Feedback Capacity of MIMO Gaussian Channels

Oron Sabag, *Member, IEEE*, Victoria Kostina, *Senior Member, IEEE*, and Babak Hassibi, *Member, IEEE*

***Abstract*— Finding a computable expression for the feedback capacity of channels with colored Gaussian, additive noise is a long standing open problem. In this paper, we solve this problem in the scenario where the channel has multiple inputs and multiple outputs (MIMO) and the noise process is generated as the output of a time-invariant state-space model. Our main result is a computable expression for the feedback capacity in terms of a finite-dimensional convex optimization. The solution to the feedback capacity problem is obtained by formulating the finite-block counterpart of the capacity problem as a *sequential convex optimization problem* which leads in turn to a single-letter upper bound. This converse derivation integrates tools and ideas from information theory, control, filtering and convex optimization. A tight lower bound is realized by optimizing over a family of time-invariant policies thus showing that time-invariant inputs are optimal even when the noise process may not be stationary. The optimal time-invariant policy is used to construct a capacity-achieving and simple coding scheme for scalar channels, and its analysis reveals an interesting relation between a smoothing problem and the feedback capacity expression.**

***Index Terms*— Multiple inputs and multiple outputs (MIMO), channel capacity, feedback, colored noise, additive white Gaussian noise (AWGN), convex optimization.**

I. Introduction

We consider the feedback capacity of a multiple-input multiple-output (MIMO) Gaussian channel

$$\mathbf{y}_i = \Lambda\, \mathbf{x}_i + \mathbf{z}_i, \tag{1}$$

where $\Lambda \in \mathbb{R}^{p\times m}$ is a deterministic matrix, $\mathbf{y}_i$ is the channel output and $\mathbf{x}_i$ is the channel input. The noise, $\mathbf{z}_i$, is a colored Gaussian process generated by a vector state-space model (a hidden Markov model)

$$\begin{aligned} \mathbf{s}_{i+1} &= F\,\mathbf{s}_i + G\,\mathbf{w}_i \\ \mathbf{z}_i &= H\,\mathbf{s}_i + \mathbf{v}_i, \end{aligned} \tag{2}$$

where the sequence $(\mathbf{w}_i, \mathbf{v}_i)$ is i.i.d. with Gaussian distribution. Our assumptions on the state-space are mild and include for instance non-stationary noise processes (when the spectral radius of F is greater than 1). Particular realizations of the state-space reveal well-known random processes such as the auto-regressive moving-average (ARMA) noise process.

Most related to our setting is the framework for channels with general additive Gaussian noise processes by Cover and Pombra [2]. They showed that the feedback capacity is equal to the limit of

$$C_n(P) = \max_{K_V \succeq 0, B} \frac{1}{2n} \log \frac{\det(K_V + (I+B)K_n^Z(I+B)^T)}{\det K_n^Z}, \tag{3}$$

where K_n^Z is the covariance of the Gaussian noise, and the maximum is subject to strictly-causal linear operators B (lower-triangular matrices) and pairs (K_V, B) that satisfy the power constraint $\mathrm{Tr}(K_V + BK_n^Z B^T) \le nP$. Their general methodology applies to arbitrary Gaussian processes, and can be extended to MIMO channels and results in a formula that is similar to (3), but the computation of such expressions remains non-trivial. In this paper, we show that imposing a state-space structure on the Gaussian noise leads to a computable characterization of the infinite-limit of the optimization problem (3).

Our main result is a computable expression for the feedback capacity, formulated as a finite-dimensional convex optimization problem. The optimization is a maximal determinant optimization problem subject to linear matrix inequalities (LMIs) constraints, a class of convex optimization problems that often appear in the control literature [3], [4], [5], [6] and recently also in information theory [7], [8], [9], [10]. The LMIs are interpretable, and one of the LMIs corresponds to a tight relaxation of a Riccati equation. Several aspects of the feedback capacity solution such as computability, comparison with non-feedback rates, and optimal inputs distribution are discusses by studying the capacity of the moving-average (MA) and the auto-regressive (AR) noise processes.

The literature on the feedback capacity of scalar Gaussian channel is rich, e.g. [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], and the focus here is on works most related to ours (a detailed survey can be found in [21]). In [22], an explicit lower bound for ARMA(1,1) noise was derived. The lower bound was shown to be optimal in [21] and [23][1]. In [21], it is also shown that stationary channel input processes achieve the feedback capacity when the noise is stationary. Based on this fundamental result, [21] studies the special case of our channel in (1)-(2) where the channel is scalar, the noise is stable, and the hidden state is available to the encoder ($\mathbf{w}_i = \mathbf{v}_i$), and formulated its capacity as a finite-dimensional, non-convex optimization problem. In contrast, we provide a convex optimization for the feedback capacity of the general

Manuscript received 3 July 2021; revised 29 March 2023; accepted 1 June 2023. Date of publication 7 June 2023; date of current version 15 September 2023. The work of Victoria Kostina was supported in part by the NSF through under Grant CCF-1751356 and Grant CCF-1956386. An earlier version of this paper was presented in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9518088]. *(Corresponding author: Oron Sabag.)*

Oron Sabag was with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA. He is now with the Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel (e-mail: oron.sabag@mail.huji.ac.il).

Victoria Kostina and Babak Hassibi are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: vkostina@caltech.edu; hassibi@caltech.edu).

Communicated by S. Yang, Associate Editor for Communications.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3283570.

Digital Object Identifier 10.1109/TIT.2023.3283570

[1]Recently, [24] identified that the proof of [21, Lemma 4.4] is invalid, which implies that the conjecture in [22] is only proven for the MA(1) noise [23]. More details in Section IV-A.

channel in (1)-(2) under mild conditions (see Section II below). In [10], a change of variable to the non-convex optimization in [21], combined with the interesting idea of using LMIs, showed that the capacity can be formulated as a convex optimization problem. However, the change of variable relied on an erroneous claim (see Remark 1). Our paper studies colored Gaussian noise described by the general state-space model where the hidden state of the noise may or may not be available to the encoder, the channel may be scalar or vector (MIMO), and the noise may be stationary or not. We express the feedback capacity as a convex optimization problem in this general setting. A recent conference paper also studies MIMO channels, and extends the convex optimization in [15] to MIMO channels with ISI [24]. The intersection of [25] and our setting is a MIMO channel with a stable, colored Gaussian noise, and the capacity results in the current paper (initially published as [1]) and [25] were developed independently and published concurrently. Each work considers a different extension of the MIMO channel with stable noise: [25] studies channels with ISI, whereas we study colored Gaussian noise that can be either stationary or non-stationary. A major technical contribution in our work is that we show the optimality of stationary inputs for non-stationary noise processes. This fact and its proof may be of independent interest since it does not rely on the frequency-based methods of [21], and it can be even used to generalize the capacity results in this paper [26].

The starting point of our derivations is the general Cover-Pombra characterization in (3), and we develop a *time-domain* methodology to provide a computable expression for the feedback capacity. We derive a novel formulation of the $n-$letter capacity in (3) as a *sequential convex optimization problem* (SCOP). In particular, we formulate an optimization problem whose decision variable is a sequence of length n, where at each time fixed-dimensional matrices are optimized. In the SCOP formulation, the LMI constraints have a sequential nature and should depend on two consecutive times only. This sequential property combined with the convexity of the problem is the key to obtain a single-letter upper bound for the limit of the n-letter capacity in (3). For the lower bound, a family of time-invariant channel inputs distributions is optimized and is shown to achieve the upper bound. An outcome of our derivation is a new methodology to show that time-invariant inputs are sufficient to achieve the feedback capacity even when noise may be non-stationary.

An optimal time-invariant policy can be computed directly from the feedback capacity convex optimization. Using this policy, we also construct an explicit coding scheme that achieves the feedback capacity for scalar channels. The derived scheme generalizes the coding proposed in [21], and simplifies its encoding by showing that the message can be encoded as a scalar rather than the multi-dimensional variant proposed in [21]. We also derive an explicit decoding rule by studying a related smoothing problem [27]. The analysis of the smoothing problem reveals an interesting relation between the volume reduction of its error covariance and the capacity solution. That analysis is performed for the general case of a MIMO channel, and a possible MIMO scheme is discussed in Section VII.

The rest of the paper is organized as follows. In Section II, we present the setting and the preliminaries. Section III includes our main result on the feedback capacity of the MIMO Gaussian channel and several examples. In Section IV, we present the optimal inputs distribution and the capacity-achieving coding scheme. In Section V, the main ideas and the technical lemmas to prove our main result are presented while their detailed proofs appear in Section VI.

II. The Setting and Preliminaries

This section includes the communication setting. We also present the Kalman filter and the Riccati equation that are required for the presentation of the main result.

A. The Setting

We consider a MIMO additive Gaussian channel

$$\mathbf{y}_i = \Lambda\,\mathbf{x}_i + \mathbf{z}_i, \tag{4}$$

where the channel input is $\mathbf{x}_i \in \mathbb{R}^m$, $\mathbf{y}_i \in \mathbb{R}^p$ is the channel output, the additive noise is $\mathbf{z}_i \in \mathbb{R}^p$, and $\Lambda \in \mathbb{R}^{p\times m}$ is a fixed known matrix. The encoder has access to noiseless, strictly-causal feedback so that the input $\mathbf{x}_i$ is a function of the message and all previous channel outputs $\mathbf{y}^{t-1} := \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$. For a fixed blocklength n, the channel input should satisfy the average power constraint $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{x}_i^T\,\mathbf{x}_i] \le P$. Definitions of the average probability of error, achievable rates and the feedback capacity are standard and can be found in [21], for instance. The feedback capacity with a power constraint P is denoted by $C_{fb}(P)$.

In the case of MIMO channels, the capacity can be expressed as the multi-letter expression in (3) by modifying all the matrices to their corresponding block matrices with appropriate dimensions. An equivalent characterization of the n-letter objective in (3) is the directed information $I(\mathbf{x}^n \to \mathbf{y}^n)$ that characterizes the feedback capacity of point to point channels [28], [29], [30], [31].

The additive noise is a colored Gaussian process generated as the output of the state-space:

$$\begin{aligned}\mathbf{s}_{i+1} &= F\,\mathbf{s}_i + G\,\mathbf{w}_i\\ \mathbf{z}_i &= H\,\mathbf{s}_i + \mathbf{v}_i,\end{aligned} \tag{5}$$

where $\mathbf{s}_i \in \mathbb{R}^n$, $\mathbf{w}_i \sim N(0, W)$ and $\mathbf{v}_i \sim N(0, V)$ are i.i.d. sequences with $\mathbb{E}[\mathbf{w}_i\,\mathbf{v}_i^T] = L$, and are independent of the initial state $\mathbf{s}_1 \sim N(0, \Sigma_1)$. Note that, due to the feedback, the encoder has a strictly causal access to the noise $\mathbf{z}_i$, but not to the hidden state $\mathbf{s}_i$. For this case, we can use Kalman filtering to write the state-space (5) in an observer form [32]. This pre-Kalman filtering step for the state-space (5) allows one to define a new channel state that is available to the encoder.

B. The Kalman Filter and the Innovations Process

The Kalman filter is a simple, recursive method to compute the maximum likelihood estimate of the hidden state $\mathbf{s}_i$ based on the measurements $\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}$. The predicted-estimate of the state and its error covariance are denoted by

$$\begin{aligned}\hat{\mathbf{s}}_i &= \mathbb{E}[\mathbf{s}_i \,|\, \mathbf{z}^{i-1}]\\ \Sigma_i &= \mathbf{cov}(\mathbf{s}_i - \hat{\mathbf{s}}_i).\end{aligned} \tag{6}$$

The Kalman filter is given by the recursion

$$\hat{\mathbf{s}}_{i+1} = F\,\hat{\mathbf{s}}_i + K_{p,i}(\mathbf{z}_i - H\,\hat{\mathbf{s}}_i) \tag{7}$$

with the initialization $\hat{\mathbf{s}}_1 = 0$, and the constants are

$$\begin{aligned} K_{p,i} &= (F\Sigma_i H^T + GL)\Psi_i^{-1} \\ \Psi_i &= H\Sigma_i H^T + V, \end{aligned} \tag{8}$$

where the error covariance Σ_i is described by the Riccati recursion

$$\Sigma_{i+1} = F\Sigma_i F^T + GWG^T - K_{p,i}\Psi_i K_{p,i}^T \tag{9}$$

with the initial condition $\Sigma_1 \succeq 0$. The estimate in (7) is updated using the *innovations process*, defined as $\mathbf{e}_i = \mathbf{z}_i - H\,\hat{\mathbf{s}}_i$, and is distributed according to $N(0, \Psi_i)$. It can be easily shown that the innovation $\mathbf{e}_i$ is orthogonal (statistically independent) of the previous instances of the measurements $\mathbf{z}^{i-1}$ [33]. Thus, we can write a new equivalent channel as

$$\begin{aligned} \hat{\mathbf{s}}_{i+1} &= F\,\hat{\mathbf{s}}_i + K_{p,i}\,\mathbf{e}_i \\ \mathbf{y}_i &= H\,\hat{\mathbf{s}}_i + \Lambda\,\mathbf{x}_i + \mathbf{e}_i \end{aligned} \tag{10}$$

where $\hat{\mathbf{s}}_i$ plays the role of the channel state and is available to the encoder. Note that this is a valid channel due to the Markov chain $(\hat{\mathbf{s}}_{i+1}, \mathbf{y}_i) - (\mathbf{x}_i, \hat{\mathbf{s}}_i) - (\mathbf{x}^{i-1}, \mathbf{y}^{i-1}, \hat{\mathbf{s}}^{i-1}, m)$.

The innovations process also characterizes the entropy rate of Gaussian random processes as

$$\begin{aligned} \frac{1}{n}h(\mathbf{z}^n) &= \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{z}_i \,|\, \mathbf{z}^{i-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{e}_i) \\ &= \frac{1}{2n}\sum_{i=1}^{n} \log\det(\Psi_i) + \frac{1}{2n}\log(2\pi e)^d. \end{aligned} \tag{11}$$

In (8), it is assumed that $\Psi_i \succ 0$ for all i. This is a natural assumption since otherwise the capacity is infinite. Namely, if Ψ_i is only positive semidefinite, a coordinate in the noise vector $\mathbf{z}_i$ is a deterministic function of the past noise instances $\mathbf{z}^{i-1}$. Building an infinite-rate scheme is straightforward: the encoder transmits $\mathbf{x}_j = 0$ so that $\mathbf{y}_j = \mathbf{z}_j$ for $j \le i-1$. Then, by having $\mathbf{z}^{i-1}$, the encoder and the decoder know a coordinate of $\mathbf{z}_i$, and can communicate an infinite number of bits on this vector coordinate (assuming the image of Λ is not degenerated at this particular direction).

C. The Riccati Equation

Consider the matrix function

$$f(\Sigma) = F\Sigma F^T - \Sigma + GWG^T - K_p(\Sigma)\Psi(\Sigma)K_p^T(\Sigma), \tag{12}$$

where $K_p(\Sigma) = (F\Sigma H^T + GL)\Psi(\Sigma)^{-1}$ and $\Psi(\Sigma) = H\Sigma H^T + V$. The Riccati equation is defined as $f(\Sigma) = 0$. The stabilizing solution to the Riccati equation (if exists) solves $f(\Sigma) = 0$, and is the unique solution such that its corresponding closed-loop system $F - K_p(\Sigma)H$ is stable. In the rest of the paper, we refer to

$$\begin{aligned} K_p &= (F\Sigma H^T + GL)\Psi^{-1} \\ \Psi &= H\Sigma H^T + V \end{aligned} \tag{13}$$

as the constants evaluated at the stabilizing solution. The corresponding time-invariant Kalman filter is

$$\hat{\mathbf{s}}_{i+1} = F\,\hat{\mathbf{s}}_i + K_p(\mathbf{z}_i - H\,\hat{\mathbf{s}}_i). \tag{14}$$

We move on to present assumptions on the state-space model. The stability of F determines the stationarity of the noise process.

Definition 1: The matrix F is stable if its spectral radius satisfies $\rho(F) < 1$.

Without further assumptions, our results hold for the stationary case, i.e., when F is stable (and $L = 0$). Thus, a reader whose interest is limited to the stationary case may skip the following assumptions.

Assumption 1: The pair (F, H) is detectable. That is, there exists a matrix K such that $\rho(F - KH) < 1$.

Assumption 2: The pair (F_s, W_s) is controllable on the unit circle where $F_s \triangleq F - GLV^{-1}H$ and $W_s \triangleq W - GLV^{-1}L^TG^T$. That is, for any x and λ such that $xF_s = x\lambda$, if $|\lambda| = 1$, then $xW_s^{1/2} \neq 0$.

Assumption 1 asserts that all eigenvectors of F that have unstable eigenvalues (outside the unit circle) can be observed via the matrix H. Indeed, without loss of generality, it can even be assumed the pair (F, H) is observable (for all eigenvectors) since the unobserved eigenvectors have no effect on the channel noise. Assumptions 1 and 2 are sufficient and necessary conditions for the existence of the unique stabilizing solution to the Riccati equation in (12).

We further assume that the initial covariance matrix Σ_1 converges to the stabilizing solution. Advanced discussions on convergence of Riccati recursions can be found in [27, App. E], and here we aim to provide several alternatives in order to obtain a general framework. The first condition is stabilizability of (F_s, W_s). That is, for any x and λ such that $xF_s = x\lambda$, if $|\lambda| \ge 1$, then $xW_s^{1/2} \neq 0$. This condition guarantees that the stabilizing solution is the only positive semidefinite solution to the Riccati equation, which implies that *any initial state covariance* converges to the stabilizing solution. Another useful condition is $\Sigma_1 \succeq \overline{\Sigma}$ where $\overline{\Sigma}$ is the stabilizing solution. Beyond these two sufficient conditions, in simple cases (such as the moving average noise in Section III-A), the convergence can be verified manually. With these assumptions, there is no loss of generality in assuming that the initial covariance Σ_1 is equal to the stabilizing solution of the Riccati equation in (12) by letting $\mathbf{x}_i = 0$ before the transmission begins.

III. Main Result and Discussion

In this section we present the feedback capacity of the MIMO channel and its particularization to scalar channels. We discuss different aspects of our main results via several examples. The following is our main result.

Theorem 1 (Feedback capacity of MIMO channels): The feedback capacity of the MIMO Gaussian channel in (4)-(5) is given by the convex optimization

$$\begin{aligned} C_{fb}(P) = \max_{\Pi,\hat{\Sigma},\Gamma} \; & \frac{1}{2}\log\det(\Psi_Y) - \frac{1}{2}\log\det(\Psi) \\ \text{s.t.} \quad & \Psi_Y = \Lambda\Pi\Lambda^T + H\hat{\Sigma}H^T + \Lambda\Gamma H^T + H\Gamma^T\Lambda^T + \Psi \end{aligned}$$

$$\begin{pmatrix} \Pi & \Gamma \\ \Gamma^T & \hat{\Sigma} \end{pmatrix} \succeq 0, \quad \mathbf{Tr}(\Pi) \leq P,$$
$$\begin{pmatrix} F\hat{\Sigma}F^T + K_p\Psi K_p^T - \hat{\Sigma} & F(\Gamma^T\Lambda^T + \hat{\Sigma}H^T) + K_p\Psi \\ (\Lambda\Gamma + H\hat{\Sigma})F^T + \Psi K_p^T & \Psi_Y \end{pmatrix} \succeq 0, \tag{15}$$

where K_p and Ψ are constants given in (13), and the optimization variables are the matrices $\Pi \in \mathbb{R}^{m\times m}$, $\hat{\Sigma} \in \mathbb{R}^{n\times n}$, and $\Gamma \in \mathbb{R}^{m\times n}$.

Note that the first LMI constraint implies that the optimization variables Π and $\hat{\Sigma}$ are positive semidefinite. The objective structure is a difference between the entropy rates of the channel outputs process and that of the noise process. Note that the objective is a concave function of the decision variables since Ψ_Y is a linear function the decision variables, while Ψ is a constant. The concavity of the objective and the linear constraints show that the optimization problem is a convex optimization that can be computed with standard software e.g. [34].

In Section IV-A, the decision variables will be given a straightforward interpretation by showing that they induce a time-invariant, optimal inputs distribution. Here, we briefly remark on the LMIs in (15). The decision variable Π corresponds to the covariance of the channel input so that the first LMI in (15) is a verification that it forms a valid covariance matrix with a correlated variable whose covariance matrix is $\hat{\Sigma}$. For the second LMI in (15), its Schur complement implies the Riccati inequality

$$\hat{\Sigma} \preceq F\hat{\Sigma}F^T + K_p\Psi K_p^T - K_Y\Psi_Y K_Y^T, \tag{16}$$

with $K_Y = (F(\Gamma^T\Lambda^T + \hat{\Sigma}H^T) + K_p\Psi)\Psi_Y^{-1}$. In Lemma 7 (Section V), it is shown that there always exist optimal decision variables $(\Pi, \hat{\Sigma}, \Gamma)$ that satisfy the Riccati inequality (16) with equality, i.e., it is a Riccati equation. This reveals that the origin for explicit capacity formulae expressed as function of roots to some polynomials [21], [22], [23] is the Riccati equation. We demonstrate this interesting fact in Section III-A for the MA noise process.

If the channel outputs, inputs, and the additive noise are scalars, but the hidden state of the noise $\mathbf{s}_i$ is possibly a vector, the capacity in Theorem 1 can be simplified as follows.

Theorem 2 (Feedback capacity of scalar channels): The feedback capacity of the scalar Gaussian channel (4)-(5) with $\Lambda = 1$ is given by the convex optimization problem

$$C_{fb}(P) = \max_{\hat{\Sigma},\Gamma} \frac{1}{2}\log\left(1 + \frac{P + H\hat{\Sigma}H^T + \Gamma H^T + H\Gamma^T}{\Psi}\right)$$
$$\text{s.t.} \quad \begin{pmatrix} P & \Gamma \\ \Gamma^T & \hat{\Sigma} \end{pmatrix} \succeq 0,$$
$$\begin{pmatrix} F\hat{\Sigma}F^T + K_p\Psi K_p^T - \hat{\Sigma} & F\Gamma^T + F\hat{\Sigma}H^T + K_p\Psi \\ \Gamma F^T + H\hat{\Sigma}F^T + \Psi K_p^T & \Psi_Y \end{pmatrix} \succeq 0,$$
$$\Psi_Y = P + H\hat{\Sigma}H^T + \Gamma H^T + H\Gamma^T + \Psi, \tag{17}$$

where K_p and Ψ are constants given in (13).

Choosing $H = 0$ in (17) recovers the capacity formula of an additive white Gaussian noise (AWGN) channel

$$C_{fb}(P) = \frac{1}{2}\log\left(1 + \frac{P}{V}\right).$$

Remark 1: The state-space studied in [10] can be recovered from the setting in Theorem 2 with $W = V = L = 1$ and a stable F. It is also assumed that $\Sigma = 0$ implying that $K_p = G, \Psi = 1$. In this case, the capacity expression in (17) and that in [10, Th. 4] are almost in full agreement. In particular, they write the optimization problem with a supremum, and there is a difference in the sign of the first LMI in (17) which reads as a strict LMI ($\succ$) in [10]. The reasoning for their strict LMI follows from an erroneous claim after their Theorem 3 that the i.i.d. component of the optimal policy should have a positive variance at all times (the policy structure appears below in (27)). This claim is utilized to show their argument on the invertibility of $\hat{\Sigma}$. In Section III-A we show, for the MA noise, that the i.i.d. component of the optimal policy is zero. In particular, we show that the optimum is achieved on the boundary of the LMI, i.e., the optimal variables induce a singular matrix in the LMI (see the proof of Theorem 3 for details).

A. Moving Average (MA) Noise

In this section, we study a scalar channel with the MA noise process

$$z_i = w_i + \alpha w_{i-1}, \tag{18}$$

with i.i.d. $w_i \sim N(0,1)$ and $\alpha \in \mathbb{R}$. In [23], the feedback capacity of the MA noise with $|\alpha| \leq 1$ was shown to be equal to

$$C_{fb}(P) = -\log x_0, \tag{19}$$

where x_0 is the unique positive root of

$$Px^2 = (1 - |\alpha|x)^2(1 - x^2). \tag{20}$$

The MA noise can be realized by the state space (5) with $F = 0, H = \alpha, G = W = V = L = 1$. We derive here the feedback capacity for all α.

Theorem 3 (Moving-average noise): The feedback capacity of the Gaussian channel with MA noise is

$$C_{fb}(\alpha, P) = \frac{1}{2}\log(1 + \mathbf{SNR}), \tag{21}$$

where $\mathbf{SNR}$ is the maximal positive root of the polynomial

$$\mathbf{SNR} = \begin{cases} \left(\sqrt{P} + |\alpha|\sqrt{\frac{\mathbf{SNR}}{1+\mathbf{SNR}}}\right)^2 & \text{if } |\alpha| \leq 1 \\ |\alpha|^{-2}\left(\sqrt{P} + \sqrt{\frac{\mathbf{SNR}}{1+\mathbf{SNR}}}\right)^2 & \text{if } |\alpha| > 1, \Sigma_1 > 0. \end{cases} \tag{22}$$

The capacity expression in (21)-(22) for $|\alpha| \leq 1$ coincides with the feedback capacity expression in (19).

For $|\alpha| \leq 1$, the feedback capacity is independent of the initial state covariance Σ_1 but, for $|\alpha| > 1$, we assume $\Sigma_1 \neq 0$. This condition is made to avoid the singularity that $\Sigma_i = 0$ for all i, and does not converge to the stabilizing solution of the Riccati equation. If $|\alpha| > 1$ and $\Sigma_1 = 0$ (i.e., the initial state is known), the capacity can still be computed but with the solution to the first polynomial in (22).

To compare the capacity expression in Theorem 3 with (19) when $|\alpha| \leq 1$, one can use the change of variable

$x^2 = (1+\mathbf{SNR})^{-1}$ to write (21) as $C_{fb}(P) = -\log x_0$ where x_0 solves $\frac{1-x^2}{x^2} = \left(\sqrt{P} + |\alpha|\sqrt{1-x^2}\right)^2$. It is interesting to note that the latter polynomial and (20) are different. However, the second part of Theorem 3 confirms that the feedback capacities are in agreement for $|\alpha| \leq 1$ by showing that their positive roots coincide. We proceed to prove Theorem 3.

Proof of Theorem 3: We compute the capacity expressions in Theorem 2, and then verify thereafter that the required conditions are met.

The Schur complement of $\begin{pmatrix} P & \Gamma \\ \Gamma^T & \hat{\Sigma} \end{pmatrix} \succeq 0$ evaluated in the optimal variables is shown to be zero using contradiction. Assume that $P - \Gamma^2\hat{\Sigma}^{-1} = p$ for some $p > 0$. Then, we can replace Γ with $\Gamma' = \Gamma(1 + \Gamma^{-2}p\hat{\Sigma})^{1/2}$ to obtain a larger objective. The Riccati LMI can be verified to be satisfied with this replacement. This step shows that the LMI cannot be strict at the optimum. For the other LMI in Theorem 2, a similar reasoning shows that the Schur complement of the Riccati LMI (16) can be achieved with equality (see Lemma 7), and the Schur complement simplifies to the Riccati equation $\hat{\Sigma} = K_p\Psi K_p - (\Psi K_p)^2\Psi_Y^{-1}$.

To derive the capacity expression in Theorem 2, we compute the Riccati constants K_p and Ψ from the stabilizing solution of the Riccati equation in (12). The Riccati equation has two solutions $\Sigma = 0, 1 - \frac{1}{\alpha^2}$. For $|\alpha| < 1$, the stabilizing solution is $\Sigma = 0$ which implies $K_p = \Psi = 1$, while for $|\alpha| > 1$, the stabilizing solution is $\Sigma = 1 - \frac{1}{\alpha^2}$ which implies $\Psi = \alpha^2$ and $K_p = \alpha^{-2}$. We note that in both cases $K_p\Psi = 1$ so that the Riccati equation above simplifies to $\hat{\Sigma} = K_p - \Psi_Y^{-1}$ where K_p is either 1 or α^{-2}. The decoder's innovation can be written as

$$\begin{aligned}\Psi_Y &= \Psi + P + \alpha^2\hat{\Sigma} + 2|\alpha|\sqrt{P\hat{\Sigma}} \\ &= \Psi + \left(\sqrt{P} + |\alpha|\sqrt{K_p - \Psi_Y^{-1}}\right)^2, \end{aligned} \tag{23}$$

where the sign of Γ was chosen to maximize Ψ_Y. We denote $\Psi_Y\Psi^{-1} = 1 + \mathbf{SNR}$ and substitute the latter in both sides of (23) to obtain the fixed-point equations in (22).

We verify the conditions for Theorem 2. The pair $(F = 0, H = \alpha)$ is detectable (Assumption 1) for all α, and $(F_s, W_s) = (-\alpha, 0)$ is controllable on the unit-circle for all $|\alpha| \neq 1$ (Assumption 2). Thus, for $|\alpha| \neq 1$, the stabilizing solution for the Riccati equation exists and is equal to $\Sigma = \max\{0, 1-\frac{1}{\alpha^2}\}$. It is easy to check that the Riccati recursion in (9), $\Sigma_{i+1} = 1 - \frac{1}{1+\alpha^2\Sigma_i}$, converges to the stabilizing solution unless $|\alpha| > 1$ and $\Sigma_1 = 0$.

If $|\alpha| = 1$, the only solution to the Riccati equation is $\Sigma = 0$, but it is not a stabilizing solution (it is the maximal solution). Although Theorem 1 concerns noise processes whose Riccati equations have stabilizing solutions, the upper bound extends to non-stabilizing solutions as well. For the particular instance of the MA noise with $|\alpha| = 1$, we verified that the lower bound in Lemma 6 holds as well.

Finally, we show the equivalence of our capacity expression and (19), for $|\alpha| < 1$, by proving that the positive roots of (20) and $\frac{1-x^2}{x^2} = \left(\sqrt{P} + |\alpha|\sqrt{1-x^2}\right)^2$ coincide. The positive root of (20) satisfies $\frac{1-x_0^2}{x_0^2} = \frac{P}{(1-|\alpha|x_0)^2}$ and $\sqrt{1-x_0^2} = \frac{\sqrt{P}x_0}{1-|\alpha|x_0}$, and by substituting these equations into the second polynomial, we get

$$\begin{aligned}&\frac{1-x^2}{x^2} - \left(\sqrt{P} + |\alpha|\sqrt{1-x^2}\right)^2 \\ &\quad = \frac{P}{(1-|\alpha|x_0)^2} - \left(\sqrt{P} + |\alpha|\frac{\sqrt{P}x_0}{1-|\alpha|x_0}\right)^2 = 0. \end{aligned} \tag{24}$$

The other direction can be shown similarly. □

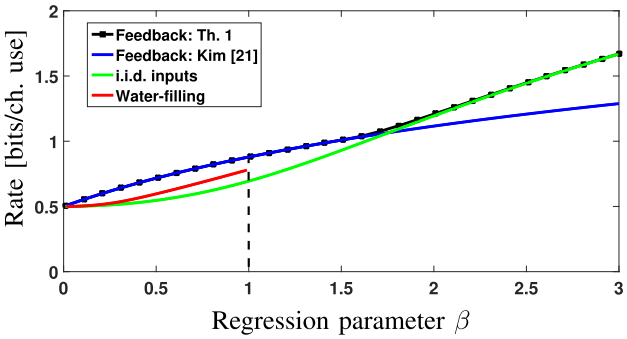

Fig. 1. The feedback capacity of the Gaussian channel with an auto-regressive noise of first order and a unit-power input constraint (black curve). The blue curve describes the feedback capacity expression for the stationary case ($|\beta| < 1$) from [21], but is plotted here for greater values of β, as it numerically coincides with the feedback capacity as long as $M = 0$ (see Fig. 2 below). The red curve corresponds to the feedforward capacity (without feedback) obtained via a water-filling solution, and the green curves corresponds to an i.i.d. coding law (feedback-independent) of the channel inputs in (26).

B. Auto-Regressive (AR) Noise

The auto-regressive (AR) process of first order is given by

$$z_i = \beta z_{i-1} + w_i, \tag{25}$$

where $w_i \sim N(0,1)$ is an i.i.d. sequence. This is one of the simplest instances of colored Gaussian noise and was studied in [11], [12], and [13]. A closed-form feedback capacity expression for the stationary case $|\beta| < 1$ was derived in [21][2]. We present next the AR noise with a general β. The AR process can be realized by (5) with $F = H = \beta$ and $G = L = W = V = 1$.

In Fig. 1, the feedback capacity in Theorem 1, the feedback capacity expression for $|\beta| < 1$ from [21], and the non-feedback capacity (using a water filling solution) are plotted. Additionally, we plot the maximal achievable rate with i.i.d. inputs by adding the constraint $\Gamma = 0$ to the optimization in (17). This rate can also be computed explicitly as

$$R_{iid}(\beta, P=1) = \frac{1}{2}\log\left(1 + \frac{\beta^2}{2}\left(1 + \sqrt{1 + \frac{4}{\beta^4}}\right)\right). \tag{26}$$

First, it is interesting to note that black and blue curves coincide for some non-stationary noise with $1 \leq \beta \lesssim 1.5$. This means that the capacity expression in [21] holds true even for some values beyond the stationary regime. The rate

[2]Recently [24] showed that [21] has an error in Corollary 4.4, which renders the feedback capacity expression in [21] for the AR noise with $|\beta| < 1$ unjustified. Particularizing our general capacity expression to AR noise and carrying out a computation similar that of the MA noise (Theorem 3) shows that the feedback capacity expression in [21] is correct for $|\beta| < 1$.

achieved with i.i.d. inputs (green curve) approaches the feedback capacity for large regression parameters. Thus, the plot shows that, as the regression parameter grows large, the rate achieved by i.i.d. inputs approaches the feedback capacity, and the feedback link has a negligible contribution in terms of capacity. From operational perspective, the feedforward capacity lies in between the i.i.d. achievable rate (green curve) and the feedback capacity (black curve). Thus, for large β, it can be well-approximated with simple i.i.d. inputs, i.e., codewords with memory are not needed. These phenomena are related to the structure of the optimal inputs distribution that is presented and discussed for the AR noise in Section IV-A.

IV. Optimal Inputs Distribution and Coding Scheme

In this section, we present a capacity-achieving, time-invariant inputs distribution that can be computed from the convex optimization in Theorem 1. We then use this inputs distribution to construct a capacity-achieving coding scheme for scalar channels, and discuss a possible extension to MIMO channels.

A. Optimal Inputs Distribution

The optimal decision variables in Theorem 1 induce a time-invariant capacity-achieving inputs distribution[3]:

$$\mathbf{x}_i = \Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i) + \mathbf{m}_i, \tag{27}$$

where $\hat{\mathbf{s}}_i$ is defined in (6), and its estimate at the decoder is defined by

$$\hat{\hat{\mathbf{s}}}_i \triangleq \mathbb{E}[\hat{\mathbf{s}}_i \,|\, \mathbf{y}^{i-1}]. \tag{28}$$

The optimal policy is composed as the sum of two signaling components. The first component, $(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$ corresponds to the decoder's estimation error, and is a function of the feedback. Its transmission refines the decoders' knowledge on $\hat{\mathbf{s}}_i$ by transmitting the states innovation (the vector $\hat{\mathbf{s}}_i$ can be regarded as the channel state, see Section II). The covariance of the innovation $(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$ is $\hat{\Sigma}$, so that the covariance of the first component is $\mathbf{cov}(\Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)) = \Gamma\hat{\Sigma}^{\dagger}\Gamma^T$. The second component, $\mathbf{m}_i$, is independent of $(\mathbf{x}^{i-1}, \mathbf{y}^{i-1})$ (and thus is feedback-independent), and has an i.i.d. distribution with the remaining covariance, i.e., $\mathbf{m}_i \sim N(0, M)$ with $M \triangleq \Pi - \Gamma\hat{\Sigma}^{\dagger}\Gamma^T$. The transmission of the vector $\mathbf{m}_i$ increases the uncertainty of the channel state $\hat{\mathbf{s}}_i$ at the decoder, but it can be used to transmit new information (on the message). In the AWGN channel, for instance, the entire power is allocated to the second component $\mathbf{m}_i$. We proceed to illustrate the policy behavior for the AR noise.

In Fig. 2, the power allocated to each of the signals in (27) is plotted for the AR noise in (25) with a power constraint $P = 1$. For $0 < \beta \leq 1$, Fig. 2 agrees with the claim in [21, Th. 4.6] on the sufficiency of inputs distribution with $\mathbf{m}_i = 0$ for scalar and stationary noise (see also next paragraph). However, beyond the stationary regime, there is a sharp phase transition, and the power allocated to $\mathbf{m}_i$ increases as β grows. The phase transition location, $\beta \approx 1.5$, explains the gap between the feedback capacity in Theorem 1 and the capacity expression in [21] since the latter used a policy with $\mathbf{m}_i = 0$. Fig. 2 also shows that the rate achieved with i.i.d. inputs in (26) approaches the feedback capacity for growing β. This implies that the Schalkwijk-Kailath (SK) encoding law is close to optimal in this regime [35].

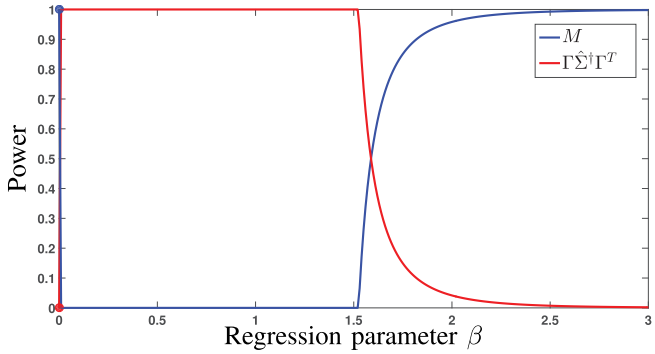

Fig. 2. The power of the signaling components in the inputs distribution (27) for AR noise. The red curve corresponds to the feedback-dependent signal $\Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$ and the blue curve corresponds to $\mathbf{m}_i$. Note that there are two phase transitions, at $\beta = 0$ and $\beta \approx 1.5$. Consistent with [21, Th. 4.6], $M = 0$ achieves the feedback capacity in the stationary regime $\beta \in (0, 1)$. Also, as β grows large, the power allocated to the i.i.d. component grows as well. Combined with the i.i.d. coding law curve in Fig. 1 (green curve), this shows that the feedback link has negligible contribution to the capacity solution.

The role of the second component $\mathbf{m}_i$ has been discussed in several papers [10], [21], [24]. In [21, Cor. 4.4], it is claimed that for scalar channels with stationary noise, the capacity can be achieved with $M = 0$. Recently, [24] showed that the proof of the claim in [21, Cor. 4.4] relies on an erroneous calculation and thus is invalid. Our capacity derivation relies on a general policy with $M \succeq 0$ (with a different coding for the first time $t = 1$). As illustrated in the examples above, for general noise processes, M can be either positive or zero; for the MA noise, we prove in Theorem 3 that $M = 0$ is necessary to achieve the capacity, and for the AR noise, it is illustrated that $M > 0$ in the non-stationary regime (Fig. 2). When specializing our capacity expression for stationary noise processes, it may be utilized to find a counterexample for [21, Cor. 4.4]. We ran extensive simulations to specialize our capacity expression in Theorem 2 to various stationary noise processes and to compute the optimal M, yet we did not find a counterexample to [21, Cor. 4.4]. Thus the claim in [21, Cor. 4.4] may be true.

As mentioned in Remark 1, the fact that $M = 0$ does not imply that the achievable rate is as erroneously claimed in [10]. If $M = 0$, it simply implies that the message is encoded at the first time with $\mathbf{m}_1 \neq 0$, and from $t > 1$ the encoder follows the rule $\mathbf{x}_i = \Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$. In the next section, we show that this explicit coding scheme is capacity-achieving with double-exponential decay in the error probability for any rate below capacity.

B. Coding Scheme for Scalar Channels

In this section, we construct a capacity-achieving coding scheme for scalar channels (with a vector hidden state) based on the optimal inputs distribution in (27). Throughout this section, it is assumed that the optimal inputs distribution in

[3]More precisely, a capacity-achieving policy in Lemma 6 is the time-invariant law in (27) for $t > 1$, and a different coding law for $t = 1$.

(27) satisfies $M = 0$. The design of an explicit coding scheme with $M \neq 0$ remains open.

Our scheme resembles the SK scheme [35], [36], and other posterior matching schemes for memoryless channels [37], [38], [39], [40] and channels with memory [21], [23], [25], [41], [42], [43], [44] in its main idea to refine the decoders' knowledge of the message (or equivalently, to refine the decoders' knowledge of the first channel noise instance in Gaussian channels).

The main difference with the SK scheme is that rather than encoding the scaled innovation of the first noise instance z_0 (our coding scheme starts at $i = 0$), our encoding follows (27) to transmit the innovation of $\hat{\mathbf{s}}_i$. This modification results in a more numerically stable encoding since our scaling factor $\Gamma\hat{\Sigma}^{\dagger}$ is a constant, while in the SK scheme the scaling of the message innovation increases with time.

A related scheme for a similar setting appears in [21]. Both schemes follow the encoding in (27) but, only our paper provides a computable expression for the coefficient matrix $\Gamma\hat{\Sigma}^{\dagger}$ (via Theorem 1). Additionally, we simplify the multidimensional encoding method in [21] by showing that, even when the hidden state is a vector, it is sufficient to encode the message in a single time instance. Additionally, we provide an explicit smoother for the maximum likelihood decoder proposed in [21] and [45].

Our coding scheme consists of three main steps:

1) **Message transmission:** At time $i = 0$, the message $m \in [1 : 2^{nR}]$ is mapped to a zero-mean, unit-separation symbol $U(m) = m - 2^{nR-1}$ whose variance is $\text{Var}(U) = (2^{2nR} - 1)/12$. The normalized symbol $\bar{U}(m) = \text{Var}(U)^{-\frac{1}{2}} U(m)$ will be transmitted at the first time instance as $\mathbf{x}_0(m)$.
2) **Refinement:** At times $i = 1, \ldots, n$, the encoding process is simply to use the optimal inputs distribution in (27). The estimates $\hat{\mathbf{s}}_i$ and $\hat{\hat{\mathbf{s}}}_i$ can be computed directly from (7) and (42) and the constant $\Gamma\hat{\Sigma}^{\dagger}$ is obtained from the optimization in Theorem 1.
3) **Decoding:** When the refinement stage ends (after the nth transmission), the decoder constructs the maximum-likelihood estimate of the noise instance $\mathbf{z}_0$ based on the measurements $\mathbf{y}_1^n$. This estimation problem corresponds to a smoothing problem analysis that is formally presented in Lemma 1. Based on the estimate of $\mathbf{z}_0$, the decoder forms an estimate of $\mathbf{x}_0$ and declares its nearest neighbour $x_0(m)$ as the decoded message point.

We are ready to present the coding scheme as Algorithm 1. The abbreviation KF-ENC stands for time-invariant Kalman filter at the encoder (14). The abbreviation KF-DEC stands for the time-varying Kalman filter at the decoder in (42) with the initial conditions $\hat{\Sigma}_1 = K_p \Psi K_p^T$ and $M_0 = V(\bar{U})$ (and $\Psi_i = \Psi, K_{p,i} = K_p$ for $i = 1, \ldots, n$). Lastly, Smooth stands for the smoothing function in Lemma 1 (Eq. (30) below).

Algorithm 1 Optimal Scheme for Scalar Channels

Inputs: m, Γ, $\hat{\Sigma}$, $\hat{z}_{0|0} = 0$
$x_0 \leftarrow \bar{U}(m)$ ▷ *Transmission*
Store $y_0 = x_0 + z_0$ ▷ *Dec.*
$\hat{\mathbf{s}}_1 \leftarrow K_p z_0$ ▷ *Enc. estimate*
$\hat{\hat{\mathbf{s}}}_1 \leftarrow 0$ ▷ *Initialization*
procedure $i = 1 : n$
  $x_i \leftarrow \Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$ ▷ *Transmission*
  $y_i = x_i + z_i$ ▷ *Ch. output*
  $\hat{\mathbf{s}}_{i+1} \leftarrow \text{KF-ENC}(\hat{\mathbf{s}}_i, z_i = y_i - x_i)$ ▷ *Eq. (14)*
  $\hat{\hat{\mathbf{s}}}_{i+1} \leftarrow \text{KF-DEC}(\hat{\hat{\mathbf{s}}}_i, y_i, i)$ ▷ *Eq. (42)*
  $\hat{z}_{0|i} \leftarrow \text{Smooth}(\hat{z}_{0|i-1}, y_i, i)$ ▷ *Dec. (Eq. (30))*
end procedure
$\hat{m} = \arg\min_{m'} |(y_0 - \hat{z}_{0|n}) - \bar{U}(m')|$ ▷ *Dec.*

An implementation of the coding scheme can be found in [46]. Note that the decoder in Algorithm 1 does not utilize y_0 to estimate $\hat{\hat{\mathbf{s}}}_1$. The choice of the initial state estimate, $\hat{\hat{\mathbf{s}}}_1 = 0$, is arbitrary, and allows us to present the analysis of the refinement stage at $i = 1, \ldots, n$ independently of the message transmission at $i = 0$. The following theorem shows the optimality of our coding scheme.

Theorem 4 (Capacity-achieving coding scheme): For any rate $R < C_{fb}(P)$, the error probability of the coding scheme for scalar channels (with $\mathbf{m}_i = 0$) in Algorithm 1 decays in a doubly-exponential rate for large n.

A simple proof of Theorem 4 appears at the end of this section. Its main building block is the analysis of a smoothing problem (Lemma 1) that corresponds to the refinement stage. We provide now explicit formulas to compute the estimate and its error covariance. The formulas are presented for the general MIMO channel, and their particularization to scalar channels will be used in the proof of Theorem 4.

Lemma 1 (The Smoothing Problem): Consider the smoothing problem of estimating $\mathbf{z}_0$ from $\mathbf{y}_1^n$ with

$$\begin{aligned} \hat{\mathbf{z}}_{0|n} &\triangleq \mathbb{E}[\mathbf{z}_0 \,|\, \mathbf{y}_1^n] \\ \hat{Z}_{0|n} &\triangleq \mathbf{cov}(\mathbf{z}_0 - \hat{\mathbf{z}}_{0|n}). \end{aligned} \tag{29}$$

Subject to the optimal inputs distribution (27), when $\mathbf{m}_i = 0$,

1) The optimal smoother can be recursively computed as

$$\hat{\mathbf{z}}_{0|i} = \hat{\mathbf{z}}_{0|i-1} + \hat{Z}_{0|i-1}\kappa_i^T \Psi_{Y,i}^{-1}(\mathbf{y}_i - H\,\hat{\hat{\mathbf{s}}}_i), \quad i = 1, \ldots, n \tag{30}$$

with $\hat{\mathbf{z}}_{0|0} = 0$ and

$$\kappa_i \triangleq (\Lambda\Gamma\hat{\Sigma}^{\dagger} + H)(F - K_p(\Lambda\Gamma\hat{\Sigma}^{\dagger} + H))^{i-1} K_p.$$

2) The error covariance can be updated as

$$\hat{Z}_{0|i} = (I - \hat{Z}_{0|i-1}\kappa_i^T \Psi_{Y,i}^{-1}\kappa_i)\hat{Z}_{0|i-1}, \quad i = 1, \ldots, n \tag{31}$$

with $\hat{Z}_{0|0} = \Psi$, and its determinant satisfies

$$\det(\hat{Z}_{0|i}) = \det(\Psi_{Y,i}^{-1}\Psi)\det(\hat{Z}_{0|i-1}). \tag{32}$$

Moreover, $\Psi_{Y,i}$ converges to Ψ_Y^*, the optimal value of Ψ_Y in Theorem 1.

- Therefore, for scalar channels, the error covariance satisfies

$$\hat{Z}_{0|i} = \Psi_{Y,i}^{-1} \Psi \hat{Z}_{0|i-1}, \tag{33}$$

and $\Psi_{Y,i}$ converges to Ψ_Y^*.

The proof of Lemma 1 appears in Section VI-A. Recall from Theorem 1 that the capacity can be expressed as $C_{fb}(P) = \frac{1}{2}\log\det(\Psi_Y^* \Psi^{-1})$ where Ψ_Y^* denotes the optimal Ψ_Y. The relation between the capacity and the smoothing problem is transparent. The volume (determinant) reduction of the error covariance in (32) is the logarithm argument in the capacity expression. For scalar channels, the volume reduces to a single dimension refinement of the noise instance $\mathbf{z}_0$. However, for MIMO channels the volume reduction is not sufficient to derive an explicit coding scheme. We provide details on a possible MIMO scheme construction.

A suggested scheme for MIMO channels is as follows: assume for simplicity $\Lambda = I$, and split the message into p independent sub-messages where p is dimension of the input, output and noise. In the first time, we normalize each sub-message, and transmit their concatenation as the vector $\mathbf{x}_0$. The encoding is identical to that in Agorithm 1, that is, it follows the policy in (27). The estimation of the vector $\mathbf{z}_0$ is based on the smoother in (30), and the decoding is carried out using coordinate-wise successive cancellation of the vector $\hat{\mathbf{z}}_{0|n}$. The analysis of such scheme can be possibly done using Lemma 1, but a finer spectral analysis is needed. In particular, the geometric reduction in (32) needs to be shown for each coordinate and not for the overall determinant. The geometric rate of the error covariance in (33) should also determine the rates allocated to each sub-message, and is the key to obtain the double-exponential decay. We proceed to show the optimality of Algorithm 1 for scalar channels using the analysis of Lemma 1.

Proof of Theorem 4: The decoder estimates x_0 from $y_0 - \hat{z}_{0|n} = \bar{U}(m) + z_0 - \hat{z}_{0|n}$. This is the problem of estimating an M-PAM signal from a Gaussian-corrupted measurement, and the probability of error can be bounded as

$$\begin{aligned} P_e^{(n)} &\le \Pr\left(|z_0 - \mathbb{E}[z_0|y^n]| \ge \frac{1}{2}\sqrt{\frac{12}{2^{2nR}-1}} \right) \\ &= 2Q(\gamma_n), \end{aligned} \tag{34}$$

where the inequality follows from the first and last messages where a large error deviation will not incur an error on one of their ends, and the equality follows from $\gamma_n \triangleq \frac{1}{2}\sqrt{12\frac{\mathrm{Var}(z_0|y^n)}{2^{2nR}-1}}$ with the standard Q-function. We can further bound γ_n as

$$\gamma_n \le \sqrt{3} \cdot 2^{-nR} \sqrt{\mathrm{Var}(z_0|y^n)}. \tag{35}$$

By Lemma 1, γ_n has a positive exponent if $R < \frac{1}{2n}\log\Psi \prod_{i=1}^n (\Psi_{Y,i}\Psi^{-1})$, which leads to the doubly-exponential decay rate. As $\Psi_{Y,i} \to \Psi_Y^*$, R can be chosen arbitrarily close to $\frac{1}{2}\log(\Psi_Y^*\Psi^{-1})$ which is precisely the feedback capacity. □

V. Proof of the Main Result

In this section we outline the proof of Theorem 1 by presenting the technical lemmas leading to tight lower and upper bounds. The proof is structured as three parts.

1. **Sequential convex optimization problem (SCOP)**: The n-letter capacity expression for the MIMO channel is defined as

$$C_n(P) = \max_{P(\mathbf{x}^n \,||\, \mathbf{y}^n):\, \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbf{x}_i^T \mathbf{x}_i] \le P} h(\mathbf{y}^n) - h(\mathbf{z}^n). \tag{36}$$

The first three lemmas formulate the n-letter capacity as a SCOP. While it is easy to show that $C_n(P)$ is concave in its decision variable $P(x^n\|y^n)$, the challenge is to formulate it as a convex optimization problem that enables one to explicitly compute the limit of $C_n(P)$. To this end, we realize a SCOP with LMI constraints that have a sequential structure.

2. **Upper bound via convexity**: The second part of the proof utilizes the SCOP structure to show that the capacity expression in Theorem 1 is an upper bound on the capacity. Since the optimization constraints contain decision variables at consecutive times, the standard time-sharing random variable argument does not apply here, and we use a different technique to show that these constraints are *asymptotically satisfied* when realized at the convex combinations of the decision variables.

3. **Lower bound using time-invariant inputs**: The last part constructs a time-invariant policy whose optimization leads to a lower bound that is expressed as the upper bound optimization problem with additional constraints. We show that the additional constraints are redundant, concluding the proof of the main result.

A. Sequential Convex Optimization Problem

The first lemma identifies an optimal structure for the inputs distribution using $\hat{\mathbf{s}}_i$ and $\hat{\hat{\mathbf{s}}}_i$ defined in (6) and (28), respectively.

Lemma 2 (The Optimal Policy Structure): For a fixed n, it is sufficient to optimize (36) with inputs of the form

$$\mathbf{x}_i = \Gamma_i \hat{\Sigma}_i^\dagger (\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i) + \mathbf{m}_i, \qquad i = 1, \ldots, n \tag{37}$$

where $\mathbf{m}_i \sim N(0, M_i)$ is independent of $(\mathbf{x}^{i-1}, \mathbf{y}^{i-1})$, $\hat{\Sigma}_i^\dagger$ is the Moore-Penrose pseudo-inverse of

$$\hat{\Sigma}_i = \mathbf{cov}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i), \tag{38}$$

Γ_i is a matrix that satisfies

$$\Gamma_i (I - \hat{\Sigma}_i \hat{\Sigma}_i^\dagger) = 0, \tag{39}$$

and the power constraint is

$$\frac{1}{n}\sum_{i=1}^n \mathbf{Tr}(\Gamma_i \hat{\Sigma}_i^\dagger \Gamma_i^T + M_i) \le P. \tag{40}$$

Lemma 2 simplifies the optimization (36) by showing that the optimization domain is over the sequence of matrices $(\Gamma_i, M_i \succeq 0)_{i=1}^n$. Note that $\hat{\Sigma}_i$ is a deterministic function of the policy up to time $i-1$ and thus is not part of the optimization. Similar policies appeared in the literature e.g. [23, Section IV] and [10] building on the ideas in [22]. Their policy reads $\mathbf{x}_i = \Gamma_i(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i) + \mathbf{m}_i$, and our policy in Lemma 2

is a subset of their policy. Specifically, if $\hat{\Sigma}_i$ is invertible, the orthogonality constraint is redundant, and one can use the change of variable $\Gamma_i' = \Gamma_i \hat{\Sigma}_i^{-1}$ to show the equivalence of the policies. However, in general, $\hat{\Sigma}_i$ may be singular, and the orthogonality constraint is required for the convex optimization formulation in Lemma 4. In the next lemma, the channel output is formalized as the measurement of a controlled state space.

Lemma 3 (Channel Outputs Dynamics): For a fixed policy $\{(\Gamma_i, M_i)\}_{i=1}^n$, the channel outputs admit the state-space model

$$\begin{aligned} \hat{\mathbf{s}}_{i+1} &= F\,\hat{\mathbf{s}}_i + K_{p,i}\,\mathbf{e}_i, \\ \mathbf{y}_i &= (\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger} + H)\,\hat{\mathbf{s}}_i - \Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}\,\hat{\mathbf{s}}_i + \Lambda\,\mathbf{m}_i + \mathbf{e}_i, \end{aligned} \tag{41}$$

where $K_{p,i}$ and $\mathbf{e}_i \sim N(0, \Psi_i)$ are defined in (8). The estimator in (28) can be written as

$$\hat{\hat{\mathbf{s}}}_{i+1} = F\,\hat{\hat{\mathbf{s}}}_i + K_{Y,i}(\mathbf{y}_i - H\,\hat{\hat{\mathbf{s}}}_i), \tag{42}$$

and its error covariance $\hat{\Sigma}_i = \mathbf{cov}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$ satisfies the Riccati recursion

$$\hat{\Sigma}_{i+1} = F\hat{\Sigma}_i F^T + K_{p,i}\Psi_i K_{p,i}^T - K_{Y,i}\Psi_{Y,i}K_{Y,i}^T \tag{43}$$

with the initial condition $\hat{\Sigma}_1 = 0$, and the constants

$$\begin{aligned} \Psi_{Y,i} &= (\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger} + H)\hat{\Sigma}_i(\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger} + H)^T + \Lambda M_i \Lambda^T + \Psi_i \\ K_{Y,i} &= (F\hat{\Sigma}_i(\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger} + H)^T + K_{p,i}\Psi_i)\Psi_{Y,i}^{-1}. \end{aligned} \tag{44}$$

Lemma 3 is a direct consequence of the policy derived in Lemma 2. As seen from (41), the policy in (37) translates into an additive measurement noise $\mathbf{m}_i$ and a modification of the observability matrix $\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger} + H$. Similar state-space structures appeared in [21] and [32], but it is interesting to realize that (41) does not fall into the classical state-space structure since the observability matrix depends on the error covariance $\hat{\Sigma}_i$ induced from our policy. Lemma 3 also reveals an objective structure that resembles the one in Theorem 1. In particular, we can use the covariance of the channel outputs innovation in (44), $\Psi_{Y,i}$, and (11) to write

$$h(\mathbf{y}_i \,|\, \mathbf{y}^{i-1}) - h(\mathbf{z}_i \,|\, \mathbf{z}^{i-1}) = \frac{1}{2}\log\det(\Psi_{Y,i}) - \frac{1}{2}\log\det(\Psi_i).$$

The next lemma summarizes the SCOP formulation.

Lemma 4 (Sequential Convex-Optimization Formulation): The n-letter capacity can be bounded by the convex optimization problem

$$\begin{aligned} C_n(P) \le \max_{\{\Gamma_i, \Pi_i, \hat{\Sigma}_{i+1}\}_{i=1}^n} \; & \frac{1}{2n}\sum_{i=1}^n \log\det(\Psi_{Y,i}) - \log\det(\Psi_i) \\ s.t. \quad & \begin{pmatrix} \Pi_t & \Gamma_t \\ \Gamma_t^T & \hat{\Sigma}_t \end{pmatrix} \succeq 0, \quad \frac{1}{n}\sum_{i=1}^n \mathbf{Tr}(\Pi_i) \le P, \\ \Psi_{Y,t} &= \Lambda\Pi_t\Lambda^T + H\hat{\Sigma}_t H^T + \Lambda\Gamma_t H^T + H\Gamma_t^T\Lambda^T + \Psi_t \\ K_{Y,t} &= (F\Gamma_t^T\Lambda^T + F\hat{\Sigma}_t H^T + K_{p,t}\Psi_t)\Psi_{Y,t}^{-1} \\ & \begin{pmatrix} F\hat{\Sigma}_t F^T + K_{p,t}\Psi_t K_{p,t}^T - \hat{\Sigma}_{t+1} & K_{Y,t}\Psi_{Y,t} \\ \Psi_{Y,t}K_{Y,t}^T & \Psi_{Y,t} \end{pmatrix} \succeq 0, \hat{\Sigma}_{n+1} \succeq 0 \end{aligned} \tag{45}$$

where the constraints hold for $t = 1, \ldots, n$, and $\hat{\Sigma}_1 = 0$.

To see that (45) is a convex optimization, note that each of the matrix constraints is a linear function of the decision variables. In the next section, we provide the single-letter upper bound on the capacity. The key to the upper bound is the concavity of the objective function and the linearity of the constraints, along with the crucial property that the Riccati LMI constraint in (45) includes decision variables of two consecutive times only.

B. Single-Letter Upper Bound

The next lemma concludes the upper bound in Theorem 1.

Lemma 5 (The Upper Bound): The feedback capacity is bounded by the convex optimization

$$\begin{aligned} C_{fb}(P) \le \max_{\Pi, \hat{\Sigma}, \Gamma} \; & \frac{1}{2}\log\det(\Psi_Y) - \frac{1}{2}\log\det(\Psi) \\ \text{s.t.} \quad & \begin{pmatrix} \Pi & \Gamma \\ \Gamma^T & \hat{\Sigma} \end{pmatrix} \succeq 0, \quad \mathbf{Tr}(\Pi) \le P, \\ \Psi_Y &= \Lambda\Pi\Lambda^T + H\hat{\Sigma}H^T + \Lambda\Gamma H^T + H\Gamma^T\Lambda^T + \Psi \\ K_Y &= (F\Gamma^T\Lambda^T + F\hat{\Sigma}H^T + K_p\Psi)\Psi_Y^{-1} \\ & \begin{pmatrix} F\hat{\Sigma}F^T + K_p\Psi K_p^T - \hat{\Sigma} & K_Y\Psi_Y \\ \Psi_Y K_Y^T & \Psi_Y \end{pmatrix} \succeq 0. \end{aligned} \tag{46}$$

The main idea behind the upper bound is to show that the objective function evaluated at the convex combination of each of the decision variables in Lemma 4 achieves a larger objective. At a high level, the idea is similar to the time-sharing random variable, but the challenge lies in the constraints. Specifically, one cannot show that the Riccati LMI constraint (46) is satisfied at all times when evaluated at the convex combination of the decision variables. To settle this point, we show that this constraint is satisfied in the asymptotics.

C. Lower Bound

In this section, we show that the upper bound in Lemma 5 is achievable. It is shown with two lemmas: the first formulates a lower bound as an optimization problem that resembles the upper bound but has two additional constraints. The second lemma shows that additional constraints are satisfied in the upper bound optimization problem.

Lemma 6 (Lower Bound): For time-invariant policies

$$\mathbf{x}_i = \Gamma(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i) + \mathbf{m}_i, \qquad i \ge 2 \tag{47}$$

with $\mathbf{m}_i \sim N(0, M)$ (and a different coding rule for $i = 1$), the maximization of (36) over (Γ, M) achieves the lower bound

$$\begin{aligned} C_{fb}(P) \ge \max_{\Gamma, \Pi, \hat{\Sigma}} \; & \log\det(\Psi_Y) - \log\det(\Psi) \\ \text{s.t.} \quad & \begin{pmatrix} \Pi & \Gamma \\ \Gamma^T & \hat{\Sigma} \end{pmatrix} \succeq 0, \quad \mathbf{Tr}(\Pi) \le P \\ K_Y &= (F\hat{\Sigma}H^T + F\Gamma^T\Lambda^T + K_p\Psi)\Psi_Y^{-1} \\ \Psi_Y &= \Lambda\Pi\Lambda^T + H\hat{\Sigma}H^T + \Lambda\Gamma H^T + H\Gamma^T\Lambda^T + \Psi \\ \hat{\Sigma} &= F\hat{\Sigma}F^T + K_p\Psi K_p^T - K_Y\Psi_Y K_Y^T \end{aligned} \tag{48}$$

$$\exists K : \rho(F - K(\Lambda\Gamma\hat{\Sigma}^{\dagger} + H)) < 1. \tag{49}$$

The optimization problem in (49) is the same as the upper bound in (46) except for the additional constraint (49) and the Riccati equation (48) which appears as an inequality in the upper bound (16). Next, we show that these two conditions are redundant concluding the proof of Theorem 1.

Lemma 7 (Equality Between the Lower and Upper Bounds): For any optimal tuple $(\Pi, \hat{\Sigma}, \Gamma)$ for the upper bound optimization problem in (46):

1) There exists an optimal tuple such that the Schur complement of the Riccati LMI (16) is achieved with equality.
2) The pair $(F, \Lambda\Gamma\hat{\Sigma}^{\dagger} + H)$ is detectable, i.e.,

$$\exists K : \rho(F - K(\Lambda\Gamma\hat{\Sigma}^{\dagger} + H)) < 1.$$

Consequently, the upper bound in Lemma 5 and the lower bound in Lemma 6 are equal to the feedback capacity.

For scalar channels with $H \neq 0$, it can be shown that the optimal tuple satisfies the first item. That is, the Schur complement of the Riccati equation evaluated at any optimal solution tuple is zero. This fact is utilized in Theorem 3.

VI. Proof of Technical Lemmas

In this section, we provide detailed proofs of Lemmas 2 - 7 consecutively. We then prove Lemma (1) on the smoothing problem in Section IV.

Proof of Lemma 2: The policy in (37) forms a subset of the maximization domain $P(\mathbf{x}^n \,||\, \mathbf{y}^n) = \prod_{i=1}^{n} P(\mathbf{x}_i \,|\, \mathbf{x}^{i-1}, \mathbf{y}^{i-1})$ in (36). Thus, our proof strategy is to construct a policy of the form (37), for any inputs distribution $P(\mathbf{x}^n \,||\, \mathbf{y}^n)$, and show that it induces the same objective. The optimality of a Gaussian inputs distribution in (36) can be shown with a standard argument of maximum entropy, e.g., [21]. We start by computing the ith objective as

$$\begin{aligned} &h(\mathbf{y}_i \,|\, \mathbf{y}^{i-1}) - h(\mathbf{z}_i \,|\, \mathbf{z}^{i-1}) \\ &\quad = \frac{1}{2}\log\det(\mathbf{cov}(\mathbf{y}_i - \hat{\hat{\mathbf{y}}}_i)) - \frac{1}{2}\log\det(\Psi_i), \end{aligned} \tag{50}$$

where $\hat{\hat{\mathbf{y}}}_i \triangleq \mathbb{E}[\mathbf{y}_i \,|\, \mathbf{y}^{i-1}]$. The covariance can be computed explicitly as

$$\begin{aligned} \mathbf{cov}(\mathbf{y}_i - \hat{\hat{\mathbf{y}}}_i) &\overset{(a)}{=} \mathbf{cov}(\mathbf{y}_i - \hat{\hat{\mathbf{z}}}_i) \\ &\overset{(b)}{=} \mathbf{cov}(\Lambda\,\mathbf{x}_i + H\,\hat{\mathbf{s}}_i - H\,\hat{\hat{\mathbf{s}}}_i + \mathbf{z}_i - H\,\hat{\mathbf{s}}_i) \\ &\overset{(c)}{=} \mathbf{cov}(\Lambda\,\mathbf{x}_i + H(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)) + \Psi_i \end{aligned} \tag{51}$$

where (a) follows from $\hat{\hat{\mathbf{z}}}_i \triangleq \mathbb{E}[\mathbf{z}_i \,|\, \mathbf{y}^{i-1}]$ and $\mathbb{E}[\mathbf{x}_i \,|\, \mathbf{y}^{i-1}] = 0$. The latter assumption is without loss of optimality since any policy with $\mathbb{E}[\mathbf{x}_i \,|\, \mathbf{y}^{i-1}] \neq 0$ can be modified to $\bar{\mathbf{x}}_i = \mathbf{x}_i - \mathbb{E}[\mathbf{x}_i \,|\, \mathbf{y}^{i-1}]$ that has zero mean without affecting the objective function in (50). Step (b) follows from the channel outputs definition in (4) and (28), and (c) follows from the independence of the innovation $\mathbf{z}_i - H\,\hat{\mathbf{s}}_i$ and the tuple $(\mathbf{x}^i, \mathbf{y}^{i-1}, \mathbf{z}^{i-1})$.

For any inputs distribution $P(\mathbf{x}^n \,||\, \mathbf{y}^n)$, denoted by P, we construct a new policy of the form (37), denoted by Q, as follows

$$\mathbf{x}_i = \Gamma_i \hat{\Sigma}_i^{\dagger}(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i) + \mathbf{m}_i, \tag{52}$$

where $\Gamma_i \triangleq \mathbb{E}_P[\mathbf{x}_i(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T]$, $\mathbf{m}_i$ is independent of $(\mathbf{x}^{i-1}, \mathbf{y}^{i-1})$ and is distributed according to $\mathbf{m}_i \sim N(0, M_i)$ with

$$M_i \triangleq \mathbb{E}_P[\mathbf{x}_i\,\mathbf{x}_i^T] - \mathbb{E}_P[\mathbf{x}_i(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T]\hat{\Sigma}_i^{\dagger}\,\mathbb{E}_P[(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)\,\mathbf{x}_i^T], \tag{53}$$

and $\hat{\Sigma}_i^{\dagger}$ is the pseudo inverse of $\hat{\Sigma}_i \triangleq \mathbf{cov}_P(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)$. The subscript P is made to emphasize the dependence on the distribution P.

We show by induction that the new policy in (52) induces the same objective as the distribution P. Consider the Gaussian vector $\Xi_i^{P/Q} \triangleq (\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i, \mathbf{x}_i, \mathbf{y}_i - \hat{\hat{\mathbf{y}}}_i)$ where the superscript indicates its distribution. If we show that Ξ_i^P has the same distribution as Ξ_i^Q for $i = 1, \ldots, n$, then their objectives are equal by (50). For the base case of the induction, we have $\Xi_1^{P/Q} = (0, 0, \mathbf{x}_1, \mathbf{y}_1)$ for both policies and our construction in (52) guarantees that $\mathbf{x}_1$ has the same distribution for both policies. For the induction step, assume that the variables $\{\Xi_i^P\}_{i \le t}$ have the same distribution as $\{\Xi_i^Q\}_{i \le t}$. We show that tuple Ξ_{t+1}^P has the same distribution as Ξ_{t+1}^Q by comparing their different components using a Bayes rule. First, the encoders' estimate $\hat{\mathbf{s}}_{t+1}$ is independent of the policy choice. The decoders' estimate $\hat{\hat{\mathbf{s}}}_{t+1} = \mathbb{E}[\hat{\mathbf{s}}_{t+1} \,|\, \mathbf{y}^t]$ is a function of the innovations $\{\mathbf{y}_i - \hat{\hat{\mathbf{y}}}_i\}_{i \le t}$, and by the induction hypothesis these innovations have the same distribution. These first two steps conclude that $\mathbf{cov}_P(\hat{\mathbf{s}}_{t+1} - \hat{\hat{\mathbf{s}}}_{t+1}) = \mathbf{cov}_Q(\hat{\mathbf{s}}_{t+1} - \hat{\hat{\mathbf{s}}}_{t+1})$. For the channel input, it can be easily verified by (52) that $\mathbb{E}_Q[\mathbf{x}_{t+1}\,\mathbf{x}_{t+1}^T] = \mathbb{E}_P[\mathbf{x}_{t+1}\,\mathbf{x}_{t+1}^T]$, and the orthogonality constraint (shown below) implies that

$$\begin{aligned} \mathbb{E}_Q[\mathbf{x}_{t+1}(\hat{\mathbf{s}}_{t+1} - \hat{\hat{\mathbf{s}}}_{t+1})^T] &= \Gamma_{t+1}\hat{\Sigma}_{t+1}^{\dagger}\hat{\Sigma}_{t+1} \\ &= \Gamma_{t+1} \\ &= \mathbb{E}_P[\mathbf{x}_{t+1}(\hat{\mathbf{s}}_{t+1} - \hat{\hat{\mathbf{s}}}_{t+1})^T]. \end{aligned}$$

The last step to complete the inductive step is for the innovation $\mathbf{y}_{t+1} - \hat{\hat{\mathbf{y}}}_{t+1}$, and we note from (51) that the distribution of the latter conditioned on $(\hat{\mathbf{s}}_{t+1} - \hat{\hat{\mathbf{s}}}_{t+1})$, $\mathbf{x}_{t+1}$ is determined by $\mathbf{z}_{t+1} - H\,\hat{\mathbf{s}}_{t+1}$.

The orthogonality constraint $\Gamma_i(I - \hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i)$ is a property of covariance matrices since $(I - \hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i)$ is the orthogonal projection onto the kernel of $\hat{\Sigma}_i$, but we prove it here for completeness. Consider the eigendecomposition of the covariance matrix

$$\begin{aligned} \hat{\Sigma}_i &= \mathbb{E}[(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T] \\ &= \begin{pmatrix} U_0 & U_1 \end{pmatrix} \begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_0^T \\ U_1^T \end{pmatrix}, \end{aligned} \tag{54}$$

where $\begin{pmatrix} U_0 & U_1 \end{pmatrix}$ is an orthogonal matrix and $\Omega \succ 0$ which imply $(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T U_1 = 0$. The Moore-Penrose pseudo inverse is

$$\hat{\Sigma}_i^{\dagger} = \begin{pmatrix} U_0 & U_1 \end{pmatrix} \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_0^T \\ U_1^T \end{pmatrix}, \tag{55}$$

and the constraint can be written as

$$\begin{aligned} &\mathbb{E}[\mathbf{x}_i(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T](I - \hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i) \\ &\quad = \mathbb{E}[\mathbf{x}_i(\hat{\mathbf{s}}_i - \hat{\hat{\mathbf{s}}}_i)^T]\left(I - \begin{pmatrix} U_0 & U_1 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_0^T \\ U_1^T \end{pmatrix}\right). \end{aligned} \tag{56}$$

To see that (56) is the zero matrix, note that if u is a column of U_0, then $\left(I-\begin{pmatrix}U_0 & U_1\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}U_0^T\\ U_1^T\end{pmatrix}\right)u = 0$. Further, if u is a column of U_1, then $\left(I-\begin{pmatrix}U_0 & U_1\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}U_0^T\\ U_1^T\end{pmatrix}\right)u = u$, but $(\hat{\mathbf{s}}_i-\hat{\hat{\mathbf{s}}}_i)^T u = 0$ by the decomposition in (54).

Finally, it can be verified that the power consumed by the new policy satisfies

$$\sum_{i=1}^{n}\mathbb{E}_Q[\mathbf{x}_i^T\,\mathbf{x}_i]=\sum_{i=1}^{n}\mathbb{E}_P[\mathbf{x}_i^T\,\mathbf{x}_i].$$

□

Proof of Lemma 3: The recursion for the predicted state $\hat{\mathbf{s}}_{i+1}$ is given in Eq. (7) where $\mathbf{e}_i$ is the innovation process. For the channel output, we use Lemma 2 to write

$$\begin{aligned}\mathbf{y}_i &= \Lambda\,\mathbf{x}_i+\mathbf{z}_i\\ &= (\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}+H)\,\hat{\mathbf{s}}_i-\Lambda\Gamma_i\,\hat{\hat{\mathbf{s}}}_i+\Lambda\mathbf{m}_i+\mathbf{e}_i.\end{aligned}\tag{57}$$

Note that the term $\hat{\hat{\mathbf{s}}}_i$ is a deterministic function of $\mathbf{y}^{i-1}$ and thus has no effect on the estimation error. To show that (41) is a state-space model that admits standard Kalman filtering, note that the measurement noise $\Lambda\mathbf{m}_i+\mathbf{e}_i$ is independent of $\mathbf{z}^{i-1}$. Thus, the measurement noise is independent of previous measurements $\mathbf{y}^{i-1}$ and the hidden states $\hat{\mathbf{s}}^{i-1}$ of the state-space model.

To obtain the optimal estimator and the error covariance recursion in (43), we apply the standard Kalman filter recursions (7)-(9) that also hold with the time-varying constants $G=K_{p,i}$, $H=\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}+H$, $W=S=\Psi_i$ and $V=\Lambda M_i\Lambda^T+\Psi_i$. □

Proof of Lemma 4: The starting point is the combination of Lemma 2 and Lemma 3 to the optimization problem of $C_n(P)$

$$\begin{aligned}&\max\frac{1}{2n}\sum_{i=1}^{n}\log\det(\Psi_{Y,i})-\log\det(\Psi_i)\\ &\quad\text{s.t.}\quad\frac{1}{n}\sum_{i=1}^{n}\mathbf{Tr}(\Gamma_i\hat{\Sigma}_i^{\dagger}\Gamma_i^T+M_i)\le P,\\ &\Gamma_i(I-\hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i)=0,\quad M_i\succeq 0\\ &\Psi_{Y,i}=(\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}+H)\hat{\Sigma}_i(\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}+H)^T+\Lambda M_i\Lambda^T+\Psi_i\\ &K_{Y,i}=(F\hat{\Sigma}_i(\Lambda\Gamma_i\hat{\Sigma}_i^{\dagger}+H)^T+K_{p,i}\Psi_i)\Psi_{Y,i}^{-1}\\ &\hat{\Sigma}_{i+1}=F\hat{\Sigma}_iF^T+K_{p,i}\Psi_iK_{p,i}^T-K_{Y,i}\Psi_{Y,i}K_{Y,i}^T\end{aligned}\tag{58}$$

with the initial condition $\hat{\Sigma}_1=0$. The maximum is over all involved variables, that is, $\{\Gamma_i,M_i,\hat{\Sigma}_{i+1}\}_{i=1}^n$.

The first step is to introduce an auxiliary decision variable

$$\Pi_i=\Gamma_i\hat{\Sigma}_i^{\dagger}\Gamma_i^T+M_i.\tag{59}$$

Then, $\Psi_{Y,i}$ can be written as

$$\Psi_{Y,i}=\Lambda\Pi_i\Lambda^T+H\hat{\Sigma}_iH^T+\Lambda\Gamma_iH^T+H\Gamma_i^T\Lambda^T+\Psi_i,$$

where we used the orthogonality constraint $\Gamma_i(I-\hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i)=0$. As a result, the Riccati recursion can also be represented with Π_i only. The power constraint can also be expressed as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Tr}(\Pi_i)\le P,\tag{60}$$

so that the variable M_i only appears in the constraints

$$\begin{aligned}&\Pi_i=\Gamma_i\hat{\Sigma}_i^{\dagger}\Gamma_i^T+M_i\\ &M_i\succeq 0,\end{aligned}\tag{61}$$

which can be reduced to the constraint

$$\Pi_i\succeq\Gamma_i\hat{\Sigma}_i^{\dagger}\Gamma_i^T.\tag{62}$$

By the Schur complement for positive semidefinite matrices [47, p. 651],

$$\begin{aligned}&\hat{\Sigma}_i\succeq 0\ \&\ \Pi_i-\Gamma_i\hat{\Sigma}_i^{\dagger}\Gamma_i^T\succeq 0\ \&\ \Gamma_i(I-\hat{\Sigma}_i^{\dagger}\hat{\Sigma}_i)=0\\ &\iff\begin{pmatrix}\Pi_i & \Gamma_i\\ \Gamma_i^T & \hat{\Sigma}_i\end{pmatrix}\succeq 0.\end{aligned}\tag{63}$$

Finally, the Riccati equation is relaxed to the Riccati inequality

$$\hat{\Sigma}_{i+1}\preceq F\hat{\Sigma}_iF^T+K_{p,i}\Psi_iK_{p,i}^T-K_{Y,i}\Psi_{Y,i}K_{Y,i}^T,\tag{64}$$

and using the Schur complement transformation, we can write

$$\begin{pmatrix}F\hat{\Sigma}_iF^T+K_{p,i}\Psi_iK_{p,i}^T-\hat{\Sigma}_{i+1} & K_{Y,i}\Psi_{Y,i}\\ \Psi_{Y,i}K_{Y,i}^T & \Psi_{Y,i}\end{pmatrix}\succeq 0.\tag{65}$$

□

Proof of Lemma 5: This is the converse proof for the capacity expression in Theorem 1. Recall that throughout the derivations, we used the n-letter capacity $C_n(P)$ in (36), but a standard converse argument can relate this quantity to the feedback capacity by showing that for any n,

$$C_{fb}(P)\le\frac{1}{n}C_n(P)+\delta_n\tag{66}$$

where $\delta_n\to 0$ is resulted from a Fano's inequality. The remaining step is to show that the SCOP formulation in Lemma 4 that serves as an upper bound to $C_n(P)$ can be further upper bounded by its single-letter counterpart, the optimization problem in Theorem 1.

Define the convex combinations of the decision variables as

$$\bar{\Pi}_n=\frac{1}{n}\sum_{i=1}^{n}\Pi_i,\quad\bar{\Gamma}_n=\frac{1}{n}\sum_{i=1}^{n}\Gamma_i,\quad\bar{\hat{\Sigma}}_n=\frac{1}{n}\sum_{i=1}^{n}\hat{\Sigma}_i,\tag{67}$$

and also let $\bar{\Sigma}_n\triangleq\frac{1}{n}\sum_{i=1}^{n}\Sigma_i$, $\bar{\Psi}_n\triangleq\frac{1}{n}\sum_{i=1}^{n}\Psi_i$ denote the averaged constants of the Riccati variables.

The concavity of the $\log\det(\cdot)$ function and Jensen's inequality imply that the convex combinations attain a greater objective than the one in Lemma 4,

$$\frac{1}{n}\sum_{i=1}^{n}\log\det(\Psi_{Y,i})\le\log\det\left(\frac{1}{n}\sum_{i=1}^{n}\Psi_{Y,i}\right),\tag{68}$$

where the argument of the right-hand side can be written as the linear function

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Psi_{Y,i}\\ &=\Lambda\bar{\Pi}_n\Lambda^T+H\bar{\hat{\Sigma}}_nH^T+\Lambda\bar{\Gamma}_nH^T+H\bar{\Gamma}_n^T\Lambda^T+H\bar{\Sigma}_nH^T.\end{aligned}$$

Next, the per-time constraints of the n-letter problem should be transformed into their single-letter counterparts, that is, the ones evaluated at the convex combinations in (67). It is straightforward to show that the power constraint and the first LMI constraint are satisfied at the convex combination by

$$\begin{aligned}\mathrm{Tr}(\bar{\Pi}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(\Pi_i)\\ &\le P,\end{aligned} \tag{69}$$

and

$$\begin{aligned}\begin{pmatrix}\bar{\Pi}_n & \bar{\Gamma}_n\\ \bar{\Gamma}_n^T & \bar{\hat{\Sigma}}_n\end{pmatrix} &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\Pi_i & \Gamma_i\\ \Gamma_i^T & \hat{\Sigma}_i\end{pmatrix}\\ &\succeq 0.\end{aligned} \tag{70}$$

We proceed to the last constraint in the optimization problem, the Riccati LMI, defined by

$$\Omega(\Pi,\hat{\Sigma},\Gamma) = \begin{pmatrix}F\hat{\Sigma}F^T - \hat{\Sigma} + K_p\Psi K_p^T & K_Y(\Gamma,\hat{\Sigma})\\ K_Y(\Gamma,\hat{\Sigma})^T & \Psi_Y(\Gamma,\hat{\Sigma},\Pi)\end{pmatrix} \tag{71}$$

with

$$\begin{aligned}K_Y(\Gamma,\hat{\Sigma}) &= F\Gamma^T\Lambda^T + F\hat{\Sigma}H^T + K_p\Psi\\ \Psi_Y(\Gamma,\hat{\Sigma},\Pi) &= \Lambda\Pi\Lambda^T + H\hat{\Sigma}H^T + \Lambda\Gamma H^T + H\Gamma^T\Lambda^T + \Psi.\end{aligned}$$

The main challenge is that the Riccati LMI does not satisfy $\Omega(\bar{\Pi}_n,\bar{\Sigma}_n,\bar{\Gamma}_n)\succeq 0$ for all n. In other words, the tuple of convex combinations in (67) does not lie in the constraint set of the convex optimization in Theorem 1. Our strategy is to show that the limiting tuple of convex combinations (as a function of n) lies in the required constraint set. This is achieved by showing that the tuple of convex combinations lies in a relaxed constraints set, parameterized with some $\epsilon>0$. We then show that ϵ can be made small as n grows large and argue that there is a limit point that lies in the constraints set that corresponds to $\epsilon=0$.

Define the ϵ-domain of the constraints set as

$$\mathcal{C}_\epsilon = \{\begin{pmatrix}\Pi & \Gamma\\ \Gamma^T & \hat{\Sigma}\end{pmatrix}\succeq 0 : \Omega(\Pi,\hat{\Sigma},\Gamma)+\epsilon I\succeq 0, \mathbf{Tr}(\Pi)\le P\},$$

and note that $\mathcal{C}_0$ is the constraints set in Theorem 1.

By summing over both sides of the Riccati inequality in (65), we have

$$\begin{aligned}&\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\hat{\Sigma}_{i+1} & 0\\ 0 & 0\end{pmatrix}\\ &\quad\preceq \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}F\hat{\Sigma}_iF^T + K_{p,i}\Psi_iK_{p,i}^T & K_{Y,i}\Psi_{Y,i}\\ \Psi_{Y,i}K_{Y,i}^T & \Psi_{Y,i},\end{pmatrix}\end{aligned}$$

Arranging both sides and using the fact that $\Sigma_{n+1}\succeq 0$,

$$\begin{aligned}\begin{pmatrix}\bar{\hat{\Sigma}}_n & 0\\ 0 & 0\end{pmatrix} &\preceq \begin{pmatrix}F\bar{\hat{\Sigma}}_nF^T + K_p\Psi K_p^T & K_Y(\bar{\Gamma}_n,\bar{\hat{\Sigma}}_n)\\ K_Y(\bar{\Gamma}_n,\bar{\hat{\Sigma}}_n)^T & \Psi_Y(\bar{\Gamma}_n,\bar{\hat{\Sigma}}_n,\bar{\Pi}_n)\end{pmatrix}\\ &\quad+\begin{pmatrix}\bar{\Psi}_n-\Psi & F(\bar{\Sigma}_n-\Sigma)H^T\\ H(\bar{\Sigma}_n-\Sigma)F^T & \bar{\Psi}_n-\Psi\end{pmatrix}.\end{aligned}$$

By our assumptions on the state-space model of the noise, we can use [27, Ch. 14] to have $\bar{\Sigma}_n\to\Sigma$ and $\bar{\Psi}_n\to\Psi$. Thus, the constraint on $\Omega(\bar{\Pi}_n,\bar{\hat{\Sigma}}_n,\bar{\Gamma}_n)$ is satisfied *asymptotically*. Specifically, for any ϵ, there exists an n_ϵ such that for all $n>n_\epsilon$

$$0\preceq\Omega(\bar{\Pi}_n,\bar{\hat{\Sigma}}_n,\bar{\Gamma}_n)+\epsilon I. \tag{72}$$

Since the set $\mathcal{C}_\epsilon$ is closed and nested (in ϵ), the sequence $\{(\bar{\Pi}_n,\bar{\hat{\Sigma}}_n,\bar{\Gamma}_n)\}_{n\in\mathbb{N}}$ has a limit point in $\bigcap_{\epsilon>0}\mathcal{C}_\epsilon=\mathcal{C}_0$. That is, there exists a sequence of times $T_1\le T_2\le T_3\dots$ such that $\lim_{i\to\infty}(\bar{\Pi}_{T_i},\bar{\hat{\Sigma}}_{T_i},\bar{\Gamma}_{T_i})\in\mathcal{C}_0$. It is important to note that the times sequence depends on the noise characteristics and not on the underlying codebooks. The proof is completed by taking the limit over the sequence $T_1,T_2,\dots$ in (66) to obtain

$$C_{fb}(P)\le\max_{(\Pi,\hat{\Sigma},\Gamma)\in\mathcal{C}_0}\frac{1}{2}\log\det(\Psi_Y(\Gamma,\hat{\Sigma},\Pi))-\frac{1}{2}\log\det(\Psi),$$

which is precisely the optimization problem in (46). □

Proof of Lemma 6: This is the achievability proof of the optimization problem in Lemma 6. The main ides is to fix a time-invariant policy and analyze the achievable rate which is determined by the asymptotic behaviour of the channel outputs process. Since the channel outputs process is described as a state-space, the asymptotic behavior of the channel outputs statistics boils down to the analysis of Riccati recursion convergence. To that end, we will use a result from [48] on certain conditions to guarantee the convergence of the Riccati recursion to the Riccati equation. Lastly, since one of the condition is given on the initial condition of the Riccati recursion (which we have no direct control over), we modify the time-invariant policy at the first time only to guarantee the convergence.

We use the policy in Lemma 2 with $\Gamma_i=\Gamma\hat{\Sigma}_i$ and $M_i=M$ such that the corresponding power satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}(\Gamma\hat{\Sigma}_i\Gamma^T+M)\le P.$$

By Lemma 3, the induced state-space is

$$\begin{aligned}\hat{\mathbf{s}}_{i+1} &= F\,\hat{\mathbf{s}}_i + K_{p,i}\,\mathbf{e}_i,\\ \mathbf{y}_i &= (\Lambda\Gamma+H)\,\hat{\mathbf{s}}_i - \Lambda\Gamma\,\hat{\hat{\mathbf{s}}}_i + \Lambda\mathbf{m}_i + \mathbf{e}_i,\end{aligned} \tag{73}$$

and the corresponding Riccati recursion is

$$\hat{\Sigma}_{i+1} = F\hat{\Sigma}_iF^T + K_{p,i}\Psi_iK_{p,i}^T - K_{L,i}\Psi_{L,i}K_{L,i}^T \tag{74}$$

with $\hat{\Sigma}_1=0$ and

$$\begin{aligned}\Psi_{L,i} &= (\Lambda\Gamma+H)\hat{\Sigma}_i(\Lambda\Gamma+H)^T + \Lambda M\Lambda^T + \Psi_i\\ K_{L,i} &= (F\hat{\Sigma}_i(\Lambda\Gamma+H)^T + K_{p,i}\Psi_i)\Psi_{L,i}^{-1}.\end{aligned} \tag{75}$$

The next step is to show the convergence of the Riccati recursion in (74) to a fixed-point solution of the Riccati equation. Since $K_{p,i}$ and Ψ_i converge to their time-invariant counterparts in (13) exponentially fast, we replace $K_{p,i}$ and Ψ_i with K_p and Ψ, respectively. This comes at the cost that the initial condition $\hat{\Sigma}_1=0$ becomes arbitrary.

Before presenting the convergence conditions, we need to modify the Riccati recursion in (74) to have an equivalent form with the property that the disturbance and the measurement of the state are independent. This is a standard modification can

be found for instance in [27, Sec. 14.7]. The equivalent form of (74) can be written as

$$\hat{\Sigma}_{i+1} = F_s \hat{\Sigma}_i F_s^T + K_p Q_s K_p^T - \bar{K}_{L,i} \Psi_{L,i} \bar{K}_{L,i}^T, \tag{76}$$

and

$$\begin{aligned} F_s &= F - K_p \Psi (\Lambda M \Lambda^T + \Psi)^{-1} (\Lambda \Gamma + H) \\ Q_s &= \Psi - \Psi (\Lambda M \Lambda^T + \Psi)^{-1} \Psi \\ \bar{K}_{L,i} &= F_s \hat{\Sigma}_i (\Lambda \Gamma + H)^T \Psi_{L,i}^{-1}. \end{aligned} \tag{77}$$

We use [48, Th. 1] for the convergence of the Riccati recursion in (76) to the maximal solution of the Riccati equation, the maximal solution $\hat{\Sigma}_s$ whose all of its closed-loop modes are inside or on the unit circle, that is, $\rho(F_s - \bar{K}_{L,i}(\Lambda\Gamma + H)) \leq 1$. The sufficient condition from [48] translates to the Riccati equation in (76) as

1) The initial state satisfies $\hat{\Sigma}_1 \succeq \hat{\Sigma}_s$.
2) The pair $(F_s, \Lambda\Gamma + H)$ is detectable.

The detectability condition guarantees the existence of the maximal solution. This condition will be carried to the lower bound optimization problem as a restriction on the optimization parameters (Γ, M). Also note that $(F_s, \Lambda\Gamma + H)$ is detectable iff $(F, \Lambda\Gamma + H)$ is detectable and thus can be expressed as $\exists K : \rho(F - K(\Lambda\Gamma + H)) < 1$. The first condition is needed for the convergence to the maximal solution and is shown next. As mentioned, the initial condition $\hat{\Sigma}_1$ is arbitrary. To this end, we modify the time-invariant policy by changing M_1 to be an identity matrix scaled with a constant α.

We proceed to show that the null-space of $\hat{\Sigma}_2$ lies in the null-space of any solution to the Riccati equation. For this proof, we use the closed-loop Lyapunov recursion of (76) can be expressed as

$$\begin{aligned} \hat{\Sigma}_2 = (F_s - \bar{K}_{L,1}(\Lambda\Gamma + H)) \hat{\Sigma}_1 (F_s - \bar{K}_{L,1}(\Lambda\Gamma + H))^T \\ + K_p Q_s K_p^T + \bar{K}_{L,1}(\Lambda M_1 \Lambda^T + \Psi) \bar{K}_{L,1}^T. \end{aligned} \tag{78}$$

Let x be an eigenvector of F with λ such that $x\hat{\Sigma}_2 = 0$. Then, pre- and post- multiplying the closed-loop Riccati equation in (76) with x and x^T we have

$$\begin{aligned} 0 = x(F_s - \bar{K}_{L,1}(\Lambda\Gamma + H)) \hat{\Sigma}_1 (F_s - \bar{K}_{L,1}(\Lambda\Gamma + H))^T x^T \\ + x K_p Q_s K_p^T x^T + x \bar{K}_{L,1}(\Lambda M_1 \Lambda^T + \Psi) \bar{K}_{L,1}^T x^T. \end{aligned} \tag{79}$$

Then, we have $xK_pQ_s = 0$, $x\bar{K}_{L,1} = 0$ which also implies $xF_s\hat{\Sigma} = 0$. By $M_1 \succ 0$, we have $Q_s \succ 0$ so that $xK_p = 0$. Now, consider any solution to the Riccati equation. Then, pre- and post- multiplying the Riccati equation with x and x^T gives

$$x\hat{\Sigma}x^T = xF_s\hat{\Sigma}F_s^T x^T + xK_pQ_sK_p^T x^T - x^T \bar{K}_L \Psi_L \bar{K}_L^T x^T,$$

which implies $x\hat{\Sigma}x^T(1 - |\lambda|^2) \preceq 0$. Finally, by the stability of $F - K_pH$, the equation $xK_p = 0$ implies $|\lambda| < 1$ and therefore $x\hat{\Sigma} = 0$. To conclude the proof of the first item, we can choose α to be large enough such that the error covariance $\hat{\Sigma}_2 \succeq \Sigma_s$. Note that the power constraint may be violated for small n but it will average out when taking n to be large enough.

To summarize, for any time-invariant policy (M, Γ) subject to the detectability condition, the channel outputs entropy rate converges to

$$\lim_{n\to\infty} \frac{1}{n} h(Y^n) = \frac{1}{2} \log\det(\Psi_{Y,s}) + \frac{1}{2} \log(2\pi e)^d \tag{80}$$

where $\Psi_{Y,s}$ is the innovation covariance of the Riccati equation in (76) evaluated at its (unique) maximal solution.

As shown in [2], the asymptotic equipartition property (AEP) holds for arbitrary Gaussian processes, so that $\lim_{n\to\infty} \frac{1}{n}(h(Y^n) - h(Z^n))$ is achievable for any policy of the form $X^n = B_n Z^n + V^n$ where $V^n \sim (0, \Sigma_{V_n})$ is independent of Z^n and B_n is a (block) lower-triangular matrix, i.e., it is a strictly causal operator. The policy considered here can be written in this form since $\hat{\mathbf{s}}_i$ is a strictly causal function of $\{\mathbf{z}_i\}_{i\geq 1}$ and $\hat{\hat{\mathbf{s}}}_i$ is a strictly causal function of $\{\mathbf{y}_i\}_{i\geq 1}$. Thus, we have that

$$C_{fb}(P) \geq \frac{1}{2} \log\det(\Psi_{Y,s}) - \frac{1}{2} \log\det(\Psi). \tag{81}$$

We formulate an optimization problem which serves as a lower bound on the feedback capacity. By taking a maximum over all valid policies, we have

$$\begin{aligned} C_{fb}(P) \geq \max_{\Gamma, M, \hat{\Sigma}_s} \; & \frac{1}{2} \log\det(\Psi_{Y,s}) - \frac{1}{2} \log\det(\Psi) \\ \text{s.t.} \;\; & \mathrm{Tr}(\Gamma \hat{\Sigma}_s \Gamma^T + M) \leq P \\ & \hat{\Sigma}_s = F\hat{\Sigma}_s F^T + K_p \Psi K_p^T - K_L \Psi_L K_L^T \\ & K_L = (F\hat{\Sigma}_s(\Lambda\Gamma + H)^T + K_p\Psi)\Psi_L^{-1} \\ & \Psi_{Y,s} = (\Lambda\Gamma + H)\hat{\Sigma}_s(\Lambda\Gamma + H)^T + \Lambda M \Lambda^T + \Psi \\ & \exists K : \rho(F - K(\Lambda\Gamma + H)) < 1, \end{aligned} \tag{82}$$

To complete the proof, change the variable $\Gamma' = \Gamma\hat{\Sigma}_s$, add the orthogonality constraint and follow the steps in Lemma 5: define $\Pi = \Gamma\hat{\Sigma}_s^\dagger \Gamma^T + M$, reduce M and apply the Schur complement to get the optimization problem (82). For consistency with the upper bound notation, we rename Γ' and $\hat{\Sigma}_s$ with Γ and $\hat{\Sigma}$ respectively. □

Proof of Lemma 7: Recall that from the upper bound optimization problem, the tuple $(\Pi, \hat{\Sigma}, \Gamma)$ satisfies

$$\hat{\Sigma} \preceq F\hat{\Sigma}F^T + K_p\Psi K_p^T - K_Y \Psi_Y K_Y^T, \tag{83}$$

with

$$\begin{aligned} \Psi_Y &= \Lambda\Pi\Lambda^T + H\hat{\Sigma}H^T + \Lambda\Gamma H^T + H\Gamma^T\Lambda^T + \Psi \\ &= (\Lambda\Gamma\hat{\Sigma}^\dagger + H)\hat{\Sigma}(\Lambda\Gamma\hat{\Sigma}^\dagger + H)^T + \Lambda(\Pi - \Gamma\hat{\Sigma}^\dagger\Gamma^T)\Lambda^T + \Psi \\ K_Y &= (F\hat{\Sigma}(\Lambda\Gamma\hat{\Sigma}^\dagger + H)^T + K_p\Psi)\Psi_Y^{-1}. \end{aligned} \tag{84}$$

We prove the claims.

1) If the optimal tuple does not satisfy the Riccati inequality (83) with equality, there exists a matrix $Q \succeq 0$ such that

$$Q \triangleq F\hat{\Sigma}F^T - \hat{\Sigma} + K_p\Psi K_p^T - K_Y\Psi_Y K_Y^T \tag{85}$$

is not the zero matrix. We let $\hat{\Sigma}' = Q + \hat{\Sigma}$, and observe that this modification satisfies the power constraint, and the LMI

$$\begin{pmatrix} \Pi & \Gamma \\ \Gamma^T & \hat{\Sigma}' \end{pmatrix} \succeq 0.$$

Then, using $\hat{\Sigma}' \succeq \hat{\Sigma}$ and the optimality of the tuple $(\Gamma, \Pi, \hat{\Sigma})$, we conclude that the objective is still equal to its optimal value after the modification.

2) If there exists an unstable mode in F that cannot be observed via $\Lambda\Gamma\Sigma^{\dagger}+H$, by our assumption that (F,H) is detectable, this mode can be observed via $\Lambda\Gamma\hat{\Sigma}^{\dagger}$. On the other hand, the instability of this mode implies that the error covariance $\hat{\Sigma}$ has an infinite value in this direction which is a contradiction to the observability of this mode via the matrix $\Lambda\Gamma\hat{\Sigma}^{\dagger}$. □

A. Proof for the Coding Scheme Analysis

Proof of Lemma 1: The proof follows a sequential estimation argument of estimating $\mathbf{z}_0$ from $\mathbf{y}_1^n = \mathbf{y}_1,\ldots,\mathbf{y}_n$. At each time, a new measurement (i.e., channel output) is made available to the decoder that improves in turn its estimate of the channel noise instance $\mathbf{z}_0$. The derivation mostly focuses on writing the channel output as a simple linear function of $\mathbf{z}_0$, and then apply known recursive formulas for updating the estimate and the error covariance with a new measurement. We iterate that the derivations here hold for general MIMO channels.

Recall that the channel output can be written as

$$\begin{aligned}\mathbf{y}_n &= \Lambda\,\mathbf{x}_n+\mathbf{z}_n\\ &\overset{(a)}{=} \Lambda\Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)+H\,\hat{\mathbf{s}}_n+(\mathbf{z}_n-H\,\hat{\mathbf{s}}_n)\\ &= (\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)+\mathbf{e}_n+H\,\hat{\hat{\mathbf{s}}}_n, \end{aligned}\tag{86}$$

where (a) follows from the channel input $\mathbf{x}_n = \Lambda\Gamma\hat{\Sigma}^{\dagger}(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)$. We now relate the channel output $\mathbf{y}_n$ and $\mathbf{z}_0$. To this end, the estimation error can be written as the recursion

$$\begin{aligned}\hat{\mathbf{s}}_{n+1}-\hat{\hat{\mathbf{s}}}_{n+1} &= F\,\hat{\mathbf{s}}_n+K_p\,\mathbf{e}_n-(F\,\hat{\hat{\mathbf{s}}}_n+K_{Y,n}(\mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n))\\ &\overset{(a)}{=} F(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)-K_{Y,n}(\mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n)\\ &\quad+K_p(\mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n-(\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n))\\ &= (F-K_p(\Lambda\Gamma\Sigma^{\dagger}+H))(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)\\ &\quad+(K_p-K_{Y,n})(\mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n)\\ &\overset{(b)}{=} F_p(\hat{\mathbf{s}}_n-\hat{\hat{\mathbf{s}}}_n)+(K_p-K_{Y,n})\tilde{\mathbf{y}}_n\\ &= F_p^n(\hat{\mathbf{s}}_1-\hat{\hat{\mathbf{s}}}_1)+\sum_{i=1}^{n}F_p^{n-i}(K_p-K_{Y,i})\tilde{\mathbf{y}}_i\\ &\overset{(c)}{=} F_p^nK_p\,\mathbf{z}_0+\mathbf{d}_n, \end{aligned}\tag{87}$$

where (a) follows from (86), (b) follows from $F_p \triangleq F-K_p(\Lambda\Gamma\Sigma^{\dagger}+H)$ and $\tilde{\mathbf{y}}_n \triangleq \mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n$, and (c) follows from $\hat{\mathbf{s}}_1 = K_p\,\mathbf{z}_0$, $\hat{\hat{\mathbf{s}}}_1 = 0$, and $\mathbf{d}_n \triangleq \sum_{i=1}^{n}F_p^{n-i}(K_p-K_{Y,i})\tilde{\mathbf{y}}_i$.

We combine (86) and (87) to write the channel output as

$$\begin{aligned}\mathbf{y}_n &= (\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)(F_p^{n-1}K_p\,\mathbf{z}_0+\mathbf{d}_{n-1})+\mathbf{e}_n+H\,\hat{\hat{\mathbf{s}}}_n\\ &= \kappa_n\,\mathbf{z}_0+(\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)\mathbf{d}_{n-1}+\mathbf{e}_n+H\,\hat{\hat{\mathbf{s}}}_n, \end{aligned}\tag{88}$$

with $\kappa_n \triangleq (\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)F_p^{n-1}K_p$. The estimation model in (88) is a sequential estimation problem, but the terms $\mathbf{d}_{n-1}$ and $H\,\hat{\hat{\mathbf{s}}}_n$ on the right-hand side depend on the previous measurement. We proceed to show that these *bias terms* have no effect on the estimation problem and thus can be ignored. Define the transformed measurements (channel outputs)

$$\begin{aligned}\mathbf{o}_n &\triangleq \tilde{\mathbf{y}}_n-(\Lambda\Gamma\hat{\Sigma}^{\dagger}+H)\mathbf{d}_{n-1}\\ &= \kappa_n\,\mathbf{z}_0+\mathbf{e}_n, \end{aligned}\tag{89}$$

in order to obtain a sequential estimation problem (without the bias terms) where the source is $\mathbf{z}_0$, and at each time we observe $\mathbf{o}_n$ with the measurement noise $\mathbf{e}_n$. The transformation $\{\tilde{\mathbf{y}}_i\}_{i=1}^n \to \{\mathbf{o}_i\}_{i=1}^n$ is linear and causal (lower-triangular). Also note that this transformation is invertible since $\mathbf{d}_{n-1}$ is a function of $\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_{n-1}$ only. Informally, the invertibility of this transformation shows that the information that can be extracted from the original channel outputs and the transformed channel outputs is the same. Formally, the innovations of both processes are the same, i.e.,

$$\mathbf{o}_n-\mathbb{E}[\mathbf{o}_n|\mathbf{o}^{n-1}] = \mathbf{y}_n-\mathbb{E}[\mathbf{y}_n\,|\,\mathbf{y}^{n-1}].\tag{90}$$

We are now ready to present the recursions for the sequential estimation problem in (89). Since the source is the same at all times (i.e., $\mathbf{z}_0$), we only need a measurement-update formula (e.g., [27, Lemma 9.3.2]) to write it recursively as

$$\begin{aligned}\hat{\mathbf{z}}_{0|n} &= \hat{\mathbf{z}}_{0|n-1}\\ &\quad+\hat{Z}_{0|n-1}\kappa_n^T\,\mathbf{cov}(\mathbf{y}_n\,|\,\mathbf{y}^{n-1})^{-1}(\mathbf{y}_n-\mathbb{E}[\mathbf{y}_n\,|\,\mathbf{y}^{n-1}])\\ \hat{Z}_{0|n+1} &= \hat{Z}_{0|n}\\ &\quad-\hat{Z}_{0|n}\kappa_n^T\,\mathbf{cov}(\mathbf{y}_n\,|\,\mathbf{y}^{n-1})^{-1}\kappa_n\hat{Z}_{0|n} \end{aligned}\tag{91}$$

with the initial conditions $\hat{\mathbf{z}}_{0|0} = 0$ and $\hat{Z}_{0|0} = \Psi$. By Lemma 3, we have $\Psi_{Y,n} = \mathbf{cov}(\mathbf{y}_n\,|\,\mathbf{y}^{n-1})$ and $\mathbb{E}[\mathbf{y}_n\,|\,\mathbf{y}^{n-1}] = H\,\hat{\hat{\mathbf{s}}}_n$ so that the recursions simplify to

$$\begin{aligned}\hat{\mathbf{z}}_{0|n} &= \hat{\mathbf{z}}_{0|n-1}+\hat{Z}_{0|n-1}\kappa_n^T\Psi_{Y,n}^{-1}(\mathbf{y}_n-H\,\hat{\hat{\mathbf{s}}}_n)\\ \hat{Z}_{0|n} &= (I-\hat{Z}_{0|n-1}\kappa_n^T\Psi_{Y,n}^{-1}\kappa_n)\hat{Z}_{0|n-1}. \end{aligned}\tag{92}$$

Furthermore, due to the optimal inputs distribution, the innovation covariance $\Psi_{Y,n}$ converges to its optimal value Ψ_Y^* (for more details, see the proof of Lemma 6). Finally, taking a determinant over (92), applying Sylvester's identity, and note that $\Psi_{Y,n} = \kappa_n\hat{Z}_{0|n-1}\kappa_n^T+\Psi$ in (90) gives (32). □

VII. Conclusion and Future Work

In this paper, we solved the feedback capacity problem of the Gaussian MIMO channel when the noise is generated from a linear dynamical system. The derivation relies on a sequential convex optimization formulation for the finite-block capacity problem using tools from control theory and convex optimization methods. Using the optimization problem convexity along with properties of Riccati recursions convergence, we provided tight lower and upper bounds that resulted a single-letter, computable capacity expression. Additionally, we showed that that the optimization problem induces a time-invariant capacity-achieving inputs distribution that was used to construct an explicit coding scheme for scalar channels.

In a broader perspective, we derived a single-letter formula for the directed information and its main steps can be summarized as follows

$$\begin{aligned}I(X^n\to Y^N) &= \sum_{i=1}^{n}I(X^i;Y_i|Y^{i-1})\\ &\overset{(a)}{=} \sum_{i=1}^{n}I(X_i,\hat{S}_i(X^{i-1},Y^{i-1});Y_i|Y^{i-1}) \end{aligned}$$

$$\stackrel{(b)}{=} \sum_{i=1}^{n} I(X_i, \hat{S}_i(X^{i-1}, Y^{i-1}); Y_i|\hat{\hat{S}}_i(Y^{i-1}))$$
$$\approx nI(X, \hat{S}; Y|\hat{\hat{S}}), \tag{93}$$

where in (a) a channel state $\hat{S}_i = \mathbb{E}[S_i|X^{i-1}, Y^{i-1}] = \mathbb{E}[S_i|Z^{i-1}]$ is defined and satisfies the channel Markov chain $(\hat{S}_{i+1}, Y_i) - (X_i, \hat{S}_i) - (X^{i-1}, Y^{i-1}, \hat{S}^{i-1})$, and in (b) $\hat{\hat{S}}_i \triangleq \mathbb{E}[\hat{S}_i|Y^{i-1}]$. Note that the channel state $\hat{S}_i$ can be computed at the encoder since it is a function of (X^{i-1}, Y^{i-1}). Also, the computation of the asymptotic behaviour at the last step was enabled due to the description of the channel outputs process structure as a hidden-Markov (Lemma 3).

The above steps are related to computations of the directed information for the discrete-alphabet counterpart of the Gaussian channel, the finite-state channel (FSC). More specifically, for FSCs with state that can be computed at the encoder, the directed information can be written as

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X_i, S_i; Y_i|Y^{i-1}), \tag{94}$$

and could be expressed with a computable expression in few instances only [41], [44], [49], [50], [51], [52], [53], [54]. In [55] and [56], it was shown that all these solutions can be unified with the single-letter expression $I(X, S; Y|Q)$, where the channel outputs is a hidden Markov model and Q serves as its hidden state with a finite, graphical structure (called the Q-graph). As the conjectured formula structure resembles the one for the Gaussian channel in (93), it should be interesting to investigate whether the techniques developed here apply also for FSCs. In particular, the main step is the formulation of the directed information as a sequential convex optimization problem in order to have an alternative feedback capacity formula that can be *single-letterized*.

Two more research directions are as follows.

1) Explicit Formulae: It may be possible to find simple capacity expressions for particular noise processes using the convex optimization in Theorem 1. For instance, the capacity of the ARMA noise of first order can be expressed as a function of the positive root to a quartic equation [21]. This implies that the two decision variables in Theorem 2 can be reduced to a single variable. Pursuing such simplifications for ARMA processes of higher order is natural [11], [12], [13], [17].

2) Scheme for MIMO Channels: In Section IV, we presented an explicit scheme for scalar channel that trivially extends to MIMO channels that can be decomposed to parallel scalar channels. However, an explicit scheme for non-trivial MIMO channels remains open. A conjectured scheme was described in Section IV, and a refinement of the spectral analysis in Lemma 1 should prove the scheme optimality.

ACKNOWLEDGMENT

The authors would like to thank an Associate Editor and the reviewers for their valuable and constructive comments, which helped to improve this article.

REFERENCES

[1] O. Sabag, V. Kostina, and B. Hassibi, "Feedback capacity of MIMO Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 7–12.
[2] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.
[3] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 2, pp. 499–533, Apr. 1998.
[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA, USA: SIAM, 1994.
[5] C. Scherer and S. Weiland, "Linear matrix inequalities in control," Dutch Inst. Syst. Control, Delft, The Netherlands, Lect. Notes, 2000, vol. 3, no. 2.
[6] R. J. Caverly and J. Richard Forbes, "LMI properties and applications in systems, stability, and control theory," 2019, *arXiv:1903.08599*.
[7] O. Sabag, P. Tian, V. Kostina, and B. Hassibi, "The minimal directed information needed to improve the LQG cost," in *Proc. 59th IEEE Conf. Decis. Control (CDC)*, Dec. 2020, pp. 1842–1847.
[8] T. Tanaka, K. K. Kim, P. A. Parrilo, and S. K. Mitter, "Semidefinite programming approach to Gaussian sequential rate-distortion trade-offs," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1896–1910, Apr. 2017.
[9] T. Tanaka, P. M. Esfahani, and S. K. Mitter, "LQG control with minimum directed information: Semidefinite programming approach," *IEEE Trans. Autom. Control*, vol. 63, no. 1, pp. 37–52, Jan. 2018.
[10] A. Gattami, "Feedback capacity of Gaussian channels revisited," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1948–1960, Mar. 2019.
[11] S. Butman, "A general formulation of linear feedback communication systems with solutions," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 3, pp. 392–400, May 1969.
[12] S. Butman, "Linear feedback rate bounds for regressive channels (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 3, pp. 363–366, May 1976.
[13] J. Tiernan and J. Schalkwijk, "An upper bound to the capacity of the band-limited Gaussian autoregressive channel with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 3, pp. 311–316, May 1974.
[14] P. M. Ebert, "The capacity of the Gaussian channel with feedback," *Bell Syst. Tech. J.*, vol. 49, no. 8, pp. 1705–1712, Oct. 1970.
[15] A. Shahar-Doron and M. Feder, "On a capacity achieving scheme for the colored Gaussian channel with feedback," in *Proc. Int. Symp. Inf. Theory*, 2004, p. 74.
[16] C. Li and N. Elia, "Youla coding and computation of Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 3197–3215, Apr. 2018.
[17] T. Liu and G. Han, "Feedback capacity of stationary Gaussian channels further examined," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2492–2506, Apr. 2019.
[18] S. Fang and Q. Zhu, "A connection between feedback capacity and Kalman filter for colored Gaussian noises," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2055–2060.
[19] N. Elia, "When Bode meets Shannon: Control-oriented feedback communication schemes," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1477–1488, Sep. 2004.
[20] L. H. Ozarow, "Random coding for additive Gaussian channels with feedback," *IEEE Trans. Inf. Theory*, vol. 36, no. 1, pp. 17–22, Jan. 1990.
[21] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.
[22] S. Yang, A. Kavcic, and S. Tatikonda, "On the feedback capacity of power-constrained Gaussian noise channels with memory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 929–954, Mar. 2007.
[23] Y.-H. Kim, "Feedback capacity of the first-order moving average Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 3063–3079, Jul. 2006.
[24] M. S. Derpich and J. Ostergaard, "Comments on 'feedback capacity of stationary Gaussian channels,'" *IEEE Trans. Inf. Theory*, early access, Jun. 10, 2022, doi: 10.1109/TIT.2022.3182270.
[25] A. Rawat and N. Elia, "Feedback capacity of ISI MIMO channel with colored noise," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.
[26] O. Sabag, V. Kostina, and B. Hassibi, "Feedback capacity of Gaussian channels with memory," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2022, pp. 2547–2552.
[27] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ, USA: Prentice-Hall, 2000.
[28] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, Signal Inf. Process. Lab., Swiss Federal Inst. Technol. (ETH) Zurich, 1998.
[29] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
[30] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.

[31] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Nov. 1990, pp. 303–305.
[32] C. D. Charalambous, C. Kourtellaris, and S. Louka, "New formulas of feedback capacity for AGN channels with memory: A time-domain sufficient statistic approach," 2020, *arXiv:2010.06226*.
[33] T. Kailath, "An innovations approach to least-squares estimation—Part I: Linear filtering in additive white noise," *IEEE Trans. Autom. Control*, vol. AC-13, no. 6, pp. 646–655, Dec. 1968.
[34] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx
[35] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback-I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 172–182, Apr. 1966.
[36] R. G. Gallager and B. Nakiboglu, "Variations on a theme by Schalkwijk and Kailath," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 6–17, Jan. 2010.
[37] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-9, no. 3, pp. 136–143, Jul. 1963.
[38] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.
[39] M. Naghshvar, T. Javidi, and M. Wigger, "Extrinsic Jensen–Shannon divergence: Applications to variable-length coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2148–2164, Apr. 2015.
[40] C. T. Li and A. El Gamal, "An efficient feedback coding scheme with low error probability for discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 2953–2963, Jun. 2015.
[41] O. Sabag, H. H. Permuter, and N. Kashyap, "Feedback capacity and coding for the BIBO channel with a no-repeated-ones input constraint," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 4940–4961, Jul. 2018.
[42] O. Sabag, H. H. Permuter, and N. Kashyap, "The feedback capacity of the binary erasure channel with a no-consecutive-ones input constraint," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 8–22, Jan. 2016.
[43] J. H. Bae and A. Anastasopoulos, "A posterior matching scheme for finite-state channels with feedback," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 2338–2342.
[44] O. Peled, O. Sabag, and H. H. Permuter, "Feedback capacity and coding for the (0,k)-RLL input-constrained BEC," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4097–4114, Jul. 2019.
[45] S. Ihara, "Upper bounds of error probabilities for stationary Gaussian channels with feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7068–7072, Dec. 2012.
[46] O. Sabag, V. Kostina, and B. Hassibi. (2023). *Implemention of 'Feedback Capacity of Gaussian Channels'*. [Online]. Available: https://github.com/oronsabag/Feedback-capacity-gaussian-coding
[47] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.
[48] G. De Nicolao and M. Gevers, "Difference and differential Riccati equations: A note on the convergence to the strong solution," *IEEE Trans. Autom. Control*, vol. 37, no. 7, pp. 1055–1057, Jul. 1992.
[49] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, pp. 780–789, Mar. 2005.
[50] O. Elishco and H. Permuter, "Capacity and coding for the ising channel with feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5138–5149, Sep. 2014.
[51] A. Sharov and R. M. Roth, "On the capacity of generalized ising channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2338–2356, Apr. 2017.
[52] Z. Aharoni, O. Sabag, and H. H. Permuter, "Computing the feedback capacity of finite state channels using reinforcement learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 837–841.
[53] H. Permuter, P. Cuff, B. Van Roy, and T. Weissman, "Capacity of the trapdoor channel with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3150–3165, Jul. 2008.
[54] J. Wu and A. Anastasopoulos, "On the capacity of the chemical channel with feedback," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 295–299.
[55] O. Sabag, H. H. Permuter, and H. D. Pfister, "A single-letter upper bound on the feedback capacity of unifilar finite-state channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1392–1409, Mar. 2017.
[56] O. Sabag, B. Huleihel, and H. H. Permuter, "Graph-based encoders and their performance for finite-state channels with feedback," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2106–2117, Apr. 2020.

Oron Sabag (Member, IEEE) received the B.Sc. (cum laude), the M.Sc. (summa cum laude), and the Ph.D. degree in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 2013, 2016, and 2019, respectively. From 2019 to 2022, he was a Postdoctoral Fellow at the Department of Electrical Engineering, Caltech. He is currently a Senior Lecturer with the School of Computer Science and Engineering, The Hebrew University of Jerusalem. He is a recipient of several awards, among them are ISEF postdoctoral fellowship, ISIT-2017 best student paper award, SPCOM-2016 best student paper award, the Feder Family Award for outstanding research in communications, the Lachish Fellowship, and the Kaufman award.

Victoria Kostina (Senior Member, IEEE) received the bachelor's degree from the Moscow Institute of Physics and Technology (MIPT) in 2004, the master's degree from the University of Ottawa in 2006, and the Ph.D. degree from Princeton University in 2013. She is a Professor of Electrical Engineering and Computing and Mathematical Sciences at Caltech. During her studies at MIPT, she was affiliated with the Institute for Information Transmission Problems of the Russian Academy of Sciences. Her research program spans finite delay theory of information, random access communications, control over communication channels, and information bottlenecks in computing systems. She has served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY AND FOR ENTROPY. She received the Natural Sciences and Engineering Research Council of Canada postgraduate scholarship during 2009 to 2012, Princeton Electrical Engineering Best Dissertation Award in 2013, Simons-Berkeley research fellowship in 2015, and the NSF CAREER award in 2017.

Babak Hassibi (Member, IEEE) was born in Tehran, Iran, in 1967. He received the B.S. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 1989, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1993 and 1996, respectively. From October 1996 to October 1998, he was a Research Associate with the Information Systems Laboratory, Stanford University. From November 1998 to December 2000, he was a member of the Technical Staff with the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ, USA. Since January 2001, he has been with the California Institute of Technology, Pasadena, CA, USA, where he is currently the Mose and Lilian S. Bohn Professor of Electrical Engineering. From 2013 to 2016, he was the Gordon M. Binder/Amgen Professor of Electrical Engineering and he was an Executive Officer of Electrical Engineering from 2008 to 2015, as well as the Associate Director of the Information Science and Technology. He has also held short-term appointments at the Ricoh California Research Center, Indian Institute of Science, and Linkoping University, Sweden. He is the coauthor of the books (both with A.H. Sayed and T. Kailath) Indefinite Quadratic Estimation and Control: A Unified Approach to H2 and H1 Theories (SIAM, 1999) and Linear Estimation (Prentice Hall, 2000). His research interests include communications and information theory, control and network science, and signal processing and machine learning. He is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck Best Paper Award of the American Automatic Control Council (with H. Hindi and S.P. Boyd), the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering, the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE), and the 2009 Al-Marai Award for Innovative Research in Communications, and he was also a participant in the 2004 National Academy of Engineering "Frontiers in Engineering" program. He has been a Guest Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY Special Issue on "Space-time transmission, reception, coding and signal processing," an Associate Editor for Communications of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2004 to 2006, and he is currently an Editor of the journal Foundations and Trends in Information and Communication and the IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING He was an IEEE Information Theory Society Distinguished Lecturer from 2016 to 2017.