barry, "The soft-feedback equalizer for turbo equalization of highly dispersive channels," *IEEE Trans. Commun.*, vol. 54, no. 5, pp. 783–788, May 2006.

[29] J. W. Choi, B. Shim, A. C. Singer, and N. I. Cho, "Low-complexity decoding via reduced dimension maximum-likelihood search," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1780–1793, Mar. 2010.

[30] J. W. Choi, B. Lee, and B. Shim, "Iterative group detection and decoding for large MIMO systems," *J. Commun. Netw.*, vol. 17, no. 6, pp. 609–621, Dec. 2015.

[31] J. W. Choi, A. C. Singer, J. W. Lee, and N. I. Cho, "Improved linear soft-input soft-output detection via soft feedback successive interference cancellation," *IEEE Trans. Commun.*, vol. 58, no. 3, pp. 986–996, Mar. 2010.

[32] M. Tüchler, A. C. Singer, and R. Kötter, "Minimum mean squared error equalization using *a priori* information," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 673–683, Mar. 2000.

[33] S. Hu, H. Kröll, Q. Huang, and F. Rusek, "Optimal channel shortener design for reduced- state soft-output Viterbi equalizer in single-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2568–2582, Jun. 2017.

[34] A. Berthet, R. Visoz, and P. Tortelier, "Sub-optimal turbo-detection for coded 8-PSK signals over ISI channels with application to EDGE advanced mobile system," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Sep. 2000, pp. 151–157.

[35] A. Duel-Hallen and C. Heegard, "Delayed decision-feedback sequence estimation," *IEEE Trans. Commun.*, vol. 37, no. 5, pp. 428–436, May 1989.

[36] F. Rusek, N. Al-Dhahir, and A. Gomaa, "A rate-maximizing channel-shortening detector with soft feedback side information," in *Proc. IEEE Global Telecom. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 1–6.

[37] C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, Jul. 2011.

[38] M. Witzke, S. Bäro, F. Schreckenbach, and J. Hagenauer, "Iterative detection of MIMO signals with linear detectors," in *Proc. Asilomar Conf. Signals, Syst. Comput. (ACSSC)*, Monterey, CA, USA, Nov. 2002, pp. 289–293.

[39] J. Zhang, H. Nguyen, and G. Mandyam, "LMMSE-based iterative and turbo equalization methods for CDMA downlink channels," in *Proc. IEEE 6th Workshop Signal Process. Adv. Wireless Commun.*, Jun. 2005, pp. 231–235.

[40] R. M. Gray, "Toeplitz and circulant matrices: A review," *Found. Trends Commun. Inf. Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[41] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[42] J. Hagenauer, "Source-controlled channel decoding," *IEEE Trans. Commun.*, vol. 43, no. 9, pp. 2449–2457, Sep. 1995.

[43] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[44] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389–399, Mar. 2003.

[45] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

[46] F. Rusek and D. Fertonani, "Bounds on the information rate of intersymbol interference channels based on mismatched receivers," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1470–1482, Mar. 2012.

[47] G. Ungerboeck, "Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems," *IEEE Trans. Commun.*, vol. COM-22, no. 5, pp. 624–636, May 1974.

[48] F. Rusek, G. Colavolpe, C. E. W. Sundberg, "40 years with the Ungerboeck model: A look at its potentialities [Lecture Notes]," *Signal Process. Mag.*, vol. 32, no. 3, pp. 156–161, May 2015.

[49] F. Rusek, M. Loncar, and A. Prlja, "A comparison of ungerboeck and Forney models for reduced-complexity ISI equalization," in *Proc. IEEE Global Telecom. (GLOBECOM)*, Washington, DC, USA, Dec. 2007, pp. 1431–1436.

[50] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[51] G. L. Turin, "An introduction to digitial matched filters," *Proc. IEEE*, vol. 64, no. 7, pp. 1092–1112, Jul. 1976.

[52] A. Kavčić and J. M. F. Moura, "Matrices with banded inverses: inversion algorithms and factorization of Gauss–Markov processes," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1495–1509, Jul. 2000.

[53] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 720–722, Aug. 2005.

[54] O. Edfors, M. Sandell, J. J. van de Beek, S. K. Wilson, and P. O. Börjesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 931–939, Jul. 1998.

[55] W. Hirt, "Capacity and information rates of discrete-time channels with memory," Ph.D. dissertation, Inst. Signal Inf. Process., Swiss Federal Inst. Technol., Zürich, Switzerland, 1988.

[56] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. Berkeley, CA, USA: Univ. California Press, 1958.

[57] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed. New York, NY, USA: McGraw-Hill, 2008.

[58] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[59] S. ten Brink, "Convergence of iterative decoding," *Electron. Lett.*, vol. 35, no. 10, pp. 806–808, May 1999.

[60] *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding, Release 12*, document TS 36.212, 3GPP, Mar. 2015.

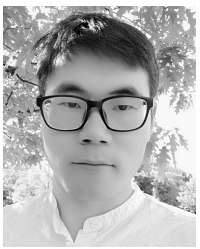
[61] F. Rusek and O. Edfors, "An information theoretic charachterization of channel shortening receivers," in *Proc. 47th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2013, pp. 2108–2112.

[62] J. R. Magnus, "On the concept of matrix derivative," *J. Multivariate Anal.*, vol. 101, no. 9, pp. 2200–2206, Oct. 2001.

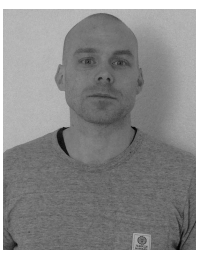
[63] P. L. Fackler, *Notes on Matrix Calculus*. Raleigh, NC, USA: North Carolina State Univ., Sep. 2005.

[64] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[65] C. Davis, "Notions generalizing convexity for functions defined on spaces of matrices," in *Proc. Symp. Pure Math.*, vol. 7. Providence, RI, USA: AMS, 1963, pp. 187–201.



SHA HU (S'15–M'18) was born in Hubei, China, in 1985. He received the Ph.D. degree in electrical engineering from Lund University, Lund, Sweden, in 2018, and the M.S. and B.S. degrees in pure mathematics from Wuhan University, China, in 2008 and 2006, respectively. Since 2008, he has been with Huawei Technologies, where he was involved in baseband algorithm research. He has co-authored 18 patents in these fields. Since 2015, he joined the Department of Electrical and Information Technology, Lund University. His research interests include applied information theory, signal processing, multi-input multi-output detection, channel shortening, and precoder design.



FREDRIK RUSEK was born in Lund, Sweden, in 1978. He received the M.S. and Ph.D. degrees in electrical engineering from Lund University, Lund, in 2003 and 2007, respectively. Since 2012, he has been an Associate Professor with the Department of Electrical and Information Technology, Lund Institute of Technology. His research interests include modulation theory, equalization, wireless communications, and applied information theory.