Capacity Achieving Distributions and Separation Principle for Feedback Gaussian Channels With Memory: the LQG Theory of Directed Information

Charalambos D. Charalambous, Christos K. Kourtellaris, and Sergey Loyka

***Abstract*— A method is developed to realize optimal channel input conditional distributions, which maximize the finite transmission feedback information (FTFI) capacity, often called $n-$block length feedback capacity, by information lossless randomized strategies. The method is applied to compute closed form expressions for the FTFI capacity and feedback capacity, of nonstationary, nonergodic, unstable, multiple input multiple output Gaussian channels with memory on past channel outputs, subject to average transmission cost constraints of quadratic form in the channel inputs and outputs. It is shown that randomized strategies decompose into two orthogonal parts-an deterministic part, which controls the channel output process, and an innovation part, which transmits new information over the channel. Then a separation principle is shown between the computation of the optimal deterministic part and the random part of the optimal randomized strategies. Finally, the ergodic theory of linear-quadratic-Gaussian stochastic optimal control theory, is applied to identify sufficient conditions, expressed in terms of solutions to matrix difference and algebraic Riccati equations, so that the optimal control part of randomized strategies induces asymptotic stationarity and ergodicity, and feedback capacity is characterized by the per unit time limit of the FTFI capacity. The method reveals an interaction of the control and the information transmission parts of the optimal randomized strategies, and that whether feedback increases capacity, is directly related to the channel parameters and the transmission cost function, through the solutions of the matrix Riccati equations. For unstable channels, it is shown that feedback capacity exists and it is strictly positive, provided the power exceeds a critical threshold.**

***Index Terms*— Channels with memory, unstable, feedback capacity, separation principle, linear quadratic control.**

I. Introduction

STOCHASTIC optimal control theory and a variational equality of directed information are previously applied in [1], to identify the information structures of optimal channel input conditional distributions with feedback, $\mathscr{P}_{[0,n]} \stackrel{\triangle}{=} \{\mathbf{P}_{A_i|A^{i-1},B^{i-1}} : i = 0,1,\ldots,n\} \subset \mathscr{P}_{[0,n]}(\kappa)$, which maximize the finite-time horizon directed information from channel inputs $A^n \stackrel{\triangle}{=} \{A_0, A_1, \ldots, A_n\}$ to channel outputs $B_0^n \stackrel{\triangle}{=} \{B_0, B_1, \ldots, B_n\}$, given the initial state B^{-1}, defined by

$$C_{A^n \to B^n}(\kappa) = \sup_{\mathscr{P}_{[0,n]}(\kappa)} I(A^n \to B^n),$$
$$I(A^n \to B^n) \stackrel{\triangle}{=} \sum_{i=0}^{n} I(A^i; B_i|B^{i-1}) \tag{I.1}$$

where $B^{i-1} = (B^{-1}, B_0, \ldots, B^{i-1})$ for each $i = 0, \ldots, n$. Here $\mathscr{P}_{[0,n]}(\kappa) \subset \mathscr{P}_{[0,n]}$ is a subset of channel input distributions, which satisfy an average transmission cost constraint with total power κ. The optimal channel input conditional distributions are characterized by conditional independence properties, called "information structures". The identification of information structures simplify the resulting finite-time horizon optimization problem called the "characterization of finite transmission feedback information (FTFI) capacity". These are analogous to those of memoryless channels without feedback, which are established via the well-known upper bounds

$$C_{A^n;B^n}^{noFB} \stackrel{\triangle}{=} \max_{\mathbf{P}_{A^n}} I(A^n; B^n) \leq \max_{\mathbf{P}_{A_i}:i=0,\ldots,n} \sum_{i=0}^{n} I(A_i; B_i)$$
$$\leq (n+1) \max_{\mathbf{P}_A} I(A; B) \equiv (n+1)C \tag{I.2}$$

which are achieved if $\mathbf{P}_{A_i|A^{i-1}} = \mathbf{P}_{A_i}, i = 0, \ldots, n$ and identically distributed, that imply the joint process $\{(A_i, B_i) : i = 0, 1, \ldots, \}$ is independent and identically distributed and ergodic (and similarly if noiseless feedback is allowed). For memoryless channels the bounds in (I.2) are applied in the converse part of the coding theorem, to obtain a tight bound on any achievable rate, while the direct part of the coding theorem is often shown by randomly generating codes independently, according to the product distribution $\mathbf{P}_{A^n}^*(da^n) \stackrel{\triangle}{=} \times_{i=0}^{n} \mathbf{P}_A^*(da_i)$, where $\mathbf{P}_A^*$ is the maximizing distribution that achieves C.

However, to make the transition from memoryless channels to channels with memory, without any á priori assumptions, such as, stationarity, ergodicity or information stability, it is often necessary to investigate the characterizations of FTFI

Manuscript received March 6, 2016; revised January 13, 2018; accepted March 11, 2018. Date of publication May 9, 2018; date of current version August 16, 2018.

C. D. Charalambous and C. K. Kourtellaris are with the Department of Electrical and Computer Engineering, University of Cyprus, 1678 Nicosia, Cyprus (e-mail: chadcha@ucy.ac.cy; kourtellaris.christos@ucy.ac.cy).

S. Loyka is with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: sergey.loyka@ieee.org).

Communicated by H. H. Permuter, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2834516

capacity $C_{A^n \to B^n}(\kappa)$, and its asymptotic properties via the per unit time limit

$$C_{A^\infty \to B^\infty}(\kappa) \stackrel{\triangle}{=} \liminf_{n \longrightarrow \infty} \frac{1}{n+1} C_{A^n \to B^n}(\kappa). \tag{I.3}$$

This follows from the fact that by the converse coding theorem, the quantity $C_{A^\infty \to B^\infty}(\kappa)$ is a tight upper bound on any achievable rate of feedback codes (see Theorem 24, (a)).

This paper builds on the following ideas put forward in [1].

1) The information structures of optimal channel input distributions and corresponding characterizations of FTFI capacity, translate into corresponding information structures for $C_{A^\infty \to B^\infty}(\kappa)$. Moreover, via the converse coding theorem, tight bounds on any achievable code rate (of feedback codes) can be obtained, while the direct part of the coding theorem can be shown, without unnecessary á priori assumptions on the channel, such as, stationarity, ergodicity, or information stability of the joint process $\{(A_i, B_i) : i = 0, 1, \ldots\}$.
2) The characterizations of the FTFI capacity reveal several hidden properties of the role of feedback to control the channel output process of nonstationary, nonergodic and unstable channels. These include fundamental properties of optimal channel input conditional distributions, which achieve $C_{A^\infty \to B^\infty}(\kappa)$, properties of channel parameters so that $C_{A^\infty \to B^\infty}(\kappa)$ corresponds to feedback capacity, and whether feedback increases capacity.

A. Contributions

The main contributions of the paper are the following.

i) Develop a methodology to realize optimal channel input conditional distributions, by information lossless randomized strategies (deterministic functions) driven by uniformly distributed Random Variables (RVs). Then apply the information lossless randomized strategies to derive alternative equivalent characterizations of FTFI capacity, using randomized strategies driven by arbitrary independent RVs. In specific application examples, such as, MIMO Gaussian channels with memory, the independent RVs transform the optimal randomized strategies into capacity achieving strategies, driven by independent Gaussian RVs.
ii) Show that optimal randomized strategies of MIMO Gaussian channels, which achieve the FTFI information capacity, decompose into two orthogonal parts. The control part which controls the channel output process, and the information transmission or innovations part, which is responsible to transmit new information. Further, show a separation principle between the computation of the optimal control part and the optimal innovations part.
iii) Identify sufficient conditions, in terms of channel parameters and transmission cost functions, so that the per unit time limit of the FTFI capacity $C_{A^n \to B^n}(\kappa)$ converges to the characterization of feedback capacity $C_{A^\infty \to B^\infty}(\kappa)$.
iv) Identify sufficient conditions, to determine whether feedback increases capacity, of MIMO Gaussian channels with memory.

The analyzed application examples of nonstationary, nonergodic, unstable, multiple input multiple output (MIMO) Gaussian channels with memory, unfold the interaction of the control and the information transmission components (strategies) of the optimal channel input distributions.

All results are developed by investigating the FTFI capacity $C_{A^n \to B^n}(\kappa)$. This is analogous to the Cover and Pombra [2] $n-$block length feedback capacity of time-varying additive Gaussian noise (AGN) channels. The separation principle is shown by directly attacking the FTFI capacity $C_{A^n \to B^n}(\kappa)$. The analysis reveals fundamental insights on the interaction of control and information transmission. Further, it is demonstrated that it is much simpler to attack $C_{A^\infty \to B^\infty}(\kappa)$, from the per unit time limit of $C_{A^n \to B^n}(\kappa)$, because sufficient conditions can be identified for $C_{A^\infty \to B^\infty}(\kappa)$ to correspond to the characterization of feedback capacity, irrespectively of whether the channel is stable or unstable. This is contrary to the recent believe in [3] (see page 57, second column, first paragraph), where it is claim that attacking directly $C_{A^n \to B^n}(\kappa)$, for each "n" is almost impossible.

The channel conditional distributions considered in this paper, include the following classes.

Channel Distributions Class A.

$$\mathbf{P}_{B_i|B^{i-1},A^i}(db_i|b^{i-1}, a^i) = \mathbf{P}_{B_i|B^{i-1},A_i}(db_i|b^{i-1}, a_i). \tag{I.4}$$

Channel Distributions Class B.

$$\mathbf{P}_{B_i|B^{i-1},A^i}(db_i|b^{i-1}, a^i) = \mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(db_i|b_{i-M}^{i-1}, a_i), \quad i = 0, \ldots, n \tag{I.5}$$

where M is a nonnegative integer. The convention for $M = 0$, is $\mathbf{P}_{B_i|B_{i-M}^{i-1},A_i}(db_i|b_{i-M}^{i-1}, a_i) = \mathbf{P}_{B_i|A_i}(db_i|a_i)$, $i = 0, 1, \ldots, n$, that is, the channel degenerates to a memoryless channel.

The average transmission cost constraint is of the form

$$\mathscr{P}_{[0,n]}(\kappa) \stackrel{\triangle}{=} \Big\{ \mathbf{P}_{A_i|A^{i-1},B^{i-1}}, i = 0, \ldots, n : \frac{1}{n+1} \mathbf{E}\Big(\sum_{i=0}^{n} \gamma_i(A_i, T^i B^n) \Big) \leq \kappa \Big\}, \quad \kappa \in [0, \infty) \tag{I.6}$$

where for each i, $T^i b^n$ is given by either $T^i b^n = b^i$ or $T^i b^n = b_{i-K}^i$, K a nonnegative finite integer, for $i = 0, \ldots, n$. Thus, the transmission cost functions are either one of the following two classes[1]

Transmission Cost Functions Class A.

$$\gamma_i(a_i, T^i b^n) = \gamma_i^{\,A}(a_i, b^i). \tag{I.7}$$

Transmission Cost Functions Class B.

$$\gamma_i(a_i, T^i b^n) = \gamma_i^{\,B}(a_i, b_{i-K}^i), \quad i = 0, \ldots, n. \tag{I.8}$$

In this paper, the dual role of optimal channel input distributions/randomized strategies, and separation principle of

[1]There is no loss of generality to consider $\gamma_i^{\,B}(a_i, b_{i-K}^i)$, because by the function restriction, such functions include $\gamma_i^{\,B}(a_i, T^i b^n) = \gamma_i(a_i, b_{i-K}^{i-L})$ and $\gamma_i(a_i)$, for any nonnegative integers $K \geq L$, and similarly for $\gamma_i^{\,A}(a_i, b^i)$.

computing the parts of the optimal strategy, are only illustrated for multiple input multiple output (MIMO) Gaussian channels with memory, via a provocative direct connection to the linear-quadratic-Gaussian (LQG) stochastic optimal control theory, stability of linear stochastic control systems, and Lyapunov and Riccati matrix equations [4], [5]. Indeed, the LQG stochastic optimal control theory generalizes, in a natural way, to directed information pay-off functionals. For the readers convenience a short summary of these concepts is given in Appendix E, while several examples are discussed to illustrate their applications. These tools are necessary to treat processes $\{(A_i, B_i) : i = 0, 1, \ldots, \}$, which are not assumed á priori to be stationary, ergodic or information stable. Rather, via these mathematical concepts, sufficient conditions are identified to show that the optimal channel input conditional distribution induces asymptotic stationarity and ergodicity of the joint process $\{(A_i, B_i) : i = 0, 1, \ldots\}$, and to show that the per unit time limit of the FTFI capacity exists and characterizes feedback capacity, irrespectively of whether the channel is stable or unstable.

For the application examples, it is further shown that the optimal channel input distributions, which achieve FTFI capacity are realized by randomized strategies, and that such strategies decompose into two orthogonal parts, the deterministic part and the random part. Through this decomposition, a separation principle is shown, between the role of randomized strategies to control the channel output process and to transmit new information over the channel. It is also shown that the deterministic part corresponds to the optimal solution of the LQG stochastic optimal control problem, while the random part is determined from water-filling type equations.

In Section I-C, a short summary of the main concepts, methods, and results obtained in this paper are presented.

B. Literature Review

Over the years many papers have been written on the characterization of capacity of channels with memory and feedback, and on the computation of capacity. This section reviews only part of the literature, with an emphasis on problems related to this paper, and channels defined on continuous alphabet spaces. Cover and Pombra [2] investigated the scalar time-varying additive Gaussian noise (AGN) channel with memory, defined by

$$B_i = A_i + V_i, \quad \frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}|A_i|^2\Big\} \leq \kappa, \quad \kappa \in [0, \infty) \tag{I.9}$$

where $V^n \stackrel{\triangle}{=} \{V_0, V_1, \ldots, V_n\}$ is a real-valued (scalar) jointly nonstationary and nonergodic Gaussian process, with covariance K_{V^n}. The underlying assumption (see [2, p. 39, Lemma 5]) is that "$A^n \stackrel{\triangle}{=} \{A_0, \ldots, A_n\}$ is causally related to V^n", which states $\mathbf{P}_{A^n,V^n}(da^n, dv^n) = \otimes_{i=0}^{n}\mathbf{P}_{A_i|A^{i-1},V^{i-1}}(da_i|a^{i-1}, v^{i-1}) \otimes \mathbf{P}_{V^n}(dv^n)$. Cover and Pombra [2] characterized the feedback capacity $C_{W;B^\infty}^{CP}(\kappa) \stackrel{\triangle}{=} \lim_{n\longrightarrow\infty}\frac{1}{n+1}C_{W;B^n}^{CP}(\kappa)$, via the following characterization of FTFI capacity.[2]

$$C_{W;B^n}^{CP}(\kappa) = \max_{(\overline{\Gamma}_n, K_{\overline{Z}^n}):\frac{1}{n+1}\mathbf{E}\left\{tr\left(A^n(A^n)^T\right)\right\}\leq\kappa} H(B^n) - H(V^n),$$

$$A_i = \sum_{j=0}^{i-1}\overline{\gamma}_{i,j}V_j + \overline{Z}_i, \quad i = 0, \ldots, n \tag{I.10}$$

$$= \max_{(\overline{\Gamma}_n, K_{\overline{Z}^n}):\ tr\left(\overline{\Gamma}_n K_{V^n}\,\overline{\Gamma}_n^T + K_{\overline{Z}^n}\right)\leq\kappa(n+1)} \left\{ \frac{1}{2}\log\frac{|\left(\overline{\Gamma}_n + I\right)K_{V^n}\left(\overline{\Gamma}_n + I\right)^T + K_{\overline{Z}^n}|}{|K_{V^n}|}\right\} \tag{I.11}$$

where $\overline{Z}^n \stackrel{\triangle}{=} \{\overline{Z}_i : i = 0, 1, \ldots, n\}$ is a correlated zero mean, covariance $K_{\overline{Z}^n}$ Gaussian process, denoted by $N(0, K_{\overline{Z}^n})$, that is orthogonal to $V^n \stackrel{\triangle}{=} \{V_i : i = 0, \ldots, n\}$, $\overline{\Gamma}_n$ is lower diagonal time-varying matrix with deterministic entries, and I is the identity matrix. Although, Cover and Pombra [2] call $\{\overline{Z}_i : i = 0, \ldots, n\}$ an innovation like process, this is not equivalent to the standard definition of an *innovation process*, which is an orthogonal process. Let $C_{A^\infty;B^\infty}^{noFB}(\kappa)$ denote the capacity without feedback. In [2] it is also shown that[3]

$$C_{A^\infty;B^\infty}^{noFB}(\kappa) \leq C_{W;B^\infty}^{CP}(\kappa) \leq 2C_{A^\infty;B^\infty}^{noFB}(\kappa)$$
$$C_{W;B^\infty}^{CP}(\kappa) \leq C_{A^\infty;B^\infty}^{noFB}(\kappa) + \frac{1}{2}. \tag{I.12}$$

The first inequalities are also obtained by Ebert [6]. The Cover and Pombra scalar time-varying AGN channel is extensively analyzed by Ihara in [7, Sec. 5.7, pp. 210–219]. Ihara in [7, Th. 5.7.3] showed that $C_{W;B^n}^{CP}(\kappa)$ is also achieved by transmitting a Gaussian message using a *linear coding scheme*. Moreover, Ihara in [7, Example 5.7.1, p. 217–218], considered a scalar, stable, first-order autoregressive AR(1) Gaussian noise represented by the recursion

$$V_i = \alpha V_{i-1} + W_i, \quad V_{-1} = 0, \quad \alpha \in (0, 1), \quad i = 0, 1, \ldots, n \tag{I.13}$$

where $\{W_i : i = 0, \ldots, n\}$ is independent and identically distributed Gaussian $W_i \sim N(0; 1)$, and applied a linear coding scheme of transmitting a Gaussian RV to derive the lower and upper bounds on feedback capacity, given by the following equations.

$$\frac{1}{2}\log x^2 = \frac{1}{2}\log\left\{1 + \left(1 + \frac{\alpha}{x}\right)^2\kappa\right\} \leq C_{W;B^\infty}^{CP}(\kappa)$$
$$\leq \frac{1}{2}\log\left\{1 + \left(1 + \alpha\right)^2\kappa\right\} \tag{I.14}$$

where x is the unique positive solution of the equation

$$x^4 - x^2 - \kappa\left(x + \alpha\right)^2 = 0. \tag{I.15}$$

[2] In [2], characterization (I.10) is obtained via the converse to the coding theorem, by showing that $C_{W;B^\infty}^{CP}(\kappa)$ is an achievable upper bound on the mutual information between uniformly distributed source messages W and channel outputs B^n, i.e., on $I(W; B^n)$. The per unit time $\frac{1}{n+1}C_{W;B^n}^{CP}(\kappa)$ is often called the $n-$block length feedback capacity.

[3] and similarly for the finite transmission versions, i.e., without taking the limit.

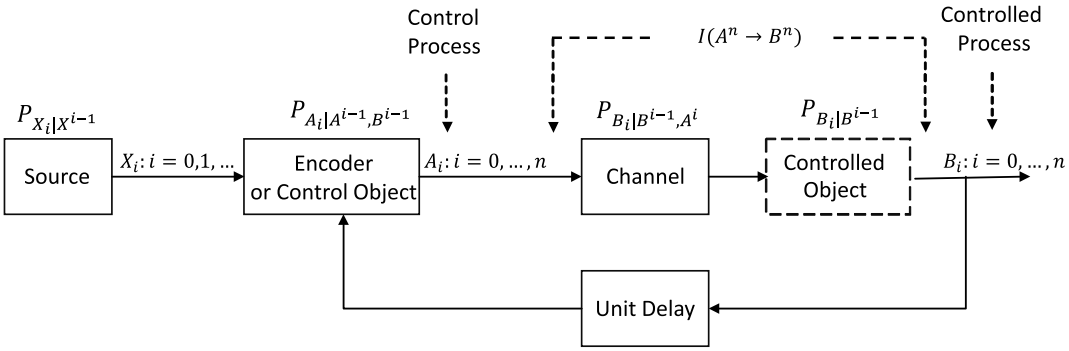

Fig. I.1. Communication block diagram and its analogy to stochastic optimal control.

The above bounds are also derived by Butman in [8, Abstract] for stationary autoregressive models with finite memory. Yang *et al.* [9, Sec. II] analyzed the stationary limited memory noise version of the Cover and Pombra [2] AGN, when the channel noise is described by a power spectral density (PSD), $S_V(\omega) = |H(e^{j\omega})|^2$, where the filter $H(z), z = e^{j\omega}, \omega \in [-\pi, \pi]$ is a proper rational polynomial in z, with stable poles and marginally stable zeros. Yang *et al.* [9] applied the inverse filter transformation $H^{-1}(z)$ to the channel output process B^n [9, Section II.C], to obtain an equivalent state space Gaussian noise channel. The main assumption in [9, Sec. II.C] is that the state of the noise uniquely defines the channel inputs and vice-versa. Yang *et al.* [9, Th. 7] state that for an autoregressive noise filter with one zero and one pole, then feedback capacity is achieved by a channel input process which is a linear function of the estimation error of the noise state, from past channel outputs, i.e., when the innovation part of the channel input process is either zero or asymptotically zero. When specialized to an AR(1) model (I.13), the feedback capacity given in [9, Corollary 7.1] is precisely the lower bound obtained in [7, Example 5.7.1] and earlier in [8], using a linear coding scheme of encoding a Gaussian message, and given by (I.14), where x is the unique positive solution of (I.15) (simply set $\Sigma_W^2 = 1$ in [9, Corollary 7.1]). Contrary to the linear coding scheme applied in [7] and [8], the input process given in [9, Th. 7, Corollary 7.1] does not include any randomization or it is asymptotically zero; however, Yang, Kavcic and Tatikonda concluded in (see [9, Sec. VII, Conclusion, last two paragraphs]) that it is still an open problem to determine whether both the scaling on the error of the noise state and the innovations part of the channel input process could take nonzero optimal values. Kim [3] analyzed the stationary limited memory noise version of the Cover and Pombra AGN channel, i.e., the noise PSD is described by Gaussian autoregressive model of order K, using mostly frequency domain techniques. Among other results, in Kim [3, Th. 6 and Lemma 6.1], it is stated that an input process which is a linear function of the estimation error of the state of the noise, from past channel outputs, with zero innovation process, as in [9, Th. 7 and Corollary 7.1], achieves feedback capacity, and further this expression of feedback capacity is also achieved by $K-$dimensional generalization of the Schalkwijk and Kailath [10] coding scheme. For the AR(1) given by (I.13), the feedback capacity given in [3, p. 58, last two equations], is precisely the lower bound given by (I.14). Kim [3] and Yang [9] do not discuss the connection to the lower and upper bounds given by (I.14). A more recent analysis of general Gaussian channels with past dependence on channel inputs and channel outputs are discussed in [11].

Recently, feedback capacity problems for certain types of channels with inputs and outputs taking values in finite alphabet spaces, without transmission cost constraints, are investigated in [12]–[16], and in [17]–[19] when transmission cost are imposed. Coding theorems for memoryless channels and channels with memory, with and without feedback, are developed extensively over the years, under various assumptions, for example, in [7] and [20]–[30] (often for stationary ergodic processes, Gaussian channels, and channel with inputs-outputs, which are defined on finite alphabet spaces).

The contributions of this paper listed in Section I-A, i)-iv) complement previous results obtained in [2], [3], [7], and [9], with respect to the methodology and the optimization procedure that is used to derive the main results of this paper.

C. Discussion of Methodology and Main Results

The results listed in Section I-A, i)-iv) are derived by applying the analogy to stochastic optimal control theory depicted in Fig. I-B, where the information measure $I(A^n \to B^n)$ is the pay-off, the channel output process $\{B_i : i = 0, 1, \ldots, n\}$ is the controlled process, the channel input process $\{A_i : i = 0, 1, \ldots, n\}$ is the control process, and $B^{-1} = b^{-1}$ is the initial state with fixed distribution, that is available to the encoder and may or may not be available to the decoder.

1) Randomized Strategies and Characterizations of FTFI Capacity: Section III, describes a methodology to derive alternative equivalent characterizations of FTFI capacity, by realizing optimal channel input conditional distributions, which maximize directed information, using information lossless randomized strategies driven by independent uniformly distributed RVs, and then by arbitrary distributed RVs.

An alternative equivalent characterization of FTFI capacity is illustrated below.

Equivalent characterizations of FTFI capacity for class B channels and transmission cost functions: Consider $\mathbf{P}_{A_i|B_{i-M}^{i-1}}$, $\gamma_i(a_i, T^i b^n) = \gamma_i^B(a_i, b_{i-M}^{i-1}), i = 0, \ldots, n$, and M a

nonnegative finite integer. In [1], it is shown that the maximization of $I(A^n \to B^n)$ over all distributions $\mathbf{P}_{A_i|A^{i-1},B^{i-1}} : i = 0, \ldots, n\}$, which satisfy the constraint $\frac{1}{n+1}\mathbf{E}\{\sum_{i=0}^{n} \gamma_i^B(A_i, B_{i-M}^{i-1})\} \leq \kappa$, satisfy conditional independence

$$\mathbf{P}_{A_i|A^{i-1},B^{i-1}} = \mathbf{P}_{A_i|B_{i-M}^{i-1}}, \quad i = 0, \ldots, n. \tag{I.16}$$

This implies the information structure of the optimal channel input distribution is B_{i-M}^{i-1} for $i = 0, 1, \ldots, n$, and the characterization of the FTFI capacity is given by the following expression.

$$C_{A^n \to B^n}^{B.M}(\kappa) = \sup_{\mathbf{P}_{A_i|B_{i-M}^{i-1}}, i=0,\ldots,n: \frac{1}{n+1}\mathbf{E}\{\sum_{i=0}^{n} \gamma_i^B(A_i, B_{i-M}^{i-1})\} \leq \kappa} \Big\{ \mathbf{E}\Big\{ \sum_{i=0}^{n} \log\Big(\frac{d\mathbf{P}_{B_i|B_{i-M}^{i-1}, A_i}(\cdot|B_{i-M}^{i-1}, A_i)}{d\mathbf{P}_{B_i|B_{i-M}^{i-1}}(\cdot|B_{i-M}^{i-1})}(B_i)\Big)\Big\}\Big\}. \tag{I.17}$$

In Theorem 8, by utilizing (I.17), the following are shown.

(i) The class of optimal channel input distributions are realized by information lossless randomized strategies defined by

$$\mathscr{E}_{[0,n]}^{B.M}(\kappa) \stackrel{\triangle}{=} \Big\{ e_i : \mathbb{B}_{i-M}^{i-1} \times \mathbb{Z}_i \longmapsto \mathbb{A}_i,\ a_i = e_i(b_{i-M}^{i-1}, z_i) : \frac{1}{n+1}\mathbf{E}^e\Big(\sum_{i=0}^{n} \gamma_i^B(e_i(B_{i-M}^{i-1}, Z_i), B_{i-M}^{i-1})\Big) \leq \kappa \Big\} \tag{I.18}$$

where $\{Z_i : i = 0, \ldots, n\}$ is an independent sequence of RVs, and Z_i is independent of B^{i-1}, for $i = 0, \ldots, n$.

(ii) An alternative equivalent characterization of the FTFI capacity (I.17), is given by[4]

$$C_{A^n \to B^n}^{B.M}(\kappa) = \sup_{\{Z_i:i=0,\ldots,n\}, \{e_i(\cdot,\cdot):i=0,\ldots,n\} \in \mathscr{E}_{0,n}^{B.M}(\kappa)} \Big\{ \mathbf{E}^e\Big\{ \sum_{i=0}^{n} \log\Big(\frac{d\mathbf{P}(\cdot|B_{i-M}^{i-1}, e_i(B_{i-M}^{i-1}, Z_i))}{d\mathbf{P}^e(\cdot|B_{i-M}^{i-1})}(B_i)\Big)\Big\}\Big\}. \tag{I.19}$$

In application examples, the maximizing randomized strategy, $\{e_i^*(\cdot,\cdot) : i = 0, \ldots, n\} \in \mathscr{E}_{[0,n]}^{B.M}(\kappa)$, and the maximizing process $\{Z_i^* : i = 0, \ldots, n\}$ (i.e., its distribution) induce the maximizing channel input conditional distribution of the characterization of FTFI capacity (I.17).

Further, the following are illustrated, via application examples of MIMO Gaussian channel models (G-CMs).

2) Dual Role of Randomized Strategies & LQG Theory: For channels with memory on past channel outputs, then randomized strategies, which realize candidates of optimal channel input distributions, have a *Dual Role*, specifically, to optimally control the channel output process $\{B_i : i = 0, 1, \ldots, n\}$, and to communicate information. In Theorem 14 (Section IV-C), the dual role of randomized strategies (and several properties), are illustrated for MIMO time-varying G-CMs with memory. The following application example illustrates this dual role.

Alternative characterization of FTFI capacity for G-CM-B.1: Consider the MIMO G-CM-B.1, corresponding to channel Class B and transmission cost Class B, defined by[5]

$$B_i = C_{i,i-1}\ B_{i-1} + D_{i,i}\ A_i + V_i,\ B_{-1} = b_{-1},\ i = 0, \ldots, n, \tag{I.20}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{ \sum_{i=0}^{n} \Big(\langle A_i, R_i A_i \rangle + \langle B_{i-1}, Q_{i,i-1} B_{i-1} \rangle \Big)\Big\} \leq \kappa, \tag{I.21}$$

$$\begin{aligned} &\mathbf{P}_{V_i|V^{i-1},A^i,B^{-1}}(dv_i|v^{i-1}, a^i, b^{-1}) = \mathbf{P}_{V_i}(dv_i), \\ &V_i \sim N(0, K_{V_i}), \quad i = 0, \ldots, n, \end{aligned} \tag{I.22}$$

$$\begin{aligned} &(C_{i,i-1}, D_{i,i}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times q}, \\ &(R_i, Q_{i,i-1}) \in \mathbb{R}^{q\times q} \times \mathbb{R}^{p\times p}, \\ &R_i = R_i^T \succ 0,\ Q_{i,i-1} = Q_{i,i-1}^T \succeq 0,\ i = 0, \ldots, n \end{aligned} \tag{I.23}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product of elements of linear spaces. In Section IV-B, the following are shown (by using (I.17), with $M = 1$).

(iii) The optimal conditional channel input distributions are conditionally Gaussian of the form $\{\mathbf{P}_{A_i|B_{i-1}}^g(da_i|b_{i-1}) : i = 0, \ldots, n\}$, which satisfy the average transmission cost constraint, and that such distributions are realized by linear randomized strategies $e(\cdot) \in \mathscr{E}_{0,n}^{B.1}(\kappa)$ driven by a Gaussian innovations process $\{Z_i : i = 0, \ldots, n\}$, defined by the set

$$\begin{aligned} &\mathscr{E}_{[0,n]}^{B.1}(\kappa) \stackrel{\triangle}{=} \Big\{ A_i^g = g_i^{B.1}(B_{i-1}^g) + Z_i = \Gamma_{i,i-1} B_{i-1}^g + Z_i, \\ &Z_i \perp B^{g,i-1},\ \{Z_i : i = 0, \ldots, n\} \text{ independent process}, \\ &Z_i \sim N(0, K_{Z_i}),\ K_{Z_i} \succeq 0, \quad i = 0, \ldots, n : \\ &\frac{1}{n+1}\mathbf{E}^{g^{B.1}}\Big\{ \sum_{i=0}^{n} \Big[\langle A_i^g, R_{i,i} A_i^g \rangle + \langle B_{i-1}^g, Q_{i,i-1} B_{i-1}^g \rangle \Big]\Big\} \leq \kappa \Big\} \end{aligned} \tag{I.24}$$

where $\cdot \perp \cdot$ means the processes are independent. Thus, randomized strategies in $\mathscr{E}_{[0,n]}^{B.1}(\kappa)$ are decomposed into two orthogonal parts, one of which is an innovations process (i.e., independent process).

(iv) The characterization of FTFI capacity of the MIMO G-CM.B.1 is given by the following expression.

$$C_{A^n \to B^n}^{B.1}(\kappa) = \sup_{\{(\Gamma_{i,i-1}, Z_i), i=0,\ldots,n\} \in \mathscr{E}_{[0,n]}^{B.1}(\kappa)} \Big\{ \frac{1}{2}\sum_{i=0}^{n} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} \Big\}. \tag{I.25}$$

The decomposition

$$A_i^g = \Gamma_{i,i-1} B_{i-1}^g + Z_i \equiv g_i^{B.1}(B_{i-1}^g) + Z_i, \quad i = 0, \ldots, n \tag{I.26}$$

implies that the feedback function $\{g_i^{B.1} \equiv \Gamma_{i,i-1} : i = 0, \ldots, n\}$ is the feedback control law or strategy, which

[4]The subscript notation on conditional distributions is suppressed, while superscript notation indicates dependence on the strategies.

[5]The fundamental difference between $Q_{i,i-1} \neq 0$ versus $Q_{i,i-1} = 0, i = 0, \ldots, n$ and its implications on the maximum rate of transmitting information over this channel, is discussed shortly.

controls the output process $\{B_i^g : i = 0, \ldots, n\}$, while the orthogonal innovations process $\{Z_i : i = 0, \ldots, n\}$ is responsible to convey new information to the output process, both chosen to maximize (I.25).

It should be noted that $K_{Z_i} = 0, i = 0, \ldots, n$ implies that $C_{A^n \to B^n}^{B.1}(\kappa) = 0$, as expected (because the randomization in (I.26) is zero), and hence the resulting optimization reduces to the minimization of the average pay-off in (I.24), over the deterministic strategies $\{g_i^{B.1}(b_{i-1}^g) : i = 0, \ldots, n\}$, which is equivalent to a LQG stochastic optimal control problem. It is also noted that the solution of this LQG problem defines the minimum cost of control, say, $\kappa_{min} \in [0, \infty)$, and that a solution to (I.25) exists for $\kappa \in [\kappa_{min}, \infty)$.

3) Separation Principle: In Theorem 14, the decomposition (I.26) is applied to show a separation principle between the computation of the optimal control part and the innovations part, and to solve the extremum problem (I.25), via its relation to the linear-quadratic-Gaussian (LQG) stochastic optimal control theory (with randomized controls). The following are shown.

(v) The characterization of FTFI capacity is given by

$$C_{A^n \to B^n}^{B.1}(\kappa) = \inf_{s \geq 0} \Big\{ -s \int_{\mathbb{R}^p} \langle b_{-1}, P(0) b_{-1} \rangle \mathbf{P}_{B_{-1}}(db_{-1}) + r(0) \Big\} \quad \text{(I.27)}$$

where $\{P(i) : i = 0, \ldots, n\}$ is a positive semi definite solution of the matrix Riccati difference equation

$$P(i) = C_{i,i-1}^T P(i+1) C_{i,i-1} + Q_{i,i-1} - C_{i,i-1}^T P(i+1) D_{i,i} \times \Big(D_{i,i}^T P(i+1) D_{i,i} + R_{i,i} \Big)^{-1} \Big(C_{i,i-1}^T P(i+1) D_{i,i} \Big)^T, \quad i = 0, \ldots, n-1, \quad P(n) = Q_{n,n-1} \quad \text{(I.28)}$$

$s \geq 0$ is the Lagrange multiplier associated with the transmission cost constraint, and *the optimal deterministic part of the randomized strategy*, $\{g_i^{B.1,*}(\cdot) : i = 0, \ldots, n\}$, is given by

$$g_i^{B.1,*}(b_{i-1}) = -\Big(D^T P(i+1) D + R \Big)^{-1} D^T P(i+1) C \, b_{i-1} \equiv \Gamma_{i,i-1}^* b_{i-1}, \quad i = 0, \ldots, n-1, \quad \text{(I.29)}$$

$$g_n^{B.1,*}(b_{n-1}) = 0. \quad \text{(I.30)}$$

The optimal random part of the strategy $\{K_{Z_i}^* : i = 0, \ldots, n\}$ (covariance of innovations process) is found from the water-filling problem

$$r(i) = r(i+1) + \sup_{K_{Z_i} \in \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(s P(i+1) \Big[D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i} \Big] \Big) - tr\Big(s \, R_{i,i} K_{Z_i} \Big) \Big\}, \quad i = 0, \ldots, n-1, \quad \text{(I.31)}$$

$$r(n) = \sup_{K_{Z_n} \in \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{n,n} K_{Z_n} D_{n,n}^T + K_{V_n}|}{|K_{V_n}|} + s(n+1)\kappa - tr\Big(s R_{n,n} K_{Z_n} \Big) \Big\}. \quad \text{(I.32)}$$

An alternative equivalent water-filling problem is given by (IV.171) and Remark 15, while the optimal $\{K_{Z_i}^* : i = 0, \ldots, n\}$ for the scalar case is given in Remark 16. The above solution illustrates the separation principle, between the computation of the deterministic part $\{g_i^{B.1}(B_{i-1}) : i = 0, \ldots, n\}$ and random part $\{Z_i \sim K_{Z_i} : i = 0, \ldots, n\}$ of the randomized strategy, in that, the latter can be found, by first computing the former. Moreover, the properties of solutions $\{P(i) : i = 0, \ldots, n\}$ to the Riccati equation, such as, $P(i) \succ 0$ or $P(i) \succeq 0, i = 0, \ldots, n$ (positive definite or positive semi definite), depend on the properties of the parameters of the channel and the transmission cost function, $\{C_{i,i-1}, D_{i,i}, R_{i,i}, Q_{i,i-1} : i = 0, \ldots, n\}$. It should be mentioned that if $P(i) = 0, i = 0, \ldots, n-1$ then $A_i^g = Z_i, i = 0, \ldots, n$ and this means feedback does not incur a higher value for $C_{A^n \to B^n}^{B.1}(\kappa)$.

(vi) The optimal strategy (I.29) is precisely the solution of the following LQG stochastic optimal control problem [4]. This connection is made explicit in Remark 15.

(vii) If the channel is time-invariant with $\{C_{i,i-1} = C, D_{i,i} = D, K_{V_i} = K_V, R_{i,i} = R, i = 0, \ldots, n, Q_{i,i-1} = Q, i = 0, \ldots, n-1, Q_{n,n-1} = M\}$, from (I.27), then whether $C_{A^\infty \to B^\infty}^{B.1}(\kappa) \stackrel{\triangle}{=} \lim_{n \longrightarrow \infty} \frac{1}{n+1} C_{A^n \to B^n}^{B.1}(\kappa)$ exists and corresponds to feedback capacity that is independent of the initial distribution $\mathbf{P}_{B_{-1}}$, is determined from the properties of solutions to the following matrix Riccati algebraic equation.

$$P = C^T P C + Q - C^T P D \Big(D^T P D + R \Big)^{-1} \Big(C^T P D \Big)^T. \quad \text{(I.33)}$$

The conditions are given in Theorem 19, in terms of the so-called detectability and stabilizability. Moreover, whether feedback increases capacity is determined from the solutions of matrix Riccati equation (I.33).

4) Application Example: Feedback versus No Feedback & The Infinite Horizon LQG Theory: In Sections V, the per unit time limit of the characterizations of FTFI capacity of time-invariant G-CM-Bs are investigated under the detectability and stabilizability conditions. It is shown that whether feedback increases capacity, is determined from the unique (stabilizing) solution of the matrix Riccati algebraic equation (I.33). This is established via direct connections to the infinite-horizon LQG stochastic optimal control theory and stability of linear stochastic controlled systems, and associated Lyapunov equations and matrix Riccati equations. Indeed, even if the channel defined by (I.20) is unstable (i.e., any of the eigenvalues of matrix C are outside the unit circle of the complex plane), under certain conditions, which are specified in terms of (C, D, R, Q, K_V), then the optimal deterministic part of the randomized strategy stabilizes the channel via feedback, and ensures asymptotic stationarity and existence of a unique invariant joint distribution of the joint process $\{(A_i, B_i) : i = 0, \ldots, n\}$, marginal distribution of the channel output process, and ensures that $C_{A^\infty \to B^\infty}(\kappa)$ exists, it is finite, and corresponds to feedback capacity.

The following simple example illustrates several hidden properties of optimal channel input distributions, and that feedback capacity and capacity without feedback are determined

from the properties of the solutions to the matrix Riccati algebraic equation (I.33).

Special case-the time-invariant scalar channel with $p = q = 1, R = 1, Q = 0$ and (C, D) arbitrary: For these choices of parameters the following are shown (and independently in Example 13). The steady state solutions of Riccati (quadratic) equation (I.33), and corresponding optimal determinist part of the randomized strategy are given by the following equations.

$$P\Big(D^2\ P + \Big[1 - C^2\Big]\Big) = 0 \Longrightarrow P_1 = 0, \quad P_2 = \frac{C^2 - 1}{D^2}, \tag{I.34}$$

$$g^{B.1,*}(b) = \Gamma^*\ b,$$

$$\Gamma^* \equiv \Gamma^*(P) = -\Big(D^2\ P + 1\Big)^{-1}$$

$$DPC = \begin{cases} 0 & \text{if } P = P_1 \\ -\frac{C^2-1}{CD} & \text{if } P = P_2. \end{cases} \tag{I.35}$$

where $P = P_1$ implies the admissible optimal channel input distribution does not use feedback. The optimal covariance K_Z^*, is obtained from the per unit time limit of (I.31), (I.32) given by (V.224), while the Lagrange multiplier, s, is found from the average constraint or by performing the infimum over $s \geq 0$ of $J^{B.1,*}$ evaluated at (P, K_Z^*) given by (V.224). The calculations give the following expressions.

If $|C| < 1$ then

$$\Gamma^* = 0, \quad K_Z^* = \kappa, \quad \kappa \in [0, \infty).$$

If $|C| > 1$ then

$$\Gamma^* = -\frac{C^2 - 1}{CD}, \quad K_Z^* = \frac{D^2\kappa + K_V(1 - C^2)}{C^2D^2} \geq 0, \quad \kappa \in [\kappa_{min}, \infty),$$

$$s^* = \frac{1}{2}\frac{D^2}{D^2\kappa + K_V} \in [s^*_{min}, \infty), \quad \kappa_{min} \stackrel{\triangle}{=} \frac{(C^2 - 1)K_V}{D^2},$$

$$s^*_{min} \stackrel{\triangle}{=} \frac{1}{2}\frac{D^2}{C^2\ K_V}.$$

Hence, if $|C| < 1$ then $P_1 = 0$ is the unique positive semidefinite solution of (I.33). If $|C| > 1$ then uniqueness of positive semidefinite solution of (I.33), which ensures asymptotic ergodicity of $B_i, i = 0, 1, \ldots,$ fails.

It is noted that $|C| = 1$ in excluded, because it implies $\Gamma^* = 0$ and the eigenvalue of the channel is on the unit circle. Asymptotic stationarity and ergodicity of $B_i, i = 0, 1, \ldots,$ i.e., uniqueness of invariant distribution is ensured if $Q > 0$, i.e., the detectability condition (V.206) also holds.

(vii) *The Feedback Capacity*. The optimal strategy which achieves feedback capacity $C_{A^\infty \to B^\infty}(\kappa)$ is given by

$$\Big(\Gamma^*, K_Z^*\Big) \equiv \Big(\Gamma^*(P), K_Z^*(P)\Big) = \begin{cases} (0, \kappa), & \kappa \in [0, \infty) \text{ if } \quad |C| < 1 \\ \Big(-\frac{C^2-1}{CD}, \frac{D^2\kappa + K_V(1-C^2)}{C^2D^2}\Big), & \kappa \in [\kappa_{min}, \infty), \text{ if } \quad |C| > 1 \end{cases} \tag{I.36}$$

and the corresponding feedback capacity is given by

$$C^{B.1}_{A^\infty \to B^\infty}(\kappa) = \begin{cases} \frac{1}{2}\ln\frac{D^2\kappa + K_V}{K_V} & \text{if } |C| < 1, \ i.e., \ K_Z^* = \kappa \\ \frac{1}{2}\ln\frac{D^2K_Z^* + K_V}{K_V} & \text{if } |C| > 1, \ \kappa \in [\kappa_{min}, \infty) \end{cases} \tag{I.37}$$

For $\kappa \in [0, \kappa_{min})$ and $|C| \geq 1$ then feedback capacity does not exists. The feedback capacity expression (I.37), illustrates that the strategies $\Big(\Gamma^*, K_Z^*\Big) \equiv \Big(\Gamma^*(P), K_Z^*(P)\Big)$ depend on the solutions P of the quadratic Riccati equation (I.34). Clearly, there are multiple regimes, depending on whether the channel is stable, that is, $|C| < 1$ or unstable $|C| > 1$. Moreover, for unstable channels $|C| > 1$, feedback capacity does not exists, unless the power κ allocated for transmission, exceeds the critical level κ_{min}. For $|C| = 1$, to ensure a strictly positive capacity or rate it is necessary to take $Q > 0$. However, for any $|C| > 1$ then $\kappa_{min} \in (0, \infty)$ and there is a threshold effect for strictly positive feedback capacity. For $Q > 0$, it can be verified that there is always a threshold effect, because $\kappa_{min} \in (0, \infty)$, i.e., it is strictly positive. It should be mentioned that, since $Q = 0$, for $|C| > 1$ then $C^{B.1}_{A^\infty \to B^\infty}(\kappa)$ is the feedback capacity, when the maximal solution to the Riccati equation P_2 is used, and $\mathbf{P}_{B_{-1}}$ is the stationary distribution (see Example 13 for discussion).

(viii) Capacity Achieving Channel Input Distributions. From the above expressions, it follows that the capacity achieving channel input distribution is

$$\mathbf{P}^*_{A_i|B_{i-1}}(da_i|b_{i-1}) = \begin{cases} \mathbf{P}^{g,*}_{A_i}(da_i) \sim N(0, \kappa), \quad \kappa \in [0, \infty) & \text{if } |C| < 1 \\ \mathbf{P}^{g,*}_{A_i|B_{i-1}}(da_i|b_{i-1}) \sim N(\Gamma^*, K_Z^*), \quad \kappa \in [\kappa_{min}, \infty) & \\ & \text{if } |C| > 1. \end{cases} \tag{I.38}$$

This shows that if the channel is stable, $|C| < 1$, then feedback does not increase capacity, for the following reasons. As far as the limit $C^{B.1}_{A^\infty \to B^\infty}(\kappa)$ is concerned, there is no incentive to apply feedback, since the controlled process-the channel output process $\{B_i : i = 0, \ldots, n\}$, does not appear, neither in the transmission cost constraint nor in the characterization of the FTFI capacity expression given by (I.25). However, if $Q \neq 0$, then the controlled process $\{B_i : i = 0, \ldots, n\}$ is represented in the pay-off, and hence there is an incentive to apply feedback.

(ix) *Capacity Without Feedback.* Clearly, for stable channels, i.e., $|C| < 1$, the capacity without feedback of channel (I.20), denoted by $C^{noFB}_{A^\infty; B^\infty}(\kappa)$, is

$$C^{noFB}_{A^\infty \to B^\infty}(\kappa) = \frac{1}{2}\log\Big(1 + \frac{D^2\kappa}{K_V}\Big), \quad \text{for } |C| < 1, \ i.e., \ K_Z^* = \kappa. \tag{I.39}$$

This is precisely the value of capacity that is obtained from (I.36), (I.37). Note that this is precisely the capacity of a memoryless channel that corresponds to (I.20) with $C = 0$, i.e., $B_i = DA_i + V_i, i = 0, \ldots,$ that is, the memory of the channel (I.20), does not increase capacity.

Thus, the expressions of capacity without feedback and feedback, coincide for the case of stable channels, i.e., $|C| < 1$. This is attributed to the dual role of randomized strategies, specifically, the role of the deterministic part to control the channel output process. Since in this case, the channel is stable and $Q = 0$, no role is assigned to the randomized strategy, except to transmit information. However, if $Q > 0$ but the channel is stable, i.e., $|C| < 1$, the above observation does not hold.

(x) *Rate Loss of Unstable Channels.* For unstable channels, there is rate loss compared to the feedback capacity of stable channels, expressed in terms of the logarithm of the unstable eigenvalues of the channel, as follows.

$$\text{If } |C| > 1 \text{ then :}$$

$$C_{A^\infty \to B^\infty}^{B.1}(\kappa) = \frac{1}{2}\log\left(1+\frac{D^2\kappa}{K_V}\right) - \ln|C|, \ \forall \kappa \in [\kappa_{min}, \infty). \tag{I.40}$$

That is, $\kappa_{min} = \frac{(C^2-1)K_V}{D^2}$ is the threshold on power beyond which a strictly positive rate is feasible.

The rate loss of unstable channels given above, in terms of the expressions of capacity with feedback and capacity without feedback, $C_{A^\infty \to B^\infty}^{B.1}(\kappa)$, $C_{A^\infty \to B^\infty}^{noFB}(\kappa)$, respectively, is fundamentally different from the bounds derived by Cover and Pombra [2] and Ebert [6], i.e., given by (I.12). Specifically, as discussed above, for unstable channels, then feedback capacity exists and it is strictly positive, if the power κ is above a critical level or threshold $\kappa_{min} > 0$, which is the minimum cost to control the channel output, when $K_Z^* = 0$, i.e., it corresponds to zero feedback capacity or rate.

Finally, it noted that for the AGN channel (I.9), with AR(1) noise model defined by (I.14), then properties analogous to the ones discussed above should be identified, in terms of Riccati equations of mean-square estimation theory. A treatment in this direction is found in [11], for general Gaussian channels with past dependence on channel inputs and channel outputs.

5) Generalizations to Gaussian Channels with Arbitrary Memory: All properties discussed above are shown to hold, for general MIMO G-CM-B.1 and G-CM-B; they are obtained by invoking properties of matrix Riccati algebraic equations. These properties illustrate fundamental connections between capacity of channels with feedback, without feedback, linear systems theory, and LQG stochastic optimal control theory.

6) Relation Between Characterizations of FTFI Capacity and Coding Theorems: In Section VI, the importance of the characterizations of FTFI capacity are discussed in the context the direct and converse parts of channel coding theorems. Specifically, sufficient conditions are identified so that the per unit time limits of the characterizations of FTFI capacity, corresponds to feedback capacity, irrespectively of whether the channel models are Gaussian.

II. Information Structures of Channel Input Distributions of Extremum Problems of Feedback Capacity

In this section, the notation used throughout the paper is established, and the information structures of optimal channel input distributions, which maximize directed information, are recalled from [31]. Table II.1 lists all considered case with corresponding notation of FTFI capacity and feedback capacity of different channels and transmission cost functions.

All spaces are assumed to be complete and separable metric spaces, i.e., standard Borel spaces, to treat simultaneously discrete, finite alphabet, real-valued $\mathbb{R}^k$ or complex-valued $\mathbb{C}^k$ random processes for any positive integer k, etc. The product measurable space of the two measurable spaces $(\mathbb{X}, \mathscr{B}(\mathbb{X}))$ and $(\mathbb{Y}, \mathscr{B}(\mathbb{Y}))$ is denoted by $(\mathbb{X}\times\mathbb{Y}, \mathscr{B}(\mathbb{X})\otimes\mathscr{B}(\mathbb{Y}))$, where $\mathscr{B}(\mathbb{X})\otimes\mathscr{B}(\mathbb{Y})$ is the smallest $\sigma-$algebra containing all rectangles $\{A\times B : A\in\mathscr{B}(\mathbb{X}), B\in\mathscr{B}(\mathbb{Y})\}$. The probability distribution $\mathbf{P}(\cdot) \equiv \mathbf{P}_X(\cdot)$ on $(\mathbb{X}, \mathscr{B}(\mathbb{X}))$ induced by a Random Variable (RV) on $(\Omega, \mathscr{F}, \mathbb{P})$ by the mapping $X : (\Omega, \mathscr{F}) \to (\mathbb{X}, \mathscr{B}(\mathbb{X}))$ is defined by[6]

$$\mathbf{P}(A) \equiv \mathbf{P}_X(A) \stackrel{\triangle}{=} \mathbb{P}\big\{\omega\in\Omega : X(\omega)\in A\big\}, \quad \forall A\in\mathscr{B}(\mathbb{X}). \tag{II.41}$$

If the cardinality of $\mathbb{X}$ is finite then the RV is finite-valued and it is called a finite alphabet valued RV. Given another RV $Y : (\Omega, \mathscr{F}) \to (\mathbb{Y}, \mathscr{B}(\mathbb{Y}))$, then $\mathbf{P}_{Y|X}(dy|X)(\omega)$ is called the conditional distribution of RV Y given RV X. The RVs X and Y are called independent if and only if $\mathbf{P}_{Y|X}(dy|X)(\omega) = \mathbf{P}_Y(dy)$, $\mathbb{P}-$almost surely (a.s.), with $\mathbb{P}$ being the measure on which X is defined. However, unless stated otherwise, the qualifying "$\mathbb{P}-a.s.$" is omitted when dealing with conditional distributions. The conditional distribution of RV Y given $X = x$ is denoted by $\mathbf{P}_{Y|X}(dy|X=x) \equiv \mathbf{P}_{Y|X}(dy|x)$, and the joint distribution by $\mathbf{P}_{X,Y}(dx,dy) = \mathbf{P}_{Y|X}(dy|x)\otimes\mathbf{P}_X(dx)$. The family of conditional distributions on $(\mathbb{Y}, \mathscr{B}(\mathbb{Y}))$ parametrized by $x\in\mathbb{X}$, is defined by

$$\mathscr{K}(\mathbb{Y}|\mathbb{X}) \stackrel{\triangle}{=} \big\{\mathbf{P}(\cdot|x)\in\mathscr{M}(\mathbb{Y}) : \quad x\in\mathbb{X} \ \text{ and } \forall F\in\mathscr{B}(\mathbb{Y}), \text{ the function } \mathbf{P}(F|\cdot) \text{ is } \mathscr{B}(\mathbb{X})\text{-measurable.}\big\}. \tag{II.42}$$

The channel input and channel output alphabets are sequences of measurable spaces $\{(\mathbb{A}_i, \mathscr{B}(\mathbb{A}_i)) : i\in\mathbb{N}\}$ and $\{(\mathbb{B}_i, \mathscr{B}(\mathbb{B}_i)) : i\in\mathbb{N}\}$, respectively, and their history spaces are the product spaces $\mathbb{A}^{\mathbb{N}} \stackrel{\triangle}{=} \times_{i\in\mathbb{N}}\mathbb{A}_i$, $\mathbb{B}^{\mathbb{N}} \stackrel{\triangle}{=} \times_{i\in\mathbb{N}}\mathbb{B}_i$. These spaces are endowed with their respective product topologies, and $\mathscr{B}(\Sigma^{\mathbb{N}}) \stackrel{\triangle}{=} \otimes_{i\in\mathbb{N}}\mathscr{B}(\Sigma_i)$, where $\Sigma_i \in \{\mathbb{A}_i, \mathbb{B}_i\}$, $\Sigma^{\mathbb{N}} \in \{\mathbb{A}^{\mathbb{N}}, \mathbb{B}^{\mathbb{N}}\}$. Similarly, for $\mathscr{B}(\Sigma^n)$, when $n\in\mathbb{N}$ is finite. Points in Σ^n are denoted by $z^n \stackrel{\triangle}{=} \{z_0, z_1, \ldots, z_n\} \in \Sigma^n$, unless stated otherwise, while points in $\Sigma_k^m \stackrel{\triangle}{=} \times_{j=k}^m \Sigma_j$ are denoted by $z_k^m \stackrel{\triangle}{=} \{z_k, z_{k+1}, \ldots, z_m\} \in \Sigma_k^m$, $(k,m)\in\mathbb{N}\times\mathbb{N}$.

Channel Distribution with Memory: A sequence of stochastic kernels or distributions defined by

$$\mathscr{C}_{[0,n]} \stackrel{\triangle}{=} \Big\{\mathbf{P}_{B_i|B^{i-1},A^i} = Q_i(db_i|b^{i-1},a^i) \in \mathscr{K}(\mathbb{B}_i|\mathbb{B}^{i-1}\times\mathbb{A}^i) : i = 0,1,\ldots,n\Big\} \tag{II.43}$$

where $A^i \stackrel{\triangle}{=} \{A_0, A_1, \ldots, A_i\}$, $B^{i-1} \stackrel{\triangle}{=} \{B^{-1}, B_0, \ldots, B_{i-1}\}$, and B^{-1} is the initial state with distribution $\mathbf{P}_{B^{-1}} = \mu(db^{-1})$.

[6]The subscript on X is often omitted.

TABLE II.1
NOTATION OF MATHEMATICAL SYMBOLS

Notation	Definition
$\mathbb{Z}$	set of integer
$\mathbb{N}$	set of nonnegative integers $\{0,1,2,\dots\}$
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{R}^n$	set of n tuples of real numbers
$\mathbb{S}_+^{p\times p}$	set of symmetric positive semidefine $p\times p$ matrices $A\in\mathbb{R}^{p\times p}$
$\langle\cdot,\cdot\rangle$	inner product of elements of vectors spaces
$\mathbb{S}_{++}^{p\times p}$	subset of positive definite matrices of the set $\mathbb{S}_+^{p\times p}$
$\mathbb{D}_o \stackrel{\triangle}{=} \{c\in\mathbb{C}:\lvert c\rvert<1\}$	open unit disc of the space of compex numbers $\mathbb{C}$
$spec(A)\subset\mathbb{C}$	spectrum of a matrix $A\in\mathbb{R}^{p\times p}$ (set of all its eigenvalues)
$(\Omega,\mathscr{F},\mathbb{P})$	probability space, where $\mathscr{F}$ is the $\sigma-$algebra generated by subsets of Ω
$\mathscr{B}(\mathbb{W})$	Borel $\sigma-$algebra of a given topological space $\mathbb{W}$
$\mathscr{M}(\mathbb{W})$	set of all probability measures on $\mathscr{B}(\mathbb{W})$ of a Borel space $\mathbb{W}$
$\mathscr{K}(\mathbb{V}\vert\mathbb{W})$	set of all stochastic kernels on $(\mathbb{V},\mathscr{B}(\mathbb{V}))$ given $(\mathbb{W},\mathscr{B}(\mathbb{W}))$ of Borel spaces $\mathbb{W},\mathbb{V}$
$X\perp Y$	Independence of RVs X and Y

TABLE II.2
NOTATION OF CAPACITY SYMBOLS

Notation	Definition
$\mathbf{P}_{B_i\vert B^{i-1},A_i}$	Class A channel; (I.4)
$\mathbf{P}_{B_i\vert B_{i-M}^{i-1},A_i}$	Class B channel; (I.5)
$\gamma_i^A(a_i,b^i)$	Class A transmission cost function; (I.7)
$\gamma_i^B(a_i,b_{i-K}^i)$	Class B transmission cost function; (I.8)
$C_{A^n\to B^n}^{A}$	FTFI capacity of class A channel without transmission cost; (II.58)
$C_{A^n\to B^n}^{A}(\kappa)$	FTFI capacity of class A channel with class A or class B transmission cost; (II.64)
$C_{A^n\to B^n}^{B.J}(\kappa)$	FTFI capacity of class B channel and class B transmission cost, $J=\min\{M,K\}$; (II.66)
$C_{A^n\to B^n}^{IL-A}(\kappa)$	FTFI capacity of class A channel and class A transmission cost and information lossless strategies; (III.90)
$C_{A^n\to B^n}^{G-A}(\kappa)$	FTFI capacity of class A Gaussian channel and quadratic transmission cost $\gamma_i(a_i,b^{i-1})$; (IV.100)
$C_{A^n\to B^n}^{G-B.1}(\kappa)$	FTFI capacity of class B Gaussian channel, $M=1$ and quadratic transmission cost $\gamma_i(a_i,b_{i-1})$; (IV.154)
$C_{A^\infty\to B^\infty}^{G-B.1}(\kappa)$	Capacity of class B Gaussian channel $M=1$ and quadratic transmission cost $\gamma(a_i,b_{i-1})$; (V.213)

Thus, for $i=0$, the initial distribution is $Q_0(db_0\vert a_0,b^{-1})$. At each time instant i the conditional distribution of channel output B_i is affected causally by previous channel output symbols $b^{i-1}\in\mathbb{B}^{i-1}$ and current and previous channel input symbols $a^i\in\mathbb{A}^i, i=0,1,\dots,n$.

Channel Input Distribution with Feedback: A sequence of stochastic kernels defined by

$$\mathscr{P}_{[0,n]} \stackrel{\triangle}{=} \Big\{\mathbf{P}_{A_i\vert A^{i-1},B^{i-1}} = P_i(da_i\vert a^{i-1},b^{i-1}) \in \mathscr{K}(\mathbb{A}_i\vert\mathbb{A}^{i-1}\times\mathbb{B}^{i-1}) : i=0,1,\dots,n\Big\}. \quad \text{(II.44)}$$

For $i=0$ the distribution is $P_0(da_0\vert a^{-1},b^{-1})=P_0(da_0\vert b^{-1})$, which means the initial state $B^{-1}=b^{-1}$ is known to the encoder. At each time instant i the conditional distribution of channel input A_i is affected causally by past channel inputs and output symbols $(a^{i-1},b^{i-1})\in\mathbb{A}^{i-1}\times\mathbb{B}^{i-1}$, $i=0,1,\dots,n$.

Transmission Cost: The cost of transmitting and receiving symbols $a^n\in\mathbb{A}^n, b^n\in\mathbb{B}^n$ over the channel is a measurable function $c_{0,n}:\mathbb{A}^n\times\mathbb{B}^n\to[0,\infty)$. The set of channel input distributions with transmission cost is defined by

$$\mathscr{P}_{[0,n]}(\kappa) \stackrel{\triangle}{=} \Big\{P_i(da_i\vert a^{i-1},b^{i-1})\in\mathscr{K}(\mathbb{A}_i\vert\mathbb{A}^{i-1}\times\mathbb{B}^{i-1}),$$
$$i=0,\dots,n : \frac{1}{n+1}\mathbf{E}_\mu^P\Big(c_{0,n}(A^n,B^n)\Big)\le\kappa\Big\}\subset\mathscr{P}_{[0,n]},$$
$$c_{0,n}(a^n,b^n)\stackrel{\triangle}{=}\sum_{i=0}^n\gamma_i(T^ia^n,T^ib^n),\quad \kappa\in[0,\infty) \quad \text{(II.45)}$$

where for each i, T^ia^n, T^ib^n, may arbitrary (for now), for example, $T^ia^n=a_{i-L}^i, T^ib^n=b^i$ or $T^ib^n=b_{i-K}^i$, with L,K nonnegative finite integers, for $i=0,\dots,n$. $\mathbf{E}_\mu^P(\cdot)$ denotes expectation with respect to the joint distribution of RVs $(A^n,B^n)\stackrel{\triangle}{=}\{A_0,\dots,A_n,B^{-1},B_0,\dots,B_n\}$, for fixed distribution $\mathbf{P}_{B^{-1}}=\mu(dy^{-1})$, and superscript "P" indicates

its dependence on the choice of conditional distribution $\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \ldots, n\} \in \mathscr{P}_{[0,n]}(\kappa)$.

FTFI Capacity: Given any channel input conditional distribution $\big\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\big\} \in \mathscr{P}_{[0,n]}(\kappa)$, any channel distribution $\big\{Q(db_i|b^{i-1}, a^i) : i = 0, 1, \ldots, n\big\} \in \mathscr{C}_{[0,n]}$, and any distribution $\mu(db^{-1})$, the induced joint distribution $\mathbf{P}^P_\mu(da^n, db^n)$ is uniquely defined, as follows.[7]

$$\mathbb{P}\big\{A^n \in da^n, B^n \in db^n\big\}$$

$$\stackrel{\triangle}{=} \mathbf{P}^P_\mu(da^n, db^n) = \mu(db^{-1}) \otimes^n_{j=0}\Big(\mathbf{P}(db_j|b^{j-1}, a^j) \otimes \mathbf{P}(da_j|a^{j-1}, b^{j-1})\Big) \tag{II.46}$$

$$= \mu(db^{-1}) \otimes^n_{j=0}\Big(Q_j(db_j|b^{j-1}, a^j) \otimes P_j(da_j|a^{j-1}, b^{j-1})\Big). \tag{II.47}$$

The joint distribution of $\big\{B_0, \ldots, B_n\big\}$, conditioned on $B^{-1} = b^{-1}$, and the conditional distribution of B_i conditioned on $B^{i-1} = b^{i-1}$, are defined by[8]

$$\mathbb{P}\big\{B^n_0 \in db^n|B^{-1} = b^{-1}\big\} \stackrel{\triangle}{=} \mathbf{P}^P(db^n_0|b^{-1}) = \int_{\mathbb{A}^n} \mathbf{P}^P(da^n, db^n_0|b^{-1}) \tag{II.48}$$

$$\equiv \Pi^P_{0,n}(db^n_0|b^{-1}) = \otimes^n_{i=0}\Pi^P_i(db_i|b^{i-1}), \tag{II.49}$$

$$\Pi^P_i(db_i|b^{i-1}) = \int_{\mathbb{A}^i} Q_i(db_i|b^{i-1}, a^i) \otimes P_i(da_i|a^{i-1}, b^{i-1}) \otimes \mathbf{P}^P(da^{i-1}|b^{i-1}),\ i = 0, \ldots, n. \tag{II.50}$$

For $i = 0$, $\Pi^P_0(db_0|b^{-1}) = \int_{\mathbb{A}_0} Q_i(db_0|b^{-1}, a_0) \otimes P_i(da_0|b^{-1})$.

Directed information from $A^n \stackrel{\triangle}{=} \{A_0, \ldots, A_n\}$ to $\{B_0, \ldots, B_n\}$ conditioned on the initial state B^{-1}, and denoted by $I(A^n \to B^n)$ (for convenience), is defined by

$$I(A^n \to B^n) \stackrel{\triangle}{=} \mathbf{E}^P_\mu\Big\{\sum_{i=0}^n \log\Big(\frac{dQ_i(\cdot|B^{i-1}, A^i)}{d\Pi^P_i(\cdot|B^{i-1})}(B_i)\Big)\Big\} \tag{II.51}$$

$$= \sum_{i=0}^n \int_{\mathbb{A}^i\times\mathbb{B}^i} \log\Big(\frac{dQ_i(\cdot|b^{i-1}, a^i)}{d\Pi^P_i(\cdot|b^{i-1})}(b_i)\Big)\mathbf{P}^P_\mu(da^i, db^i) \tag{II.52}$$

Note that directed information is expressed in terms of the Radon-Nikodym derivatives between $Q_i(\cdot|b^{i-1}, a^i)$ and $\Pi^P_i(\cdot|b^{i-1})$, for $i = 0, \ldots, n$ (see [7], [32] or [28], [31]). If for any i, the distribution $Q_i(\cdot|b^{i-1}, a^i)$ is not absolutely continuous with respect to the distribution $\Pi^P_i(\cdot|b^{i-1})$, then directed information takes the value $+\infty$. Strictly speaking the proper definition of directed information is via relative entropy [7], which admits the value $+\infty$. If the probability distributions can be expressed in terms of probability density functions, then $Q_i(db_i|b^{i-1}, a^i) = f_{B_i|B^{i-1},A^i}(b_i|b^{i-1}, a^i)db_i$, $\Pi^P_i(db_i|b^{i-1}) = f^P_{B_i|B^{i-1}}(b_i|b^{i-1})db_i$, $i = 0, \ldots, n$, etc., and (II.51) reduces to

$$I(A^n \to B^n) = \mathbf{E}^P_\mu\Big\{\sum_{i=0}^n \log\Big(\frac{f_{B_i|B^{i-1},A^i}(B_i|B^{i-1}, A^i)}{f^P_{B_i|B^{i-1}}(B_i|B^{i-1})}\Big)\Big\}. \tag{II.53}$$

The FTFI capacity with and without transmission cost constraints, $C_{A^n\to B^n}(\kappa)$ and $C_{A^n\to B^n}$, respectively, are defined as follows.

$$C_{A^n\to B^n}(\kappa) \stackrel{\triangle}{=} \sup_{\mathscr{P}_{[0,n]}(\kappa)} I(A^n \to B^n), \tag{II.54}$$

$$C_{A^n\to B^n} \stackrel{\triangle}{=} \sup_{\mathscr{P}_{[0,n]}} I(A^n \to B^n). \tag{II.55}$$

Throughout the paper it is assumed that the supremum in (II.55) exists and it is achieved in the sets (see [31] for sufficient conditions based weak compactness of probability measures). That is, $\kappa \in [0, \infty)$ is sufficiently large for $\mathscr{P}_{[0,n]}(\kappa)$ to be non-empty.

For the per unit time limiting version $C_{A^\infty\to B^\infty}(\kappa)$ of $C_{A^n\to B^n}(\kappa)$, to be a candidate of feedback capacity, and thus a candidate of the characterization of the supremum of all achievable rates (via direct and converse channel coding theorems), the following assumption is imposed throughout the paper. For any process $\{X_i : i = 0, \ldots, \}$, which may represent the source process to be encoded and transmitted over the channel, the following conditional independence, pointed out by Massey [33] holds.

$$\mathbf{P}_{B_i|B^{i-1},A^i,X^k} = \mathbf{P}_{B_i|B^{i-1},A^i} \Longleftrightarrow X^k \leftrightarrow (A^i, B^{i-1}) \leftrightarrow B_i, \quad \forall k \in \{0, 1, \ldots, \}, \quad i = 0, \ldots, \tag{II.56}$$

The next theorem summarizes certain results derived in [1]; they are recalled, because they are extensively used in this paper.

Theorem 1 (Characterization of FTFI Capacity for Channels of Class A or B and Transmission Cost of Class A or B [1]):

(1) Suppose the channel distribution is of Class A defined by (I.4). Define the restricted class of channel input distributions $\overline{\mathscr{P}}^A_{[0,n]} \subset \mathscr{P}_{[0,n]}$ by

$$\overline{\mathscr{P}}^A_{[0,n]} \stackrel{\triangle}{=} \Big\{\big\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\big\} \in \mathscr{P}_{[0,n]} : P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i(da_i|b^{i-1}),\ i = 0, 1, \ldots, n\Big\}. \tag{II.57}$$

Then the following hold.

(1.a) The maximization of $I(A^n \to B^n)$ defined by (II.51) over $\mathscr{P}_{[0,n]}$ occurs in $\overline{\mathscr{P}}^A_{[0,n]} \subset \mathscr{P}_{[0,n]}$ and the characterization of FTFI capacity is given by the following expression.

$$C^A_{A^n\to B^n} = \sup_{\{\pi_i(da_i|b^{i-1})\in\mathscr{M}(\mathbb{A}_i):i=0,\ldots,n\}} \Big\{\mathbf{E}^\pi_\mu\Big\{\sum_{i=0}^n \log\Big(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi^{\pi_i}_i(\cdot|B^{i-1})}(B_i)\Big)\Big\}\Big\} \tag{II.58}$$

[7]Notation $\otimes$ is used to denote compound probability distributions generated by multi-fold integrals.

[8]Throughout the paper the superscript notation on P in $\mathbf{P}^P_\mu(\cdot)$, $\Pi^P_{0,n}(\cdot)$, *etc.*, indicates the dependence of these distributions on the channel input conditional distribution, while subscript notation on μ indicates the dependence on the distribution μ of the initial state B^{-1}.

where

$$\Pi_i^{\pi}(db_i|b^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b^{i-1}, a_i) \otimes \pi_i(da_i|b^{i-1}), \quad i = 0, \ldots, n, \tag{II.59}$$

$$\mathbf{P}_{\mu}^{\pi}(da^i, db^i) = \mu(db^{-1}) \otimes_{j=0}^{i} \Big(Q_j(db_j|b^{j-1}, a_j) \otimes \pi_j(da_j|b^{j-1}) \Big). \tag{II.60}$$

(1.b) Suppose the following two conditions hold.

$$(b.1) \quad \gamma_i(T^i a^n, T^i b^n) = \gamma_i^{A}(a_i, b^i) \quad or$$
$$\gamma_i(T^i a^n, T^i b^n) = \gamma_i^{B.K}(a_i, b_{i-K}^i), \ i = 0, \ldots, n, \tag{II.61}$$

$$(b.2) \quad C_{A^n \to B^n}^{A}(\kappa) \stackrel{\triangle}{=} \sup_{\mathscr{P}_{[0,n]}(\kappa)} I(A^n \to B^n) \tag{II.62}$$

$$= \inf_{s \geq 0} \sup_{\left\{P_i(da_i|a^{i-1}, b^{i-1}): i=0,\ldots,n\right\} \in \mathscr{P}_{[0,n]}} \Big\{ I(A^n \to B^n) - s\Big\{ \mathbf{E}_{\mu}^{P}\Big(c_{0,n}(A^n, B^n) \Big) - \kappa(n+1) \Big\} \Big\}. \tag{II.63}$$

Then the maximization of $I(A^n \to B^n)$ defined by (II.51) over $\left\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, \ldots, n\right\} \in \mathscr{P}_{[0,n]}(\kappa)$ occurs in $\overline{\mathscr{P}}_{[0,n]}^{A} \bigcap \mathscr{P}_{[0,n]}(\kappa)$ and the FTFI capacity is given by the following expression.

$$C_{A^n \to B^n}^{A}(\kappa) = \sup_{\pi_i(da_i|b^{i-1}) \in \mathscr{M}(\mathbb{A}_i), i=0,\ldots,n: \frac{1}{n+1}\mathbf{E}_{\mu}^{\pi}\left\{c_{0,n}(A^n, B^n)\right\} \leq \kappa} \Big\{ \mathbf{E}_{\mu}^{\pi}\Big\{ \sum_{i=0}^{n} \log\Big(\frac{dQ_i(\cdot|B^{i-1}, A_i)}{d\Pi_i^{\pi_i}(\cdot|B^{i-1})}(B_i) \Big) \Big\} \Big\}. \tag{II.64}$$

(2) Suppose the channel distribution is of Class B defined by (I.5), that is, $Q_i(db_i|b_{i-M}^{i-1}, a_i), i = 0, \ldots, n$, a transmission cost is imposed $\mathscr{P}_{0,n}(\kappa)$, corresponding to $\left\{\gamma_i^{B}(a_i, b_{i-K}^i), i = 0, \ldots, n\right\}$, and the analogue of (1), (b.2) holds. Define the restricted class of channel input distributions $\overset{\circ}{\mathscr{P}}{}_{[0,n]}^{B.J} \subset \mathscr{P}_{[0,n]}$ by

$$\overset{\circ}{\mathscr{P}}{}_{[0,n]}^{B.J} \stackrel{\triangle}{=} \Big\{ \Big\{P_i(da_i|a^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\Big\} \in \mathscr{P}_{[0,n]} : P_i(da_i|a^{i-1}, b^{i-1}) = \pi_i(da_i|b_{i-J}^{i-1}) : i = 0, 1, \ldots, n \Big\} \tag{II.65}$$

where $J \stackrel{\triangle}{=} \max\{K, M\}$. Then the following hold.

(2.a) The maximization of $I(A^n \to B^n)$ defined by (II.51) over $\left\{P_i(da_i|a^{i-1}, b^{i-1}), i = 0, \ldots, n\right\} \in \mathscr{P}_{0,n}(\kappa)$ occurs in $\overset{\circ}{\mathscr{P}}{}_{[0,n]}^{B.J} \bigcap \mathscr{P}_{[0,n]}(\kappa)$, and the characterization of FTFI capacity is given by the following expression.

$$C_{A^n \to B^n}^{B.J}(\kappa) = \sup_{\pi_i(da_i|b_{i-J}^{i-1}) \in \mathscr{M}(\mathbb{A}_i), i=0,\ldots,n: \frac{1}{n+1}\mathbf{E}_{\mu}^{\pi}\left\{c_{0,n}(A^n, B^n)\right\} \leq \kappa} \Big\{ \mathbf{E}_{\mu}^{\pi}\Big\{ \sum_{i=0}^{n} \log\Big(\frac{dQ_i(\cdot|B_{i-M}^{i-1}, A_i)}{dv_i^{\pi}(\cdot|B_{i-J}^{i-1})}(B_i) \Big) \Big\} \Big\} \tag{II.66}$$

where

$$\mathbf{P}_{\mu}^{\pi}(db^i, da^i) = \mu(db_{-J}^{-1}) \otimes_{j=0}^{i} \Big(Q_j(db_j|b_{j-M}^{j-1}, a_j) \otimes \pi_j(da_j|b_{j-J}^{j-1}) \Big), \quad i = 0, \ldots, n, \tag{II.67}$$

$$v_i^{\pi}(db_i|b_{i-J}^{i-1}) = \int_{\mathbb{A}_i} Q_i(db_i|b_{i-M}^{i-1}, a_i) \otimes \pi_i(da_i|b_{i-J}^{i-1}). \tag{II.68}$$

(2.b) Suppose the channel distribution is of Class B, and the maximization of $I(A^n \to B^n)$ defined by (II.51), is over channel input conditional distributions with transmission cost $\mathscr{P}_{0,n}(\kappa)$, corresponding to $\{\gamma_i^{A}(a_i, b^i) : i = 0, \ldots, n\}$, and the analogue of (1), (b.2) holds.

Then the maximization of $I(A^n \to B^n)$ over $\left\{P_i(da_i|a^{i-1}, b^{i-1}), i = 0, \ldots, n\right\} \in \mathscr{P}_{[0,n]}(\kappa)$ occurs in $\overline{\mathscr{P}}_{[0,n]}^{A} \bigcap \mathscr{P}_{[0,n]}(\kappa)$.

Remark 2 (Equivalence of Constraint and Unconstraint Problems): Assuming existence of maximizing elements, then the equivalence of constraint and unconstraint problems in Theorem 1, can be shown by invoking Lagrange's duality theory of globally optimizing convex functionals over convex sets [34]. Specifically, by [31] the set of distributions $\mathbf{P}^{C1}(da^n|b^{n-1}) \stackrel{\triangle}{=} \otimes_{i=0}^{n} P_i(da_i|a^{i-1}, b^{i-1}) \in \mathscr{M}(\mathbb{A}^n)$ is convex, and this uniquely defines $\mathscr{P}_{[0,n]}$ and vice-versa, directed information as a functional of $\mathbf{P}^{C1}(da^n|b^{n-1}) \in \mathscr{M}(\mathbb{A}^n)$ is convex, and by the linearity the constraint set $\mathscr{P}_{[0,n]}(\kappa)$ expressed in $\mathbf{P}^{C1}(da^n|b^{n-1})$, is convex. Hence, if a maximizing distribution exists, and the so-called Slater condition holds (i.e., a sufficient condition is the existence of an interior point to the constraint set), then for every $s \geq 0$ (not equal to zero), then the constraint and unconstraint problems are equivalent.

III. Realization of Channel Input Distributions by Information Lossless Randomized Strategies

In this section, alternative characterizations of FTFI capacity given in Theorem 1 are obtained by inducing, i.e., realizing the optimal channel input conditional distributions by information lossless randomized strategies, driven by independent RVs, using as an intermediate step, independent uniformly distributed RVs. Application examples to Gaussian channel models with memory are given in Section IV and Section V.

The principle idea exploited is based on Lemma 3 (see [35]). This lemma states that, for any family of conditional distributions (on Polish spaces), conditioned on an information structure (i.e., parametrized by the conditioning variables), there exist deterministic functions, measurable with respect to the conditioning information structure and an additional

real-valued uniform RV taking values in $[0,1]$, such that conditional distributions can be replaced by the Lebesgue measure of such deterministic functions.

Lemma 3 (Realization of Conditional Distributions by Randomized Strategies Lemma 1.2 in [35]): Let $\mathbf{P}(\cdot|w)$ be a family of measures on Polish space $(\mathbb{V},\mathscr{B}(\mathbb{V}))$, $w\in\mathbb{W}$, (i.e., $(\mathbb{W},\mathscr{B}(\mathbb{W}))$ a measurable space).

Let $\mathscr{B}([0,1])$ be the $\sigma-$algebra of Borel sets on $[0,1]$ and $\mathbf{m}(\cdot)$ the Lebesgue measure on $[0,1]$.

If $\mathbf{P}(A|w)$ is $\mathscr{B}(\mathbb{W})-$measurable in $w\in\mathbb{W}$ for all $A\in\mathscr{B}(\mathbb{V})$, then there exists a function $f:\mathbb{W}\times[0,1]\to\mathbb{V}$, $(w,t)\mapsto a\stackrel{\triangle}{=}f(w,t)$, measurable with respect to $\mathscr{B}(\mathbb{W})\otimes\mathscr{B}([0,1])$ such that

$$\mathbf{m}\Big\{t\in[0,1]: f(w,t)\in A\Big\}=\mathbf{P}(A|w),\quad \forall A\in\mathscr{B}(\mathbb{V}) \quad \text{and} \quad \forall w\in\mathbb{W}. \tag{III.69}$$

Since Lemma 3 holds for general distributions defined on complete separable metric spaces $\mathbb{V},\mathbb{W}$, these also include distributions induced by arbitrary RVs, such as, continuous, countable, finite etc., valued RVs. The function $f(w,\cdot)$ in Lemma 3 is a randomization with respect to a uniform RV taking values in $[0,1]$. The derivation of Lemma 3 utilizes the fact that, for any Polish space $(\mathbb{V},\mathscr{B}(\mathbb{V}))$, there exists a one-to-one function $\lambda:(\mathbb{V},\mathscr{B}(\mathbb{V}))\to([0,1],\mathscr{B}([0,1]))$ such that $\lambda(A)$ is a Borel set on $[0,1]$ for all $A\in\mathscr{B}(\mathbb{V})$ (i.e., $\lambda(A)$ is the image of A under λ). Any distribution function on $\mathbb{V}$ is a candidate for such a function $\lambda(x)$, and hence by introducing the quantile representation of $\lambda(x)$, then one can assume without loss of generality that $\mathbb{V}$ coincides with $[0,1]$.

The definition of the quantile representation of distributions and some of their properties are presented in the next remark, for subsequent application.

Remark 4 (Quantile Representation of Distributions): Let $\mathscr{F}(\mathbb{X})$ denote the space of probability distributions $F(x)$ on $\mathbb{X}$, and define the corresponding set of quantile functions $\mathscr{Q}_F([0,1])$ by

$$\mathscr{Q}_F([0,1]) \stackrel{\triangle}{=} \Big\{G:(0,1)\to\mathbb{X}: G(\cdot)\ \textit{is monotonically increasing, continuous on the left, with limits on the right}\Big\}.$$

For any $F(\cdot)\in\mathscr{F}(\mathbb{X})$ define

$$\mathbf{I}[F](u)\stackrel{\triangle}{=}\inf\Big\{x\in\mathbb{X}: F(x)\geq u\Big\},\quad \forall u\in(0,1). \tag{III.70}$$

Then the quantile function of $F(\cdot)\in\mathscr{F}(\mathbb{X})$ is $G(\cdot)\stackrel{\triangle}{=}\mathbf{I}[F](\cdot)$ and the following well-known properties hold.

(1) $\mathbf{I}:\mathscr{F}(\mathbb{X})\to\mathscr{Q}_F([0,1])$ is a bijection map.

(2) For any integrable measurable function $\varphi:\mathbb{X}\to\mathbb{V}$ then

$$\mathbf{E}\Big\{\varphi(X)\Big\}=\int_{\mathbb{X}}\varphi(x)dF(x)=\int_{(0,1)}\varphi(G(u))du. \tag{III.71}$$

If $F(\cdot)\in\mathscr{F}(\mathbb{X})$ is strictly increasing then $G(\cdot)\stackrel{\triangle}{=}\mathbf{I}[F](\cdot)=F^{-1}(\cdot), u\in(0,1)\mapsto G(u)=\mathbf{I}[F](u)=F^{-1}(u)$, which is the usual inverse of a distribution function $F(\cdot)$, and $du=\frac{du}{dx}dx=\frac{dF(x)}{dx}dx=f(x)dx$, where the last equality holds provided the density function $f(\cdot)$ of the distribution $F(\cdot)$ exists.

A. Recursive Nonlinear Channel Models A and B

The material of this section are developed for nonlinear channel models, which induce channel distributions of Class A, B, as defined below.

Definition 5 (Nonlinear Channel Models and Transmission Costs A and B):

(1) NCM-A. Nonlinear channel models A are defined by nonlinear recursive models and transmission cost functions, as follows.

$$B_i=h_i^A(B^{i-1},A_i,V_i),\quad B^{-1}=b^{-1},\quad i=0,\ldots,n,$$
$$\frac{1}{n+1}\mathbf{E}_\mu\Big\{\sum_{i=0}^{n}\gamma_i^A(A_i,B^i)\Big\}\leq\kappa \tag{III.72}$$

where $\{V_i: i=0,1,\ldots,n\}$ is the noise process. The underlying assumptions are the following.

(1.a) $h_i^A:\mathbb{B}^{i-1}\times\mathbb{A}_i\times\mathbb{V}_i\to\mathbb{B}_i$, $\gamma_i^A:\mathbb{A}_i\times\mathbb{B}^i\to\mathbb{A}_i$ and $h_i^A(\cdot,\cdot,\cdot)$, $\gamma_i^A(\cdot,\cdot)$ are measurable functions, for $i=0,1,\ldots,n$.

(1.b) The noise process $\{V_i: i=0,\ldots,n\}$ satisfies

$$\mathbf{P}_{V_i|V^{i-1},A^i,B^{-1}}(dv_i|v^{i-1},a^i,b^{-1})=\mathbf{P}_{V_i}(dv_i),\quad i=0,\ldots,n. \tag{III.73}$$

(2) NCM-B. Nonlinear channel models B are defined as follows.

$$B_i=h_i^{B.M}(B_{i-M}^{i-1},A_i,V_i),\quad B_{-M}^{-1}=b_{-M}^{-1},\quad i=0,\ldots,n,$$
$$\frac{1}{n+1}\mathbf{E}_\mu\Big\{\sum_{i=0}^{n}\gamma_i^{B.K}(A_i,B_{i-K}^i)\Big\}\leq\kappa \tag{III.74}$$

where $\{V_i: i=0,1,\ldots,n\}$ is the noise process, and the underlying assumptions are the following.

(2.a) Conditions (1.a), (1.b) hold with appropriate changes.

In the above definition, there is no assumption on the alphabet spaces of the RVs, i.e., whether these RVs are continuous-valued, finite or countable-valued, or combinations of them. To ensure Theorem 1 applies to the NCMs of Definition 5, it is sufficient to show such NCMs induce any of the channel distributions described in Theorem 1. The following calculation verifies that model (III.72) induces a conditional channel distribution of Class A. By (III.73), the following consistency condition holds.

$$\mathbb{P}\Big\{B_i\in\Gamma\Big|B^{i-1}=b^{i-1},A^i=a^i\Big\}$$
$$=\mathbf{P}_{V_i}\Big(v_i: h_i^A(b^{i-1},a_i,v_i)\in\Gamma\Big),\quad \Gamma\in\mathscr{B}(\mathbb{B}_i) \tag{III.75}$$
$$=Q_i(\Gamma|b^{i-1},a_i),\quad i=0,1,\ldots,n. \tag{III.76}$$

The convention for model (III.72) is that transmission starts at time $i=0$, and the initial data $B^{-1}=b^{-1}\equiv b_{-\infty}^{-1}$ are specified, and their distribution is fixed. Another alternative convention is to assume

$$B_0=h_0(B^{-1},A_0,V_0)\equiv h_0(A_0,V_0),$$
$$\gamma_0^A(A_0,B^0)\equiv\gamma_0^A(A_0,B_0),$$
$$B_1=h_1(B^0,A_1,V_1)\equiv h_1(B_0,A_1,V_1),$$
$$\gamma_1^A(A_1,B^1)=\gamma_1(A_1,B_0,B_1),\ldots,n.$$

This alternative convention means that no information is available prior to transmission, that is, the sigma algebra generated by B^{-1} is $\sigma\{B^{-1}\} = \{\Omega, \emptyset\}$ (the trivial $\sigma-$algebra), and by Theorem 1, the optimal channel input distribution at time $i = 0$ is $\pi_0(da_0|b^{-1})) = \pi_0(da_0)$.

Similarly, for NCM-B defined by (III.74), by using (III.73), the following consistency condition holds.

$$\mathbb{P}\Big\{B_i \in \Gamma \Big| B^{i-1} = b^{i-1}, A^i = a^i\Big\}$$

$$= \mathbf{P}_{V_i}\Big(v_i : h_i^{B.M}(b_{i-M}^{i-1}, a_i, v_i) \in \Gamma\Big), \quad \Gamma \in \mathscr{B}(\mathbb{B}_i) \quad \text{(III.77)}$$

$$= Q_i(\Gamma|b_{i-M}^{i-1}, a_i), \quad i = 0, 1, \ldots, n. \quad \text{(III.78)}$$

Hence, NCM-B given by (III.74) induces a channel distribution of Class B. It is not necessary to introduced additional NCMs which are combinations of channels of Class A or B and transmission costs of Class A or B, because these are included in the above models.

B. Alternative Characterization of FTFI Capacity for NCM-A

Consider the NCM-A given by (III.72), i.e., Definition 5, (1). By invoking Lemma 3 and Remark 4, then a property called information lossless of randomized strategies is identified, and an alternative characterization of the FTFI capacity given in Theorem 1, (1), is obtained, as stated in the next theorem.

Theorem 6 (Characterization of FTFI Capacity for NCM-A by Information Lossless Randomized Strategies):
Consider the characterization of FTFI capacity, $C^{A}_{A^n \to B^n}(\kappa)$, of Theorem 1, (1), for the NCM-A of Definition 5, (1). Then the following hold on some constructed probability space $(\overline{\Omega}, \overline{\mathscr{F}}, \overline{\mathbb{P}})$.

(a) For each channel input distribution from the set $\overline{\mathscr{P}}^{A}_{[0,n]}$ defined by (II.57), the consistency conditions CON(a.1), (a.2) stated below hold.

CON(a.1). There exist functions $\overline{e}_i^A(\cdot,\cdot)$ measurable with respect to the information structure $\mathscr{I}_i^{\overline{e}^A} \stackrel{\triangle}{=} \{b^{i-1}, u_i\}$, $i = 0, 1, \ldots, n$ and defined by

$$\overline{e}_i^A : \mathbb{B}^{i-1} \times \mathbb{U}_i \to \mathbb{A}_i, \quad \mathbb{U}_i \stackrel{\triangle}{=} [0,1],$$
$$a_i = \overline{e}_i^A(b^{i-1}, u_i), \quad i = 0, 1, \ldots, n \quad \text{(III.79)}$$

such that $\{U_i : i = 0, \ldots, n\}$ are uniform RVs on $\mathbb{U}^{n+1}$ and

$$\mathbf{P}_{A_i|B^{i-1}}(da_i|b^{i-1}) \equiv \pi_i(da_i|b^{i-1}), \quad i = 0, 1, \ldots, n$$
$$= \mathbf{P}_{U_i}\Big(u_i : \overline{e}_i^A(b^{i-1}, u_i) \in da_i\Big). \quad \text{(III.80)}$$

CON(a.2). i) A_i is conditionally independent of A^{i-1} given B^{i-1}, for each $i = 0, 1, \ldots, n$, ii) U_i is independent of $\big(U^{i-1}, V^{i-1}\big), i = 0, 1, \ldots, n$, and iii) V_i is independent of $\big(V^{i-1}, U^{i}\big), i = 0, 1, \ldots, n$.

Moreover,[9]

$$A_i = \overline{e}_i^A(B^{i-1}, U_i), \quad i = 0, 1, \ldots, n, \quad \text{(III.81)}$$

$$B_i = h_i^A(B^{i-1}, A_i, V_i), \quad B^{-1} = b^{-1}, \quad i = 0, 1, \ldots, n, \quad \text{(III.82)}$$

$$\frac{1}{n+1}\mathbf{E}_{\mu}^{\overline{e}^A}\Big\{\sum_{i=0}^{n} \gamma_i^A(\overline{e}_i^A(B^{i-1}, U_i), B^i)\Big\} \le \kappa. \quad \text{(III.83)}$$

(b) Let $\{Z_i : i = 0, \ldots, n\}$ be a sequence of independent RVs taking values in $\{\mathbb{Z}_i : i = 0, \ldots, n\}$ with corresponding sequence of distributions $\{F_{Z_i}(\cdot) \in \mathscr{F}(\mathbb{Z}_i) : i = 0, \ldots, n\}$.
Then (a) holds with the following changes.

$$\mathbb{U}_i \longmapsto \mathbb{Z}_i, \quad (u_i, U_i) \longmapsto (z_i, Z_i),$$
$$\overline{e}_i^A(b^{i-1}, u_i) \longmapsto e_i^A(b^{i-1}, z_i),$$
$$\mathbf{P}_{U_i} \longmapsto \mathbf{P}_{Z_i}, \quad i = 0, \ldots, n. \quad \text{(III.84)}$$

Moreover,

$$C^{A}_{A^n \to B^n}(\kappa)$$
$$= \sup_{\{\mathbf{P}_{Z_i}:i=0,\ldots,n\}, \{e_i^A(\cdot,\cdot):i=0,\ldots,n\} \in \mathscr{E}^A_{[0,n]}(\kappa)} \Big\{$$
$$\mathbf{E}_{\mu}^{e^A}\Big\{\sum_{i=0}^{n} \log\Big(\frac{dQ_i(\cdot|B^{i-1}, e_i^A(B^{i-1}, Z_i))}{d\Pi_i^{e^A}(\cdot|B^{i-1})}(B_i)\Big)\Big\}\Big\} \quad \text{(III.85)}$$

where

$$\mathscr{E}^A_{[0,n]}(\kappa)$$
$$\stackrel{\triangle}{=} \Big\{e_i^A(b^{i-1}, z_i), i = 0, \ldots, n :$$
$$\frac{1}{n+1}\mathbf{E}_{\mu}^{e^A}\Big(\sum_{i=0}^{n} \gamma_i^A(e_i^A(B^{i-1}, Z_i), B^i)\Big) \le \kappa\Big\}. \quad \text{(III.86)}$$

Define the restricted class of randomized strategies called information lossless randomized strategies, as follows.

$$\mathscr{E}^{IL-A}_{[0,n]}$$
$$\stackrel{\triangle}{=} \Big\{e_i^A : \mathbb{B}^{i-1} \times \mathbb{Z}_i \to \mathbb{A}_i,\ a_i = e_i^A(b^{i-1}, z_i),\ i = 0, \ldots, n :$$
$$\{Z_i : i = 0, \ldots, n\} \ \text{an independent process,}$$
$$\text{for fixed } b^{i-1}, \text{ the map } e_i^A(b^{i-1}, \cdot) : \mathbb{Z}_i \to \mathbb{A}_i \ \text{is a}$$
$$\text{bijection and its inverse measurable}$$
$$\text{for } i = 0, \ldots, n\Big\}, \quad \text{(III.87)}$$

$$\mathscr{E}^{IL-A}_{[0,n]}(\kappa)$$
$$\stackrel{\triangle}{=} \Big\{\big\{e_i^A(b^{i-1}, z_i) : i = 0, \ldots, n\big\} \in \mathscr{E}^{IL-A}_{[0,n]} :$$
$$\frac{1}{n+1}\mathbf{E}_{\mu}^{e^A}\Big(\sum_{i=0}^{n} \gamma_i^A(e_i^A(B^{i-1}, Z_i), B^i)\Big) \le \kappa\Big\}. \quad \text{(III.88)}$$

Then an alternative equivalent characterization of FTFI capacity $C^{FB,A}_{A^n \to B^n}(\kappa)$ defined by (II.58), is given by the

[9] Superscript notation "$\mathbf{E}_{\mu}^{\overline{e}^A}\{\cdot\}$" indicates the dependence of the joint distribution on the strategy $\{\overline{e}_i^A(\cdot,\cdot) : i = 0, \ldots, n\}$.

following expression.

$$C^{A}_{A^n\to B^n}(\kappa) = C^{IL-A}_{A^n\to B^n}(\kappa) \stackrel{\triangle}{=} \sup_{\{\mathbf{P}_{Z_i}:i=0,\ldots,n\},\{e^A_i(\cdot,\cdot):i=0,\ldots,n\}\in\mathscr{E}^{IL-A}_{[0,n]}(\kappa)} \Big\{ \mathbf{E}^{e^A}_{\mu}\Big\{\sum_{i=0}^{n}\log\Big(\frac{dQ_i(\cdot|B^{i-1},e^A_i(B^{i-1},Z_i))}{d\Pi^{e^A}_i(\cdot|B^{i-1})}(B_i)\Big)\Big\}\Big\} \tag{III.89}$$

$$\equiv \sup_{\{\mathbf{P}_{Z_i}:i=0,\ldots,n\},\{e^A_i(\cdot,\cdot):i=0,\ldots,n\}\in\mathscr{E}^{IL-A}_{[0,n]}(\kappa)} \sum_{i=0}^{n} I^{e^A}(Z_i;B_i|B^{i-1}) \tag{III.90}$$

and superscript e^A in $I^{e^A}(\cdot;\cdot|\cdot)$ indicates the dependence of the distributions on the strategies $\{e^A_i(\cdot,\cdot): i=0,\ldots,n\}\in \mathscr{E}^{IL-A}_{[0,n]}(\kappa)$.

Proof: See Appendix A. □

An application of Theorem 6, (b) is illustrated in Theorem 10. It is noted that the context of Theorem 6 is different from posterior matching schemes of memoryless channels [36].

Remark 7: (Comments on Theorem 6)

(a) Given a specific NCM-A, then Theorem 6, (b) shows that, any candidate of optimal channel input distribution $\pi_i(da_i|b^{i-1})$ can be replaced by a randomized strategy $\{e^A_i(b^{i-1},Z_i): i=0,\ldots,b\}$ driven by independent RVs $\{Z_i: i=0,\ldots,n\}$, from the set of randomized strategies $\mathscr{E}^A_{[0,n]}(\kappa)$. Moreover, for such channels, directed information $I(A^n\to B^n)=\sum_{i=0}^n I(A_i;B_i|B^{i-1})$, which is a functional of channel input conditional distributions $\{\pi_i(da_i|b^{i-1}): i=0,\ldots,n\}$, i.e., $I(A_i;B_i|B^{i-1}) \equiv I^{\pi}(A_i;B_i|B^{i-1})$, can be expressed as a functional of randomized strategies $\{e^A_i(b^{i-1},z_i): i=0,\ldots,n\}\in\mathscr{E}^A_{[0,n]}$, i.e., $I^{\pi}(A_i;B_i|B^{i-1}) = I^{e^A}(A_i;B_i|B^{i-1}), i=0,\ldots,n$. However, under appropriate conditions, i.e., for information lossless randomized strategies $\{e^A_i(b^{i-1},z_i): i=0,\ldots,n\}\in\mathscr{E}^{IL-A}_{[0,n]}$ then $I^{\pi}(A_i;B_i|B^{i-1}) = I^{e^A}(Z_i;B_i|B^{i-1}), i=0,\ldots,n$.

(b) By Remark 4, (1) and (2), then an alternative characterization of the FTFI capacity can be obtained with respect to uniform RVs $\{U_i: i=0,\ldots,n\}$, as follows. For any $\{e^A_i(b^{i-1},z_i): i=0,\ldots,n\}\in\mathscr{E}^A_{[0,n]}$, let $z_i=G(u_i)=\mathbf{I}[F_{Z_i}](u_i)$, and define $a_i=\overline{e}^A_i(b^{i-1},u_i) = e^A_i(b^{i-1},G(u_i)), i=0,\ldots,n$. By Theorem 6, then

$$I(A^n\to B^n) = \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) = \sum_{i=0}^{n} I(e^A_i(B^{i-1},G(U_i));B_i|B^{i-1}) \tag{III.91}$$

(c) For memoryless channels, i.e., $Q_i(db_i|b^{i-1},a_i) = Q_i(db_i|a_i)$ and $\gamma_i{}^A(a_i,b^{i-1})=\gamma_i(a_i)$, $i=0,\ldots,n$, the distribution, which maximizes the characterization of FTFI capacity is $\mathbb{P}\{A_i\le a_i\}\stackrel{\triangle}{=}F_{A^*_i}(a_i), i=0,\ldots,n$. In this case, $a_i=e_i(G(u_i))$, and the optimal randomized strategies are given by $e^*_i(G(u_i)) = F^{-1}_{A^*_i}(u_i)$, that is, the optimal process is $A^*_i=Z^*_i, i=0,\ldots,n$. This is due to the fact an arbitrary distributed RVs can be generated from uniform RVs.

(d) To show the fundamental difference between Theorem 6 and posterior matching schemes of memoryless channels [36], consider channels with memory, say, $Q_i(db_i|b^{i-1},a_i) = Q_i(db_i|b^{i-1},a_i)$ and $\gamma_i{}^A(a_i,b^i)$, $i=0,\ldots,n$. Then any candidate of optimal distribution corresponding to the characterization of FTFI capacity, in general, satisfies

$$\mathbb{P}\{A_i\le a_i|A^{i-1}=a^{i-1},B^{i-1}=b^{i-1}\} = \mathbb{P}\{A_i\le a_i|B^{i-1}=b^{i-1}\}\neq F_{A_i}(a_i) \tag{III.92}$$

hence

$$a_i = e^A_i(b^{i-1},G(u_i))\neq \tilde{e}_i(G(u_i)),\quad i=0,\ldots,n. \tag{III.93}$$

Moreover, $a_i=e^A_i(b^{i-1},G(u_i))\neq\tilde{e}_i(G_{b^{i-1}}(u_i)), G_{b^{i-1}}(\cdot)\stackrel{\triangle}{=}\mathbf{I}[F_{A_i|B^{i-1}}](\cdot|b^{i-1})$, because such a construction violates Lemma 3, i.e., the information structure of randomized strategies. This subtle issue is further clarified via application examples of Gaussian channels with memory, in Section IV.

C. Alternative Characterization of FTFI Capacity for NCM-B

Consider the NCM-B defined by (III.74), i.e., Definition 5, (2). By Theorem 1, (2), the corresponding optimal channel input distribution are of the form $\{\pi_i(da_i|b^{i-1}_{i-J}): i=0,1,\ldots,n\}$, $J\stackrel{\triangle}{=}\max\{M,K\}$. Clearly, all the material of Section III-B apply to NCM-B. The analog of Theorem 6 is stated below for future reference.

Theorem 8 (Characterization of FTFI Capacity for NCM-B by Information Lossless Randomized Strategies):

Consider the characterization of FTFI capacity, $C^{B.J}_{A^n\to B^n}(\kappa)$, given in Theorem 1, (2), for the NCM-B of Definition 5, (2). Then the following hold on some constructed probability space $(\overline{\Omega},\overline{\mathscr{F}},\overline{\mathbb{P}})$.

(a) For each channel input distribution $\overset{\circ}{\mathscr{P}}{}^{B.J}_{[0,n]}$ defined by (II.65), the consistency conditions CON(b.1), (b.2) stated below hold.

CON(b.1). There exists a function $\overline{e}^{B.J}_i(\cdot)$, $J\stackrel{\triangle}{=}\max\{M,K\}$ measurable with respect to the information structure $\mathscr{I}^{\overline{e}^{B.J}}_i\stackrel{\triangle}{=}\{b^{i-1}_{i-J},u_i\}, i=0,1,\ldots,n$ defined by

$$\overline{e}^{B.J}_i : \mathbb{B}^{i-1}_{i-J}\times\mathbb{U}_i\to\mathbb{A}_i,\quad \mathbb{U}_i\stackrel{\triangle}{=}[0,1],$$
$$a_i=\overline{e}^{B.J}_i(b^{i-1}_{i-J},u_i), i=0,1,\ldots,n \tag{III.94}$$

such that $\{U_i: i=0,1,\ldots,n\}$ are uniform distributed on $[0,1]^{n+1}$ and

$$\mathbf{P}_{A_i|B^{i-1}_{i-J}}(da_i|b^{i-1}_{i-J}) = \mathbf{P}_{U_i}\Big(u_i:\overline{e}^{B.J}_i(b^{i-1}_{i-J},u_i)\in da_i\Big),$$
$$i=0,1,\ldots,n,\quad J\stackrel{\triangle}{=}\max\{M,K\}. \tag{III.95}$$

CON(b.2). i) A_i is conditionally independent of $\{A^{i-1},B^{i-J-1}\}$ given $\{B^{i-1}_{i-J}: i=0,\ldots,n\}$ for $i=0,\ldots,n$, ii) U_i is independent of $\Big(U^{i-1},V^{i-1}\Big), i=0,\ldots,n$, iii) V_i is independent of $\Big(V^{i-1},U^{i}\Big), i=0,\ldots,n$.

Moreover, the analog of (III.81)-(III.83) hold.

(b) Let $\{Z_i : i = 0, \ldots, n\}$ *be a sequence of independent RVs taking values in* $\{\mathbb{Z}_i : i = 0, \ldots, n\}$ *with corresponding sequence of distributions* $\{F_{Z_i}(\cdot) \in \mathscr{F}(\mathbb{Z}_i) : i = 0, \ldots, n\}$.

Then the analog of Theorem 6, (b) holds.

Proof: See Appendix B. □

Remark 9 (Alternative Characterizations): The main point to be made is that, the characterizations of FTFI capacity, which are extremum problems with respect to channel input conditional distributions, can be transformed into alternative equivalent characterizations, which are extremum problems over randomized strategies driven by independent random variables.

IV. Characterizations of FTFI Capacity of Gaussian CMs A and B & Connections to the LQG Theory

In this section, the characterizations of FTFI capacity given in Section II and Section III are applied to Gaussian channel models (G-CMs), which are special cases of NCM-A, NCM-B of Definition 5, to obtain the following.

(a) Characterizations of FTFI capacity for multiple input multiple output (MIMO) G-CMs;
(b) characterizations of FTFI capacity for MIMO G-CMs via connections to finite horizon linear-quadratic-Gaussian (LQG) stochastic optimal control theory, matrix Riccati difference equations, and water-filling solutions of MIMO channels;
(c) unfold the dual role of randomized strategies, which realize optimal channel channel input processes that correspond to the characterizations of FTFI capacity, to control the channel output process and to transmit new information over the channel.

The characterizations of feedback capacity and its connections to the infinite horizon LQG stochastic optimal control theory and stability theory of linear control systems is treated in Section V, by using per unit time limiting versions of the results obtained in this section.

A. Characterizations of FTFI Capacity for Gaussian Channel Models A

Consider a Gaussian channel model A (G-CM-A), i.e., a special case of the NCM-A given by (III.72), and defined as follows[10]

$$B_i = \sum_{j=0}^{i-1} C_{i,j} B_j + D_{i,i}\ A_i + V_i, \quad B_0 = D_{0,0} A_0 + V_0, \quad i = 1, \ldots, n, \tag{IV.96}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{ \sum_{i=0}^{n} \Big(\langle A_i, R_{i,i} A_i \rangle + \langle B^{i-1}, Q_i(i-1) B^{i-1} \rangle \Big) \Big\} \leq \kappa, \tag{IV.97}$$

$$C_{i,j} \in \mathbb{R}^{p \times p}, \quad D_{i,i} \in \mathbb{R}^{p \times q}, \quad R_{i,i} \in \mathbb{S}_{++}^{q \times q},$$

$$Q_0(-1) = 0,\ Q_i(i-1) \in \mathbb{S}_{+}^{ip \times ip}, \quad i = 0, \ldots, n, \quad j = 0, \ldots, i-1 \tag{IV.98}$$

where $B^i \stackrel{\triangle}{=} (B_0, B_1, \ldots, B_i)$. Thus, at time $i = 0$, A_0 does not use feedback, equivalently, $\sigma\{B^{-1}\} = \{\Omega, \emptyset\}$. The following assumption holds.

Assumption A: 1) Definition 5, (1.a), (1.b) hold, and 2) the noise process $\{V_i : i = 0, \ldots, n\}$ is Gaussian distributed, specified by

$$V_i \sim N(0, K_{V_i}), \quad K_{V_i} \succ 0, \quad i.e., \quad \mu_{V_i} \stackrel{\triangle}{=} \mathbf{E}\{V_i\} = 0,$$
$$K_{V_i} \stackrel{\triangle}{=} Cov(V_i, V_i) = \mathbf{E}\{V_i V_i^T\}, \quad i = 0, 1, \ldots, n. \tag{IV.99}$$

The next theorem states that the optimal channel input distribution is Gaussian, and it is realized by an information lossless Gaussian randomized strategy that is expressed via the decomposition $A_i = g_i(B^{i-1}) + Z_i$, in which $g_i(B^{i-1}) \perp Z_i, i = 0, \ldots, n$, $\{g_i(\cdot) : i = 0, \ldots, n\}$ is a deterministic function of the feedback or output process, and $\{Z_i : i = 0, \ldots, n\}$ is an orthogonal innovations process.

Theorem 10 (Characterization of FTFI Capacity for G-CM-A):

Consider the G-CM-A defined by (IV.96)-(IV.98), and suppose Assumption A holds. Let $\{(A_i^g, B_i^g) : i = 0, \ldots, n\}$ *denote a jointly Gaussian process.*

Then the following hold.

(a) The optimal channel input distribution $\{\pi(da_i|b^{i-1}) \equiv \pi^g(da_i|b^{i-1}) : i = 0, \ldots, n\}$ *is conditionally Gaussian, with conditional mean which is linear in* $\{b_i : i = 0, \ldots, n\}$ *and conditional covariance which is non-random, and the characterization of FTFI capacity is given by the following expression.*

$$C_{A^n \to B^n}^{G-A}(\kappa) \stackrel{\triangle}{=} \sup_{\mathscr{P}_{[0,n]}^{G-A}(\kappa)} H(B^{g,n}) - H(V^n) \tag{IV.100}$$

where

$$\mathscr{P}_{[0,n]}^{G-A}(\kappa) \stackrel{\triangle}{=} \Big\{ \pi_i^g(da_i|b^{i-1}), i = 0, \ldots, n : \frac{1}{n+1} \sum_{i=0}^{n} \mathbf{E}^{\pi^g} \Big(\langle A_i^g, R_{i,i} A_i^g \rangle + \langle B^{g,i-1}, Q_i(i-1) B^{g,i-1} \rangle \Big) \leq \kappa \Big\}. \tag{IV.101}$$

(b) An alternative equivalent characterization of the FTFI capacity is given by the following expressions.

$$C_{A^n \to B^n}^{G-A}(\kappa) \stackrel{\triangle}{=} \sup_{\{(\Gamma_i(i-1), K_{Z_i}), i=0,\ldots,n\} \in \mathscr{C}_{[0,n]}^{IL-G-A}(\kappa)} \Big\{ H(B^{g,n}) - H(V^n) \Big\} \tag{IV.102}$$

where

$$\mathscr{C}_{[0,n]}^{IL-G-A}(\kappa) \stackrel{\triangle}{=} \Big\{ (\Gamma_i(i-1), K_{Z_i}), i = 0, \ldots, n : \frac{1}{n+1} \mathbf{E}\Big\{ \sum_{i=0}^{n} \big(\langle A_i^g, R_{i,i} A_i^g \rangle$$

[10] There is no loss of generality of not considering the more general transmission cost function $\langle A_i, R_{i,i} A_i \rangle + \langle B^i, Q_i(i) B^i \rangle$, because we can substitute B_i by (IV.96) and re-define the cost function.

$$+\langle B^{g,i-1}, Q_i(i-1)B^{g,i-1}\rangle\Big) \leq \kappa\Big\}\Big\} \tag{IV.103}$$

$$\begin{aligned} H(B^{g,n}) - H(V^n) &= \sum_{i=0}^{n} H(B_i^g|B^{g,i-1}) - H(V^n) \\ &= \frac{1}{2}\sum_{i=0}^{n} \log \frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|}, \end{aligned} \tag{IV.104}$$

$$A_i^g = \sum_{j=0}^{i-1} \Gamma_{i,j} B_j^g + Z_i, \quad i = 0, 1, \ldots, n, \tag{IV.105}$$

$$\equiv \Gamma_i(i-1)B^{g,i-1} + Z_i, \tag{IV.106}$$

$$\begin{aligned} B_i^g &= \sum_{j=0}^{i-1} C_{i,j} B_j^g + D_{i,i} A_i^g + V_i \\ &= \sum_{j=0}^{i-1} \Big(C_{i,j} + D_{i,i}\Gamma_{i,j}\Big) B_j^g + D_{i,i} Z_i + V_i, \end{aligned} \tag{IV.107}$$

$$\equiv \Big(C_i(i-1) + D_{i,i}\Gamma_i(i-1)\Big)B^{g,i-1} + D_{i,i}Z_i + V_i, \tag{IV.108}$$

$$\begin{aligned} &i)\ \ Z_i \text{ is independent of } (A^{g,i-1}, B^{g,i-1}),\ i = 0, \ldots, n, \\ &ii)\ \ Z^i \text{ is independent of } V^i, \quad i = 0, \ldots, n, \\ &iii)\ \ \Big\{Z_i \sim N(0, K_{Z_i}) : i = 0, 1, \ldots, n\Big\} \text{ is an} \\ &\qquad \text{orthogonal innovations Gaussian process.} \end{aligned} \tag{IV.109, IV.110}$$

Proof: The derivation is based on the maximum entropy property of Gaussian distribution, and the decomposition (IV.106) expressed in terms of an orthogonal process $\{Z_i : i = 0, \ldots, n\}$. The details are given in Appendix C. □

Remark 11 (Extremum Solution of the G-CM-A):
(a) The connection of decomposition (IV.105) to the Cover and Pombra [2] realization of Gaussian channel input process given by (I.10) is done as follows. Substituting in the right hand side of (IV.105) the output process (IV.107), then the process $\{A_i^g : i = 0, \ldots, n\}$ is expressed in terms of the channel noise process $\{V_i : i = 0, \ldots, n\}$ and a linear combination of the process $\{Z_i : i = 0, \ldots, n\}$, by

$$\begin{aligned} &A^{g,n} = \overline{\Gamma}^n V^n + \overline{Z}^n, \quad \text{where} \\ &\qquad \Big\{\overline{Z}_i : i = 0, \ldots, n\Big\} \text{ is Gaussian and Correlated} \\ &\qquad \overline{Z}^n \text{ is Gaussian } N(0, K_{\overline{Z}^n}), \quad V^n \perp \overline{Z}^n \end{aligned} \tag{IV.111}$$

and $\overline{\Gamma}^n$ is a lower diagonal matrix with time-varying deterministic entries. It should be noted that, for the above equivalent realization is very difficult to optimize the corresponding characterization of FTFT capacity given by (IV.102), even in the special case, $Q_i(i-1) = 0, i = 0, \ldots, n$, because the process $\{\overline{Z}_i : i = 0, \ldots, n\}$ is not an orthogonal innovations process. Any past attempts to solve the Cover and Pombra [2], characterization given by (I.10), for any n, that is, corresponding to the nonstationary nonergodic case, have been unsuccessful. Previous attempts are extensively elaborated in [3].

(b) Although, at first glance, the problem of determining the optimal matrices $\{\Gamma_i^*(i-1), K_{Z_i}^*), i = 0, \ldots, n\}$, which correspond to the extremum problem (IV.102), appears difficult, even in special cases, one possible re-formulation, is to compactly represent (IV.102), as follows.

From (IV.105), (IV.107), it is always possible to find lower diagonal matrices $\{(C_{[i,i]}, \Gamma_{[i,i]}) : i = 0, \ldots, n\}$ and matrices $\{D_{[i,i]} : i = 0, \ldots, n\}$, such that the following hold.

$$A^{g,i} = \Gamma_{[i,i]} B^{g,i} + Z^i, \quad i = 0, \ldots, n, \tag{IV.112}$$

$$B^{g,i} = C_{[i,i]} B^{g,i} + D_{[i,i]} A^i + V^i, \quad i = 0, \ldots, n. \tag{IV.113}$$

By the above expressions, then the covariance of the channel output process is given by

$$K_{B^{i-1}} \stackrel{\triangle}{=} \mathbf{E}\Big\{B^{g,i-1}\big(B^{g,i-1}\big)^T\Big\}, \quad i = 0, 1, \ldots, n, \tag{IV.114}$$

$$\begin{aligned} &= \Big(I - C_{[i-1,i-1]} - D_{[i-1,i-1]}\Gamma_{[i-1,i-1]}\Big)^{-1} \\ &\quad .D_{[i-1,i-1]}\Big(K_{Z^{i-1}} + K_{V^{i-1}}\Big)D_{[i-1,i-1]}^T \\ &\quad .\Big(I - C_{[i-1,i-1]} - D_{[i-1,i-1]}\Gamma_{[i-1,i-1]}\Big)^{-1,T}, \\ &\quad spec\Big(C_{[i-1,i-1]} + D_{[i-1,i-1]}\Gamma_{[i-1,i-1]}\Big) < 1. \end{aligned} \tag{IV.115}$$

The condition $spec\Big(C_{[i-1,i-1]} + D_{[i-1,i-1]}\Gamma_{[i-1,i-1]}\Big) < 1$, $i = 0, \ldots, n$ is equivalent to the existence of a sequence $\{\Gamma_{i,j} : i = 0, \ldots, n, j = 0, \ldots, i-1\}$, which ensures the eigenvalues of the channel output process lie in the open unit disc in the space of complex numbers $\mathbb{C}$. Utilizing the above representations, the average transmission cost constraint is given by

$$\begin{aligned} &\mathscr{C}_{[0,n]}^{IL-G-A}(\kappa) \\ &= \Big\{\Big(\Gamma_i(i-1), K_{Z_i}\Big), i = 0, \ldots, n : \\ &\quad \frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}\Big(\langle A_i^g, R_{i,i}A_i^g\rangle \\ &\quad + \langle B^{g,i-1}, Q_i(i-1)B^{g,i-1}\rangle\Big)\Big\}\Big\} \\ &= \frac{1}{n+1}\sum_{i=0}^{n} tr\Big(R_{i,i}\Gamma_i(i-1)K_{B^{i-1}}\Gamma_i^T(i-1) \\ &\quad + R_{i,i}K_{Z_i} + Q_i(i-1)K_{B^{i-1}}\Big) \leq \kappa\Big\}. \end{aligned} \tag{IV.116}$$

Hence, the FTFI capacity is characterized by

$$\begin{aligned} &C_{A^n \to B^n}^{G-A}(\kappa) \\ &= \max_{\Big\{\Big(\Gamma_i(i-1), K_{Z_i}\Big), i=0,\ldots,n\Big\} \in \mathscr{C}_{[0,n]}^{IL-G-A}(\kappa),\ \textit{(IV.115) holds}} \Big\{ \\ &\quad \frac{1}{2}\sum_{i=0}^{n} \log \frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|}\Big\} \end{aligned} \tag{IV.117}$$

Extremum problem (IV.117) is a deterministic optimization problem. However, although compactly represented and attractive, it is not at all easy to optimize, because the functional dependence of $\{K_{B^{i-1}} : i = 0, 1, \ldots, n\}$ *on* $\{\Gamma_i(i-1), K_{Z_i} : i = 0, \ldots, n\}$*, is very complex. Hence, this re-formulation is not pursued any further. Rather, extremum problem (IV.117) is re-visited in Section IV-E, where closed form expressions are obtained via direct connections to linear-quadratic-Gaussian (LQG) stochastic optimal control problems.*

B. Characterizations of FTFI Capacity for Gaussian Channel Models B.1

Consider the Gaussian channel model B.1 (G-CM-B.1), i.e., a special case of NCM-B with $M = 1$, and defined by

$$B_i = C_{i,i-1}B_{i-1} + D_{i,i}A_i + V_i, \quad B_{-1} = b_{-1}, \quad i = 0, \ldots, n, \tag{IV.118}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}\langle A_i, R_{i,i}A_i\rangle + \langle B_{i-1}, Q_{i,i-1}B_{i-1}\rangle\Big\} \leq \kappa,$$
$$R_{i,i} \in \mathbb{S}_{++}^{q\times q}, \quad Q_{i,i-1} \in \mathbb{S}_{+}^{p\times p}, \quad i = 0, \ldots, n \tag{IV.119}$$

under the following assumption.

Assumption B: 1) Definition 5, (2.a) holds, and 2) the noise $\{V_i \sim N(0, K_{V_i}) : i = 0, 1, \ldots, n\}$ is independent and Gaussian distributed, as in (IV.99) and independent of the zero mean Gaussian RV B_{-1} (and a maximizing element exists in the set of channel input distributions).

It should be noted that one can consider the case when $B_{-1} = b_{-1}$ is fixed, hence all expressions below are replaced by expectations for fixed $B_{-1} = b_{-1}$.

Clearly, all statements regarding the G-CM-A, defined by (IV.96)-(IV.98) (given in Section IV-A), can be specialized to G-CM-B.1. The following statements are listed for future reference.

1) Characterization of the FTFI Capacity: The characterization of the FTFI capacity of G-CM.B.1 is given by

$$C_{A^n\to B^n}^{G-B.1}(\kappa) = \sup_{\{\pi_i^g(da_i|b_{i-1}), i=0,\ldots,n\}\in \mathring{\mathscr{P}}_{[0,n]}^{G-B.1}(\kappa)} \sum_{i=0}^{n} H(B_i^g|B_{i-1}^g) - H(V^n) \tag{IV.120}$$

where

$$\mathring{\mathscr{P}}_{[0,n]}^{G-B.1}(\kappa) \triangleq \Big\{\pi_i^g(da_i|b_{i-1}), i = 0, \ldots, n : \frac{1}{n+1}\mathbf{E}^{\pi^g}\sum_{i=0}^{n}\Big(\langle A_i^g, R_{i,i}A_i^g\rangle + \langle B_{i-1}^g, Q_{i,i-1}B_{i-1}^g\rangle\Big) \leq \kappa\Big\} \tag{IV.121}$$

$$\mathbb{P}\{B_i^g \leq b_i | B_{i-1}^g = b_{i-1}\} = \int_{\mathbb{A}_i}\mathbb{P}\{V_i \leq b_i - C_{i,i-1}b_{i-1} - D_{i,i}a_i\}\,\pi_i^g(da_i|b_{i-1}), \quad i = 0, 1, \ldots, n \tag{IV.122}$$

that is, $\{\pi_i^g(da_i|b_{i-1}) \equiv \mathbf{P}_{A_i|B_{i-1}}^g(a_i|b_{i-1}) : i = 0, 1, \ldots, n\}$ is conditionally Gaussian, satisfying the average transmission cost constraint, implying $\{\mathbf{P}_{B_i|B_{i-1}}(b_i|b_{i-1}) \equiv \mathbf{P}_{B_i|B_{i-1}}^g(b_i|b_{i-1}) : i = 0, 1, \ldots, n\}$ is also conditionally Gaussian, both with conditional mean which is linear in $\{b_i : i = 0, \ldots, n\}$ and conditional covariance which is non-random.

2) Alternative Characterization of FTFI Capacity: By Theorem 10, the set of all channel input conditional distribution is realized by randomized strategies, as follows.

$$A_i^g = e_i^{B.1}(B_{i-1}^g, Z_i) = \Gamma_{i,i-1}B_{i-1}^g + Z_i, \quad i = 0, \ldots, n, \tag{IV.123}$$

$$B_i^g = \Big(C_{i,i-1} + D_{i,i}\Gamma_{i,i-1}\Big)B_{i-1}^g + D_{i,i}Z_i + V_i, \quad B_{-1}^g = b_{-1}, \quad i = 0, \ldots, n, \tag{IV.124}$$

$$i)\ \ Z_i \text{ independent of } \Big(A^{g,i-1}, B^{g,i-1}\Big), \quad i = 0, \ldots, n,$$
$$ii)\ \ Z^i \text{ independent of } V^i, \quad \text{for } i = 0, \ldots, n,$$
$$iii)\Big\{Z_i \sim N(0, K_{Z_i}) : i = 0, \ldots, n\Big\} \text{ independent Gaussian process.} \tag{IV.125}$$

The following are easily obtained, from the above equation.

$$\mu_{B_i|B_{i-1}} \triangleq \mathbf{E}\Big\{B_i^g\Big|B_{i-1}^g\Big\} = \Big(C_{i,i-1} + D_{i,i}\Gamma_{i,i-1}\Big)B_{i-1}^g, \quad i = 0, \ldots, n, \tag{IV.126}$$

$$K_{B_i|B_{i-1}} \triangleq \mathbf{E}\Big\{\Big(B_i^g - \mu_{B_i|B_{i-1}}\Big)\Big(B_i^g - \mu_{B_i|B_{i-1}}\Big)^T\Big|B_{i-1}^g\Big\} = D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}, \quad i = 0, \ldots, n, \tag{IV.127}$$

$$K_{B_i} \triangleq \mathbf{E}\Big\{B_i^g\big(B_i^g\big)^T\Big\}, \quad \text{satisfies the discrete time-varying Lyapunov equation} \tag{IV.128}$$

$$K_{B_i} = \Big(C_{i,i-1} + D_{i,i}\Gamma_{i,i-1}\Big)K_{B_{i-1}}\Big(C_{i,i-1} + D_{i,i}\Gamma_{i,i-1}\Big)^T + D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}, \quad i = 0, \ldots, n, \tag{IV.129}$$

$$K_{B_{-1}} = \text{Given}. \tag{IV.130}$$

Consequently, the alternative characterization of the FTFI capacity is given, as follows.

$$C_{A^n\to B^n}^{G-B.1}(\kappa) = C_{A^n\to B^n}^{IL-G-B.1}(\kappa) \triangleq \sup_{\{(\Gamma_{i,i-1}, K_{Z_i}), i=0,\ldots,n\}\in\mathscr{C}_{[0,n]}^{IL-G-B.1}(\kappa)} \sum_{i=0}^{n} H(B_i^g|B_{i-1}^g) - H(V^n) \text{ and (IV.129), (IV.130) hold} \tag{IV.131}$$

where

$$\sum_{i=0}^{n} H(B_i^g|B_{i-1}^g) - H(V^n) = \frac{1}{2}\sum_{i=0}^{n}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|}, \tag{IV.132}$$

$$\mathscr{C}_{[0,n]}^{IL-G-B.1}(\kappa) \triangleq \Big\{\big(\Gamma_{i,i-1}, K_{Z_i}\big), i=0,\ldots,n : \frac{1}{n+1}\mathbf{E}\sum_{i=0}^{n}\Big(\langle A_i^g, R_{i,i}A_i^g\rangle + \langle B_{i-1}^g, Q_{i,i-1}B_{i-1}^g\rangle\Big) = \frac{1}{n+1}\sum_{i=0}^{n} tr\Big(R_{i,i}\Gamma_{i-1,i}K_{B_{i-1}}\Gamma_{i,i-1}^T + R_{i,i}K_{Z_i} + Q_{i,i-1}K_{B_{i-1}}\Big) \leq \kappa\Big\}. \quad \text{(IV.133)}$$

This is a classical deterministic optimization problem of a dynamical system, described by the covariance of the channel output process $\{K_{B_i} : i = 0,\ldots,n\}$, and satisfying the discrete time-varying Lyapunov type difference equation (IV.129), (IV.130), where $\{K_{B_i} : i = 0,\ldots,n\}$ is the controlled object, while the control object is $\{(\Gamma_{i,i-1}, K_{Z_i}) : i = 0,\ldots,n\}$, and it is chosen to maximize the pay-off. Discrete time-varying Lyapunov type difference equations are extensively utilized in stability analysis of time-varying linear controlled systems (see Appendix E or [4], Chapter 7).

The next remark elaborates further on the direct connection between the characterization of FTFI capacity and Discrete-time Lyapunov matrix equations its per unit time limiting version, and linear stochastic controlled systems.

Remark 12 (Relations of FTFI Capacity and Feedback Capacity of G-CM-B.1 & Linear Stochastic Controlled Systems):
(a) The recursive equation (IV.129) satisfied by the covariance $\{K_{B_i} : i = 0,\ldots,n\}$ of the output process $\{B_i^g : i = 0,\ldots,n\}$ is a Lyapunov type matrix difference equation. It is possible to apply calculus of variations to determine the pair $\{(\Gamma_{i,i-1}, K_{Z_i}) \in \mathbb{R}^{q\times p}\times\mathbb{S}_+^{q\times q} : i = 0,\ldots,n\}$, which incurs the maximum in (IV.131). However, since this is done in a subsequent section via dynamic programming, this direction is not pursued any further.

(b) Suppose the coefficients of the G-CM-B.1 defined by (IV.118), (IV.119) are time-invariant, and the parameters of the optimal channel input distributions induced by (IV.123), are restricted to time-invariant, i.e.,

$$C_{i,i-1} = C,\ D_{i,i} = D,\ K_{V_i} = K_V,\ R_{i,i} = R,\quad i = 0,\ldots,n,\qquad Q_{i,i-1} = Q,\ i = 0,\ldots,n-1,\ Q_{n,n-1} = M, \quad \text{(IV.134)}$$

$$(\Gamma_{i,i-1}, K_{Z_i}) = (\Gamma, K_Z),\quad i = 0,\ldots,n. \quad \text{(IV.135)}$$

Recursive substitution gives

$$K_{B_i} = \Big(\big[C+D\Gamma\big]^i\Big)K_{B_0}\Big(\big[C+D\Gamma\big]^i\Big)^T + \sum_{j=0}^{i-1}\Big(\big[C+D\Gamma\big]^j\Big)\Big(DK_ZD^T + K_V\Big)\times\Big(\big[C+D\Gamma\big]^j\Big)^T,\quad i = 1,\ldots,n. \quad \text{(IV.136)}$$

Next, the properties of time-invariant Lyapunov difference and algebraic equations, given in Appendix E, Theorem 27 are utilized to analyze the FTFI capacity and feedback capacity of the G-CM-B.1.

Suppose the set of all eigenvalues of $(C + D\Gamma)$ lie in the open unit disc of the space of complex numbers $\mathbb{C}$, i.e., $spec\big(C + D\Gamma\big) \subset \mathbb{D}_o$. Then, irrespectively of the initial covariance K_{B_0}, the limit, $\lim_{n\longrightarrow\infty} K_{B_i} = K_B$ exists and satisfies the Lyapunov algebraic matrix equation

$$K_B = \Big(C + D\Gamma\Big)K_B\Big(C + D\Gamma\Big)^T + DK_ZD^T + K_V \quad \text{and } K_B \succeq 0 \text{ is a unique solution} \quad \text{(IV.137)}$$

However, if $K_{B_0} = K_B$, then the solution of the Lyapunov matrix difference equation (IV.129) with time-invariant coefficients is time-invariant.

The per unit time limiting version of the characterization of the FTFI capacity is given by the following expression.

$$C_{A^\infty\to B^\infty}^{G-B.1}(\kappa) \triangleq \sup_{\mathscr{C}_{[0,\infty]}^{IL-G-B.1}(\kappa),\ \text{(IV.137) holds}}\Big\{\frac{1}{2}\log\frac{|DK_ZD^T + K_V|}{|K_V|}\Big\}, \quad \text{(IV.138)}$$

$$\mathscr{C}_{[0,\infty]}^{IL-G-B.1}(\kappa) \triangleq \Big\{(\Gamma, K_Z) \in \mathbb{R}^{q\times p}\times\mathbb{S}_+^{q\times q} : tr\Big(R\Gamma K_B\Gamma^T + RK_Z + QK_B\Big) \leq \kappa\Big\},\quad spec\big(C + D\Gamma\big) \subset \mathbb{D}_o. \quad \text{(IV.139)}$$

If $spec\big(C + D\Gamma\big) \subset \mathbb{D}_o$, and there is a unique invariant distribution of the transition kernel $\mathbf{P}_{B_i|B_{i-1}}$, then the joint distribution of the joint process $\{A_i^g, B_i^g) : i = 0,\ldots,\}$ and its marginals are asymptotically ergodic, and hence (IV.138) is the feedback capacity. Appendix E, Theorem 27, gives sufficient conditions, which imply $spec\big(C + D\Gamma\big) \subset \mathbb{D}_o$, and existence of per unit time limiting version of the characterization of the FTFI capacity, and existence of unique invariant distribution of the joint process $\{(A_i^g, B_i^g) : i = 0,\ldots,\}$. The complete analysis is done in Section V via dynamic programming.

In the next example it is demonstrated that the statements described by (I.34)-(I.40), which are derived by invoking of the algebraic Riccati equation (I.33), can also be derived by an alternative method that uses (IV.137)-(IV.139).

Example 13 (Scalar Channel $p = q = 1$ and $R = 1$, $Q = 0$): The explicit solution of feedback capacity (IV.138) is obtained below. From (IV.137), then

$$K_B = \frac{D^2K_Z + K_V}{1 - \big(C + D\Gamma\big)} \quad \text{if } |C + D\Gamma| < 1. \quad \text{(IV.140)}$$

The constraint optimization problem (IV.138) is convex, and by substituting (IV.140) into (IV.139), it is equivalent to the following unconstraint optimization (see [34]).

$$J(K_Z^*, s^*) \triangleq \inf_{s\geq 0}\sup_{\Gamma\in\mathbb{R}, K_Z\geq 0}\Big\{\frac{1}{2}\log\frac{D^2K_Z + K_V}{K_V} - s\Big(\Gamma^2\frac{D^2K_Z + K_V}{1 - \big(C + D\Gamma\big)} + K_Z - \kappa\Big)\Big\},\quad |C + D\Gamma| < 1 \quad \text{(IV.141)}$$

where $s \geq 0$ is the Lagrange multiplier associated with the constraint. The above problem gives the following

optimal solution.

If $|C| < 1$ *then:*

$$\Gamma^* = 0, \quad K_Z^* = \kappa, \quad \kappa \in [0, \infty). \tag{IV.142}$$

If $|C| > 1$ *then:*

$$\Gamma^* = -\frac{C^2 - 1}{CD}, \quad K_Z^* = \frac{D^2\kappa + K_V(1 - C^2)}{C^2 D^2} \geq 0, \quad \kappa \in [\kappa_{min}, \infty), \tag{IV.143}$$

$$s^* = \frac{1}{2}\frac{D^2}{D^2\kappa + K_V} \in [s^*_{min}, \infty), \quad \kappa_{min} \stackrel{\triangle}{=} \frac{(C^2 - 1)K_V}{D^2}, \quad s^*_{min} \stackrel{\triangle}{=} \frac{1}{2}\frac{D^2}{C^2 K_V}. \tag{IV.144}$$

Hence, for $|C| > 1$, *then* $\kappa_{min} = \frac{(C^2-1)K_V}{D^2}$ *is the threshold on power that ensures a strictly positive rate is feasible, i.e., exists. Note that at* $\kappa = \kappa_{min}$, *then* $K_Z^* = 0$, *hence the rate is zero, i.e.,* $\frac{1}{2}\log\frac{D^2\ K_Z^* + K_V}{K_V} = 0$, *and* κ_{min} *is precisely the minimum cost incurred of the problem of controlling the channel output process, when the rate is zero.*

The feedback capacity is obtained by substituting the optimal values (Γ^*, K_Z^*) *into (IV.138) to deduce the following expression.*

$$C_{A^\infty \to B^\infty}^{G-B.1}(\kappa) = \begin{cases} \frac{1}{2}\ln\frac{D^2\,\kappa + K_V}{K_V} & \text{if } |C| < 1, i.e.,\ K_Z^* = \kappa \\ \frac{1}{2}\ln\frac{D^2 K_Z^* + K_V}{K_V} & \text{if } |C| > 1, \kappa \in [\kappa_{min}, \infty) \end{cases} \tag{IV.145}$$

If $C = 1$ *then* $\Gamma^* = 0$ *and* $\left(C + D\Gamma^*\right) = 1 \notin \mathbb{D}_o$, *hence* $C_{A^\infty \to B^\infty}^{G-B.1}(\kappa)$ *does not exists. To cover this case, one needs to take* $Q > 0$.

If $|C| > 1$ *and* $\kappa \in [0, \kappa_{min})$ *then* $C_{A^\infty \to B^\infty}^{G-B.1}(\kappa)$ *does not exists. This is precisely the feedback capacity obtained in (I.37), using the solutions of the Riccati equation.*

Let $C_{A^\infty \to B^\infty}^{Stable}(\kappa)$ *denote the feedback capacity if the channel is stable, i.e.,* $|C| < 1$ *and* $C_{A^\infty \to B^\infty}^{Unstable}(\kappa)$ *denote the feedback capacity if the channel is unstable, i.e.,* $|C| > 1$. *Then, the corresponding feedback capacity is given by the following expressions.*

For $|C| < 1$*:*

$$C_{A^\infty \to B^\infty}^{G-B.1}(\kappa) \stackrel{\triangle}{=} C_{A^\infty \to B^\infty}^{Stable}(\kappa) = \frac{1}{2}\log\left(1 + \frac{\kappa}{K_V}\right), \quad \kappa \in [0, \infty) \tag{IV.146}$$

For $|C| > 1$*:*

$$C_{A^\infty \to B^\infty}^{G-B.1}(\kappa) \stackrel{\triangle}{=} C_{A^\infty \to B^\infty}^{Unstable}(\kappa) = \begin{cases} \frac{1}{2}\log\left(1 + \frac{\kappa}{K_V}\right) - \log|C| & \text{if } \kappa \in [\kappa_{min}, \infty) \\ \text{does not exists} & \text{if } \kappa \in [0, \kappa_{min}). \end{cases} \tag{IV.147}$$

Then it is clear from (IV.146) *and* (IV.147)*, that*

$$C_{A^\infty \to B^\infty}^{Unstable}(\kappa) = C_{A^\infty \to B^\infty}^{Stable}(\kappa) - \log|C|, \qquad \kappa \in [\kappa_{min}, \infty). \tag{IV.148}$$

Therefore, the rate loss due to the instability of the channel is given by

Rate Loss of Unstable Channels

$$\stackrel{\triangle}{=} C_{A^\infty \to B^\infty}^{Stable}(\kappa) - C_{A^\infty \to B^\infty}^{Unstable}(\kappa) = \log|C|, \quad \kappa \in [\kappa_{min}, \infty). \tag{IV.149}$$

In view of the above expressions, for unstable channels, there is rate a loss, expressed in terms of the logarithm of the unstable eigenvalue of the channel.

The above example illustrates the direct connection to linear stochastic systems and stability theory via Lyapunov equations. The general MIMO G-CM-B.1 is addressed in Section IV-C, by invoking dynamic programming.

C. Characterization of FTFI Capacity of G-CM-B.1 and the LQG Theory of Directed Information

The objective of this section is to completely solve the extremum problem corresponding to the characterization of FTFI capacity of the G-CM-B.1, and to gain insight on how to solve more general versions, such as, the G-CM-B (i.e., when the channel distribution depends on arbitrary memory), and the G-CM-A. This is done by re-formulating such extremum problems, using LQG stochastic optimal control theory, with randomized strategies (instead of deterministic as in the standard LQG theory [4, Ch. 6] or [5, Ch. 5]. Via this re-formulation, the optimal deterministic part of the randomized strategy, $\{\Gamma^*_{i,i-1} : i = 0, \ldots, n\}$, is found explicitly, in terms of solutions of Riccati matrix difference equations, while the random part $\{K^*_{Z_i} : i = 0, \ldots, n\}$, is determined from a water filling optimization problem, similar to that of MIMO memoryless channels [37].

The subsequent methodology is based the following simple observations.

(i) Define the randomized strategy of the equivalent characterization of FTFI capacity given by (IV.123)-(IV.125), as follows.

$$A_i^g \stackrel{\triangle}{=} U_i^g + Z_i, \quad i = 0, \ldots, n \tag{IV.150}$$

$$U_i^g \stackrel{\triangle}{=} g_i^{B.1}(B_{i-1}^g) \equiv \Gamma_{i,i-1} B_{i-1}^g \tag{IV.151}$$

where $\{U_i^g : i = 0, \ldots, n\}$ is the deterministic part of the strategy and $\{Z_i : i = 0, \ldots, n\}$ its random part. Then $\{U_i^g : i = 0, \ldots, n\}$ is the control process, chosen to control the channel output process $\{B_i^g : i = 0, \ldots, n\}$, and $\{Z_i : i = 0, \ldots, n\}$ is the innovations process, chosen to transmit new information over the channel.

(ii) Apply dynamic programming to show a separation principle and to determine recursively the optimal deterministic strategy $\{g_i^{B.1,*}(\cdot) : i = 0, \ldots, n\}$ and the optimal randomized process $\{Z_i : i = 0, \ldots, n\}$ (i.e., $\{K^*_{Z_i} : i = 0, \ldots, n\}$), from which the optimal solution $\left\{(\Gamma^*_{i,i-1}, K^*_{Z_i}) : i = 0, \ldots, n\right\}$, can be constructed.

Indeed, this methodology unfolds all consequences and the role of the control process $\{U_i^g : i = 0, \ldots, n\}$ to affect the controlled process $\{B_i^g : i = 0, \ldots, \}$, for the extremum problem of FTFI capacity characterization, and its per unit time limiting version, the feedback capacity.

The next theorem establishes the separation principle and the direct connection between LQG stochastic optimal control theory and the characterization of FTFI capacity, for MIMO G-CM-B.1.

Theorem 14 (Optimal Strategies of FTFI Capacity of G-CM-B.1): Consider the G-CM-B.1 defined by (IV.118), (IV.119), under Assumptions B, and consider $\kappa \in [\kappa_{min}, \infty)$.

(a) Define

$$A_i^g \stackrel{\triangle}{=} U_i^g + Z_i, \quad U_i^g = g_i^{B.1}(B_{i-1}^g) \equiv \Gamma_{i,i-1} B_{i-1}^g, \quad i = 0, \ldots, n \tag{IV.152}$$

where $\{U_i^g : i = 0, \ldots, n\}$ *is the deterministic part of the randomized strategy (control part) and* $\{Z_i : i = 0, \ldots, n\}$ *is the random part. Then*

$$B_i^g = C_{i,i-1} B_{i-1}^g + D_{i,i} U_i^g + D_{i,i} Z_i + V_i, \quad i = 0, \ldots, n, \quad B_{i-1}^g = b_{-1} \tag{IV.153}$$

and the equivalent characterization of the FTFI capacity is given by

$$C_{A^n \to B^n}^{G-B.1}(\kappa) = C_{A^n \to B^n}^{IL-G-B.1}(\kappa) = \sup_{\left\{(g_i^{B.1}(\cdot), K_{Z_i}), i=0,\ldots,n\right\} \in \mathscr{C}_{[0,n]}^{B.1}(\kappa)} \sum_{i=0}^{n} H(B_i^g | B_{i-1}^g) - H(V^n) \tag{IV.154}$$

where

$$\sum_{i=0}^{n} H(B_i^g | B_{i-1}^g) - H(V^n) = (IV.132), \tag{IV.155}$$

$$\mathscr{C}_{[0,n]}^{B.1}(\kappa) \stackrel{\triangle}{=} \Big\{ g_i^{B.1} : \mathbb{R}^p \to \mathbb{R}^q, u_i = g_i^{B.1}(b_{i-1}),$$

$$K_{Z_i} \in \mathbb{S}_+^{q \times q},\ i = 0, \ldots, n : \frac{1}{n+1} \mathbf{E}^{g^{B.1}} \Big(\sum_{i=0}^{n} \Big[\langle A_i^g, R_{i,i} A_i^g \rangle + \langle B_{i-1}^g, Q_{i,i-1} B_{i-1}^g \rangle \Big] \Big) \leq \kappa \Big\}. \tag{IV.156}$$

For the rest of the statements assume there exist an $\big\{(B_i^g, g_i^{B.1}(\cdot), Z_i) : i = 0, \ldots, \big\}$ *in the Hilbert space of square summable sequences, such that the feasible set in (IV.156) has an interior point (convexity of pay-off functional and constraint set can be shown).*

(b) The cost-to-go $C_i^{B.1} : \mathbb{R}^p \longmapsto \mathbb{R}$ *(corresponding to (IV.154)), from time "i" to the terminal time "n" for a fixed the value of the output* $B_{i-1}^g = b_{i-1}$ *is defined by*

$$C_i^{B.1}(b_{i-1}) \stackrel{\triangle}{=} \sup_{\left\{(U_j^g, K_{Z_j}) \in \mathbb{R}^q \times \mathbb{S}_+^{q \times q}, U_j^g = g_j^{B.1}(B_j^g),\ j=i,\ldots,n\right\}} \Bigg\{ \frac{1}{2} \sum_{j=i}^{n} \log \frac{|D_{j,j} K_{Z_j} D_{j,j}^T + K_{V_j}|}{|K_{V_j}|} - \sum_{j=i}^{n} tr\Big(s R_{j,j} K_{Z_j}\Big) + s(n+1)\kappa - s\mathbf{E}^{g^{B.1}} \Big\{ \sum_{j=i}^{n} \Big[\langle U_j^g, R_{j,j} U_j^g \rangle + \langle B_{j-1}^g, Q_{j,j-1} B_{j-1}^g \rangle \Big] \Big| B_{i-1}^g = b_{i-1} \Big\} \Bigg\} \tag{IV.157}$$

where $s \geq 0$ *is the Lagrange multiplier associated with the average transmission cost constraint (IV.156).*

(c) The dynamic programming recursions are given by the following equations.

$$C_n^{B.1}(b_{n-1}) = \sup_{(u_n, K_{Z_n}) \in \mathbb{R}^q \times \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{n,n} K_{Z_n} D_{n,n}^T + K_{V_n}|}{|K_{V_n}|} - tr\Big(s R_{n,n} K_{Z_n}\Big) + s(n+1)\kappa - s\Big[\langle u_n, R_{n,n} u_n \rangle + \langle b_{n-1}, Q_{n,n-1} b_{n-1} \rangle \Big] \Big\}, \tag{IV.158}$$

$$C_i^{B.1}(b_{i-1}) = \sup_{(u_i, K_{Z_i}) \in \mathbb{R}^q \times \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(s R_{i,i} K_{Z_i}\Big) - s\Big[\langle u_i, R_{i,i} u_i \rangle + \langle b_{i-1}, Q_{i,i-1} b_{i-1} \rangle \Big] + \mathbf{E}^{g^{B.1}} \Big\{ C_{i+1}^{B.1}(B_i^g) \Big| B_{i-1}^g = b_{i-1} \Big\} \Big\},\ i = 0, \ldots, n-1. \tag{IV.159}$$

(d) Separation Principle. The optimal deterministic part of the randomized strategy, $\{g_i^{B.1,*}(\cdot) : i = 0, \ldots, n\}$ *is independent of* $\{K_{Z_i} : i = 0, \ldots, n\}$, *and the corresponding covariance* $K_{B_i} \stackrel{\triangle}{=} \mathbf{E}\big\{B_i^g \big(B_i^g\big)^T\big\}, i = 0, 1, \ldots, n$, *are given by the following equations.*

$$g_i^{B.1,*} : \mathbb{R}^p \to \mathbb{R}^q, \quad i = 0, \ldots, n,$$
$$\Gamma^* : \{0, 1, \ldots, n\} \to \mathbb{R}^{q \times p}, \ P : \{0, 1, \ldots, n\} \to \mathbb{S}_+^{p \times p}, \tag{IV.160}$$

$$g_i^{B.1,*}(b_{i-1}) = F^*(i)\, b_{i-1} \equiv \Gamma_{i,i-1}^* b_{i-1}, \quad i = 0, \ldots, n, \tag{IV.161}$$

$$F^*(n) = \Gamma_{n,n-1}^* = 0, \quad F^*(i) = -H_{22}^{-1}(i) H_{12}^T(i), \tag{IV.162}$$

$$H_{11}(i) = C_{i,i-1}^T P(i+1) C_{i,i-1} + Q_{i,i-1},$$
$$H_{12}(i) = C_{i,i-1}^T P(i+1) D_{i,i},$$
$$H_{22}(i) = D_{i,i}^T P(i+1) D_{i,i} + R_{i,i},$$
$$P(i) = H_{11}(i) - H_{12}(i) H_{22}^{-1}(i) H_{12}^T(i), \quad i = 0, \ldots, n-1, \tag{IV.163}$$

$$= C_{i,i-1}^T P(i+1) C_{i,i-1} + Q_{i,i-1} - C_{i,i-1}^T P(i+1) D_{i,i} \cdot \Big(D_{i,i}^T P(i+1) D_{i,i} + R_{i,i} \Big)^{-1} \Big(C_{i,i-1}^T P(i+1) D_{i,i} \Big)^T \quad \text{(IV.164)}$$

$$P(n) = Q_{n,n-1}, \quad \text{(IV.165)}$$

$$K_{B_i} = \Big(C_{i,i-1} + D_{i,i} \Gamma_{i,i-1}^* \Big) K_{B_{i-1}} \Big(C_{i,i-1} + D_{i,i} \Gamma_{i,i-1}^* \Big)^T + D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}, \quad i = 0, \ldots, n, \quad \text{(IV.166)}$$

$$K_{B_{-1}} = Given. \quad \text{(IV.167)}$$

(e) The solution of the dynamic programming equations is given by the following equations.

$$C_i^{B.1}(b_{i-1}) = -s \langle b_{i-1}, P(i) b_{i-1} \rangle + r(i), \quad i = 0, \ldots, n \quad \text{(IV.168)}$$

where $\{P(i) : i = 0, \ldots, n\}$ *satisfies the backward recursive Riccati equation (IV.164), (IV.165), and the process* $\{r(i) : i = 0, \ldots, n\}$ *satisfies the backward recursion*

$$r(i) = r(i+1) + \sup_{K_{Z_i} \in \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(s P(i+1) \Big[D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i} \Big] \Big) - tr\Big(s \ R_{i,i} K_{Z_i} \Big) \Big\}, \quad i = 0, \ldots, n-1, \quad \text{(IV.169)}$$

$$r(n) = \sup_{K_{Z_n} \in \mathbb{S}_+^{q \times q}} \Big\{ \frac{1}{2} \log \frac{|D_{n,n} K_{Z_n} D_{n,n}^T + K_{V_n}|}{|K_{V_n}|} + s(n+1)\kappa - tr\Big(s R_{n,n} K_{Z_n} \Big) \Big\} \quad \text{(IV.170)}$$

or equivalently

$$r(0) = \sup_{K_{Z_i} \in \mathbb{S}_+^{q \times q} : i = 0, \ldots, n} \Big\{ \sum_{i=0}^{n-1} \Big[\frac{1}{2} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(s \ R_{i,i} K_{Z_i} \Big) - tr\Big(s P(i+1) \Big[D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i} \Big] \Big) \Big] + \frac{1}{2} \log \frac{|D_{n,n} K_{Z_n} D_{n,n}^T + K_{V_n}|}{|K_{V_n}|} + s(n+1)\kappa - tr\Big(s R_{n,n} K_{Z_n} \Big) \Big\} \quad \text{(IV.171)}$$

where s *is the Lagrange multiplier.*

Moreover, the optimal deterministic part of the randomized strategy is given by

$$g_i^{B.1,*}(b_{i-1}) = -\Big(D_{i,i}^T P(i+1) D_{i,i} + R_{i,i} \Big)^{-1} \times D_{i,i}^T P(i+1) C_{i,i-1} \, b_{i-1} \equiv \Gamma_{i,i-1}^* b_{i-1}, \quad i = 0, \ldots, n-1, \quad \text{(IV.172)}$$

$$g_n^{B.1,*}(b_{n-1}) = 0. \quad \text{(IV.173)}$$

(f) The optimal covariance (the random part of the randomized strategy) $\{K_{Z_i}^* : i = 0, \ldots, n\}$ *and* $s^* \equiv s^*(\kappa) \geq 0$ *are found from the problem*

$$\sup_{s \geq 0} \Big\{ - s \langle b_{-1}, P(0) b_{-1} \rangle + r(0) \Big\} \quad \textit{subject to (IV.171).} \quad \text{(IV.174)}$$

(g) The characterization of FTFI capacity (for any $s \geq 0$ *corresponding to* κ*) is given by*

$$C_{A^n \to B^n}^{G-B.1}(\kappa) = -s \int_{\mathbb{R}^p} \langle b_{-1}, P(0) b_{-1} \rangle \mathbf{P}_{B_{-1}}(db_{-1}) + r(0). \quad \text{(IV.175)}$$

(h) κ_{min} *is given by*

$$\frac{1}{n+1} \mathbf{E}^{g^{B.1,*}} \Big\{ \sum_{i=0}^{n} \Big[\langle A_i^g, R_{i,i} A_i^g \rangle + \langle B_{i-1}^g, Q_{i,i-1} B_{i-1}^g \rangle \Big] \Big\} \Big|_{K_Z = 0} \quad \text{(IV.176)}$$

that is, it is the solution of the LQG stochastic optimal control problem with $K_{Z_i} = 0, i = 0, \ldots, n$.

Proof: See Appendix D. □

The closed form expressions given in Theorem 14, for the G-CM.B.1 is attributed to the decomposition of the randomized strategies (IV.150), where the innovation process is an orthogonal process, which then implies the separation principle can be established. It appears this orthogonal decomposition and separation principle are vital and should be incorporated in other extremum problems of feedback capacity, such as, the Cover and Pombra [2] characterization of FTFI capacity given by (I.9) or any of its variants [3], [7], [9]. These points are further elaborated below.

Remark 15 (Connections to LQG Stochastic Optimal Control Theory): Theorem 14 illustrates the separation principle and the dual role of the randomized strategies (IV.152) in extremum problems of directed information. Specifically, the optimal deterministic part (IV.172), (IV.173) controls the channel output process, precisely as in LQG stochastic optimal control theory [4]. However, its optimal random part $\{Z_i : i = 0, \ldots, n\}$ *that is found from (IV.171), ensures an optimal innovations process with covariance* $\{K_{Z_i}^* : i = 0, \ldots, n\}$ *is transmitted over the channel, to achieve the characterization of FTFI capacity, and to meet the average transmission cost constraint.*

(a) The separation principle of Theorem 14 is derived via dynamic programming. However, this is not the only choice, as it is demonstrated below. Moreover, the main reasons which lead to the separation principle are (i) the decomposition of the channel input process $A_i^g = U_i^g + Z_i, i = 0, \ldots, n$ *given by (IV.152) and (ii) the independence of the directed information pay-off* $I(A^{g,n} \to B^{g,n}) = \frac{1}{2} \sum_{j=0}^{n} \log \frac{|D_{j,j} K_{Z_j} D_{j,j}^T + K_{V_j}|}{|K_{V_j}|}$ *on the predictable process* $\{U_i^g : i = 0, \ldots, n\}$*. To illustrate this, consider the optimization problem of Theorem 14.*

$$\sup_{\{(U_i^g, K_{Z_i}), i = 0, \ldots, n\} \in \mathcal{C}_{[0,n]}^{B.1}(\kappa)} \frac{1}{2} \sum_{j=0}^{n} \log \frac{|D_{j,j} K_{Z_j} D_{j,j}^T + K_{V_j}|}{|K_{V_j}|}. \quad \text{(IV.177)}$$

The Lagrangian of the unconstraint optimization problem is given by

$$\begin{aligned}
&\mathscr{L}^s_{0,n}(K_{Z_i}, U_i^g : i=0,\ldots,n)\\
&\quad\stackrel{\triangle}{=} \frac{1}{2}\sum_{i=0}^{n}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|}\\
&\qquad - s\Big(\mathbf{E}^{g^{B.1}}\Big\{\sum_{i=0}^{n}\Big[\langle A_i^g, R_{i,i}A_i^g\rangle + \langle B_{i-1}^g, Q_{i,i-1}B_{i-1}^g\rangle\Big]\Big\}\\
&\qquad - \kappa(n+1)\Big), \quad s\geq 0 \qquad \text{(IV.178)}\\
&\quad\stackrel{(a)}{=} \frac{1}{2}\sum_{i=0}^{n}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - s\sum_{i=0}^{n}Tr\Big(R_{i,i}K_{Z_i}\Big)\\
&\qquad - s\Big(\mathbf{E}^{g^{B.1}}\Big\{\sum_{i=0}^{n}\Big[\langle U_i^g, R_{i,i}U_i^g\rangle + \langle B_{i-1}^g, Q_{i,i-1}B_{i-1}^g\rangle\Big]\Big\}\\
&\qquad - \kappa(n+1)\Big) \qquad \text{(IV.179)}
\end{aligned}$$

where $\{B_i^g : i=0,\ldots,n\}$ *satisfies (IV.153), and* (a) *is due to the orthogonality of* U_i^g *and* Z_i. *Clearly, for a fixed* s, *the Lagrangian* $\mathscr{L}^s_{0,n}(K_{Z_i}, U_i^g : i=0,\ldots,n)$ *needs to be maximized over* $(U_i^g, K_{Z_i}), i=0,\ldots,n$.

But, for a fixed $\{K_{Z_i} : i=0,\ldots,n\}$, *the maximization of* $\mathscr{L}^s_{0,n}(K_{Z_i}, U_i^g : i=0,\ldots,n)$ *over* $\{U_i^g : i=0,\ldots,n\}$ *is equivalent to the following problem.*

$$\begin{aligned}
&J_{0,n}(U^{g,*})\\
&\quad\stackrel{\triangle}{=}\inf_{U_i^g : i=0,\ldots,n}\mathbf{E}^{g^{B.1}}\Big\{\sum_{j=0}^{n}\Big[\langle U_j^g, R_{j,j}U_j^g\rangle\\
&\qquad + \langle B_{j-1}^g, Q_{j,j-1}B_{j-1}^g\rangle\Big]\Big\}, \qquad \text{(IV.180)}\\
&\qquad \text{subject to } \{B_i^g : i=0,\ldots,n\} \text{ satisfying (IV.153).} \qquad \text{(IV.181)}
\end{aligned}$$

Note that optimization problem $J_{0,n}(U_i^{g,*} : i=0,\ldots,n)$ *is a classical LQG stochastic optimal control problem with complete information. Hence, its solution can be derived using*

1) the completion of squares method, or

2) the stochastic Pontryagin's maximum principle method, or

3) dynamic programming method.

The important observation is that, the optimal strategy of a classical LQG stochastic optimal control problem with complete information, and hence that of $J_{0,n}(U^{g,*})$, *is independent of the noise processes* $\{(K_{Z_i}, K_{V_i}) : i=0,\ldots,n\}$ *which drive the dynamics of the process* $\{B_i^g : i=0,\ldots,n\}$. *Hence, the strategy of the optimal process* $U_i^{g,*} = g_i^{B.1,*}(B_{i-1}^{g,*}), i=0,\ldots,n$ *is independent of the innovations process* $\{K_{Z_i} : i=0,\ldots,n\}$, *and it is given as stated in Theorem 14. Thus, the fact that separation holds, is not related to the method of dynamic programming applied to find the optimal solution.*

Moreover, by the above discussion or Theorem 14, it follows that the optimal pay-off $J_{0,n}(U^{g,*})$, *for fixed* $B_{-1}=b_{-1}$ *is given by*

$$\begin{aligned}
&J_{0,n}(U^{g,*}, b_{-1})\\
&\quad=\sum_{i=0}^{n-1}tr\Big(P(i+1)\big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\big] + R_{i,i}K_{Z_i}\Big)\\
&\qquad + tr\Big(R_{n,n}K_{Z_n}\Big) + \langle b_{-1}, P(0)b_{-1}\rangle \qquad \text{(IV.182)}
\end{aligned}$$

Hence, the characterization of FTFI capacity for fixed $B_{-1}=b_{-1}$ *is given by*

$$\begin{aligned}
C_0^{B.1}(b_{-1}) = \sup_{\substack{K_{Z_i}\geq 0,\ i=0,\ldots,n:\\ J_{0,n}(U^{g,*},b_{-1})\leq(n+1)\kappa}}\Big\{&\\
\frac{1}{2}\sum_{j=0}^{n}\log\frac{|D_{j,j}K_{Z_j}D_{j,j}^T + K_{V_j}|}{|K_{V_j}|}\Big\}.& \qquad \text{(IV.183)}
\end{aligned}$$

Clearly, (IV.171) is the unconstraint optimization problem corresponding to (IV.183), where s *is the Lagrange multiplier associated with the average cost constraint. It is easy to verify this is a convex optimization problem.*

(b) From (IV.152)-(IV.154) it follows directly, as expected, that

$$\text{if}\quad K_{Z_i}^*=0,\quad i=0,\ldots,n \text{ then } C_{A^n\to B^n}^{G-B.1}(\kappa)=0. \qquad \text{(IV.184)}$$

Hence, the FTFI capacity is zero and consequently, its per unit time limit the feedback capacity is zero, although the output process can be stabilized (under appropriate conditions, given in Theorem 19). This re-confirms and strengthens the following well-known fact of LQG stochastic optimal control or decision theory. Among all (not necessarily Markov) randomized strategy $\pi^{RS}\stackrel{\triangle}{=}\Big\{\mathbf{P}_{A_i|A^{i-1},B^{i-1}} : i=0,\ldots,n\Big\}$, *the optimal strategy of the LQG stochastic optimal control problem*

$$\begin{aligned}
&J_{0,n}(\pi^{RS,*})\\
&\quad\stackrel{\triangle}{=}\inf_{\mathbf{P}_{A_i|A^{i-1},B^{i-1}}:i=0,\ldots,n}\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}\Big(\langle A_i, R_{i,i}A_i\rangle\\
&\qquad + \langle B_{i-1}, Q_{i,i-1}B_{i-1}\rangle\Big)\Big\}, \qquad \text{(IV.185)}\\
&\quad\text{subject to}\\
&\qquad B_i = C_{i,i-1}\ B_{i-1} + D_{i,i}\ A_i + V_i,\quad B_{-1}=b_{-1},\\
&\qquad\qquad i=0,\ldots,n \qquad \text{(IV.186)}
\end{aligned}$$

is Gaussian and Markov of the form $A_i = g_i^M(B_{i-1}) + Z_i, Z_i\sim N(0,K_{Z_i}), Z_i\perp B^{i-1}, i=0,\ldots,n$, $\{Z_i : i=0,\ldots,n\}$ *an orthogonal process, and occurs in the subclass of nonrandom or deterministic policies*

$$\Big\{\big(g_i^M(b_{i-1}), Z_i\big) = \big(g_i^{B.1,*}(b_{i-1}), 0\big) : i=0,\ldots,n\Big\} \qquad \text{(IV.187)}$$

i.e., $\mathbf{P}^*_{A_i|A^{i-1},B^{i-1}} = \mathbf{P}^*_{A_i|B_{i-1}} = \delta_{A_i}(g_i^{B.1,*}(b_{i-1}))$, *is a delta measure concentrated at* $g_i^{B.1,*}(\cdot), i=0,\ldots,n$). *In view of (IV.184), then* $J_{0,n}(\pi^{RS,*})$ *is the minimum cost,* $\kappa_{min}\in[0,\infty)$ *for existence of the solution to the FTFI capacity, and for any* $\kappa\in(\kappa_{min},\infty)$ *its value is strictly positive.*

It should be mentioned that the above observations imply that the FTFI capacity is attractive for designing controllers,

which stabilize controlled dynamical systems, and ensure information transfer or signalling from, say, the control process to the controlled process, or from one decision maker to another.

(b) The optimal random part of the strategy is found from (IV.170), that depends on the solution of a Riccati difference equation.

(c) The extremum solution illustrates a separation between the role of control (deterministic part of the strategy) and the role of information transmission (random part of the strategy).

(d) The material discussed in Section I-C, regarding the G-CM-B.1, given by (I.20)-(I.40), that relate feedback capacity, capacity without feedback and LQG stochastic optimal control theory, are direct consequences of the above theorem, specifically, the per unit time limiting version of Theorem 14, which is investigated in Section V.

The solution of the information transmission optimization problem is presented in the next remark for the scalar case.

Example 16 (Solution of Information Transmission Problem): From (IV.171) or the equivalent formulation (IV.183), it follows that both optimization problems are convex, that is, both pay-offs in (IV.171) or (IV.183) are concave with respect to randomized part of the strategy $\{K_{Z_i} : i = 0, \ldots, n\}$. *Also, by (IV.183) then* $C_0^{B.1}(b_{-1})$ *is nondecreasing and concave in* $\kappa \in (\kappa_{min}, \infty)$, *and hence continuous on this interval, and also zero at* $\kappa = \kappa_{min}$. *These imply* $C_0^{B.1}(b_{-1})$ *is strictly increasing in* $\kappa \in (\kappa_{min}, \infty)$, *and the constraint is satisfied with equality. In fact, (IV.171) is the unconstraint version of (IV.183), where* s *is the Lagrange multiplier. Hence, the optimal strategy* $\{K^*_{Z_i} \succeq 0 : i = 0, \ldots, n\}$ *is found by using the Kuhn-Tucker conditions as follows. Write the information rate as a function of the power allocated to the optimal strategy at each time instant as follows.*

$$C_0^{B.1}(b_{-1}) = C_{0,n}^{b_{-1}}(\kappa_0^*, \ldots, \kappa_n^*) \stackrel{\triangle}{=} \sum_{i=0}^{n} C_i^{b_{-1}}(\kappa_i^*) \tag{IV.188}$$

$$\equiv \sup_{K_{Z_i} \succeq 0, i=0,\ldots,n: \sum_{i=0}^{n} \kappa_i(K_{Z_i}) = \kappa(n+1)} \sum_{i=0}^{n} C_i^{b_{-1}}(\kappa_i) \tag{IV.189}$$

where

$$C_i^{b_{-1}}(\kappa_i) \stackrel{\triangle}{=} \frac{1}{2} \log \frac{|D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|}, \quad i = 0, \ldots, n, \tag{IV.190}$$

$$\kappa_i \equiv \kappa_i(K_{Z_i}) \stackrel{\triangle}{=} \begin{cases} tr\Big(R_{n,n} K_{Z_n}\Big), \ i = n \\ tr\Big(P(i+1)\big[D_{i,i} K_{Z_i} D_{i,i}^T + K_{V_i}\big] + R_{i,i} K_{Z_i}\Big), \\ \quad i = 1, \ldots, n-1 \\ tr\Big(P(1)\big[D_{0,0} K_{Z_0} D_{0,0}^T + K_{V_0}\big] + R_{0,0} K_{Z_0}\Big) \\ \quad + \langle b_{-1}, P(0) b_{-1} \rangle, \ i = 0. \end{cases} \tag{IV.191}$$

Note that as expected, if $K_{Z_i} = 0 : i = 0, \ldots, n$, *then*

$$\kappa_{0,n}^{b_{-1}}(0) \stackrel{\triangle}{=} J_{0,n}(U^{g,*}, b_{-1})\Big|_{K_{Z_i}=0, i=0,\ldots,n} = \sum_{i=0}^{n-1} tr\Big(P(i+1) K_{V_i}\Big) + \langle b_{-1}, P(0) b_{-1} \rangle \tag{IV.192}$$

which is the optimal pay-off of the LQG problem (IV.185) or (IV.182) with deterministic strategies.

1) Special Case $p = q = 1$: *Since* $K^*_{Z_i}$ *must be nonnegative, by invoking the Kuhn-Tucker conditions, after some algebra the following are obtained.*

$$K^*_{Z_n} = \Big\{\frac{1}{2s R_{n,n}} - \frac{K_{V_n}}{D_{n,n}^2}\Big\}^+, \quad \{x\}^+ \stackrel{\triangle}{=} \max\{0, x\} \tag{IV.193}$$

$$K^*_{Z_i} = \Big\{\frac{1}{2s\Big(P(i+1) D_{i,i}^2 + R_{i,i}\Big)} - \frac{K_{V_i}}{D_{i,i}^2}\Big\}^+, \quad i = n-1, n-2, \ldots, 0 \tag{IV.194}$$

where $s = s_n(\kappa, b_{-1}) \geq 0$ *is chosen to satisfy the average constraint with equality given by*

$$\sum_{i=0}^{n-1} \Big\{\Big\{\frac{1}{2s} - \frac{\Big(P(i+1) D_{i,i}^2 + R_{i,i}\Big) K_{V_i}}{D_{i,i}^2}\Big\}^+ + P(i+1) K_{V_i}\Big\} + \Big\{\frac{1}{2s} - \frac{R_{n,n} K_{V_n}}{D_{n,n}^2}\Big\}^+ + b_{-1}^2 P(0) = \kappa(n+1). \tag{IV.195}$$

The characaterization of FTFI capacity for fixed $B_{-1} = b_{-1}$ *is given by*

$$\begin{aligned} C_{A^n \to B^n}^{G-B.1}(\kappa, b_{-1}) &\equiv C_0^{B.1}(b_{-1}) = \frac{1}{2} \sum_{i=0}^{n} \log \frac{|D_{i,i} K^*_{Z_i} D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} \\ &= \frac{1}{2} \sum_{i=0}^{n-1} \Big\{ \log\Big(\frac{D_i^2}{2s\Big(P(i+1) D_{i,i}^2 + R_{i,i}\Big) K_{V_i}}\Big)\Big\}^+ \\ &\quad + \frac{1}{2}\Big\{ \log\Big(\frac{D_{i,i}^2}{2s R_{n,n} K_{V_n}}\Big)\Big\}^+ \\ &= \sum_{i=0}^{n} C_i^{b_{-1}}(\kappa_i^*). \end{aligned} \tag{IV.196}$$

Over the time horizon, $\{0, 1, \ldots, n\}$, *for a given* κ *and initial state* $B_{-1} = b_{-1}$, *then the optimal level* $s = s_n(\kappa, b_{-1})$ *is found from (IV.195) and then substituted it into (IV.193), (IV.194) and (IV.196), to determine whether at each time* i, *the information rate* $C_i^{b_{-1}}(\kappa_i^*)$ *is either positive or zero, for* $i = 0, \ldots, n$. *Clearly, in general, for each* i, *then* $C_i^{b_{-1}}(\kappa_i^*) > 0$ *provided* $\kappa_i^* \in (\kappa_{min,i}, \infty)$ *and these critical values depend on whether the coefficients of the channel model are* $|C_{i,i-1}| \geq 1$ *or* $|C_{i,i-1}| < 1$, *for* $i = 0, \ldots, n$. *In principle, the general MIMO case is solved similarly, although it is much more involved because there is a water-filling both in time and dimension (spatial).*

Remark 17 (Relation to Cover and Pombra [2]):

(a) As pointed out in Remark 11, it is difficult to obtain

closed form solutions to the extremum problem of the Cover and Pombra [2] scalar AGN channel, without, re-visiting the derivation to obtain a realization of optimal channel input distribution, based on an orthogonal decomposition similar to (IV.150), and without showing an analogous separation principle. In fact, the only known explicit calculations of feedback capacity to the Cover and Pombra [2] scalar AGN channel, are the ones presented in Section I-B (as documented in [3], [7] and [9]), for scalar, stable and stationary Gaussian noise models, such as, the AR(1) model $|\alpha| < 1$ *[3], [7], [9]. The unstable AR(1) model is not addressed [3], [7], [9], hence it remains to be determined whether the cost* $\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}|A_i|^2\Big\} \leq \kappa$ *is sufficient to ensure asymptotic ergodicity. Moreover, the tools applied in [3], [7] and [9] are based on Power Spectral densities, and hence these need to be re-visited, to be able to deal with non-stationary noise processes. Also the analysis in [3], [7] and [9] does not include a direct connection to mean-square estimation theory, which the analog (dual) of the LQG stochastic optimal control theory, for models presented in this paper. As illustrated for the scalar channel by (IV.145), whether feedback capacity exists, and whether feedback increases capacity, depends on the á priori assumptions imposed on the channel, and the type of transmission cost imposed. This point should be accounted for when analyzing feedback channels, with memory, especially for unstable channels or noise models.*

(b) For more general channels, which also depend on past channel inputs, a decomposition analogous to (IV.150) can be derived, which includes additional components. Some of these components can be obtained, independently of others, and hence a separation similar to the one obtained in Theorem 14, can be shown. In general, it is expected that extremum problems of FTFI capacity can be decomposed into sub-optimization problems, each associated with either determining one of the components of the optimal strategy, or determining two or more components which interact. However, the critical level κ_{min} *needs to be identified (see Example 13) to ensure a non-zero rate.*

Finally, it should be mentioned that for the Cover and Pombra [2] scalar AGN channel (I.9), with AR(1) noise model defined by (I.14), the statements in Kim [3, Th. 6 and Lemma 6.1] and [9, Corollary 7.1] do not include the analog of (IV.184).

(c) For non-Gaussian channels it remains to be determined whether a separation similar to the one obtained in Theorem 14, can be established.

D. Characterization of FTFI Capacity of G-CM-B and the LQG Theory

Consider the G-CM-B.J (a generalization of the G-CM-B.1), defined by

$$B_i = \sum_{j=1}^{M} C_{i,i-j}\,B_{i-j} + D_{i,i}\,A_i + V_i, \quad B_{-M}^{-1} = b_{-M}^{-1}, \quad i = 0,\ldots,n, \quad \text{(IV.197)}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{\sum_{i=0}^{n}\Big(\langle A_i, R_{i,i}A_i\rangle + \langle B_{i-K}^{i-1}, Q_K(i-1)B_{i-K}^{i-1}\rangle\Big)\Big\} \leq \kappa, \quad \text{(IV.198)}$$

$$J \stackrel{\triangle}{=} \max\{M,K\}, \quad R_{i,i} \in S_+^{q\times q}, \quad Q_K(-1) = 0, \quad Q_K(i-1) \in S_+^{Kp\times Kp}, \quad i = 0,\ldots,n, \quad \text{Assumption B holds.} \quad \text{(IV.199)}$$

It can be verified, by repeating the derivation of Theorem 10, if necessary, that the optimal channel input distribution is conditionally Gaussian of the form $\{\pi_i^g(da_i|b_{i-J}^{i-1}) : i = 0,\ldots,n\}$, with linear conditional mean and non-random conditional covariance, and that all material presented in Section IV-C, generalize to G-CM-B.J.

E. Characterization of FTFI Capacity of G-CM-A and the LQG Theory

Consider the G-CM-A, i.e., (IV.96), which is not of limited memory. By Theorem 10, if $\sigma\{B^{g,-1}\} = \{\emptyset,\Omega\}$ then

$$A_i^g \stackrel{\triangle}{=} U_i^g + Z_i, \quad A_0 = Z_0, \quad i = 1,\ldots,n \quad \text{(IV.200)}$$

$$U_i^g = g_i^A(B^{g,i-1}) \equiv \Gamma_i(i-1)B^{g,i-1} \quad \text{(IV.201)}$$

where $\{U_i^g : i = 0,\ldots,n\}$ is the deterministic part of the randomized strategy and $\{Z_i : i = 0,\ldots,n\}$ is the random part. Then

$$B_i^g = C_i(i-1)B^{g,i-1} + D_{i,i}U_i^g + D_{i,i}Z_i + V_i, \quad B_0^g = D_{0,0}A_0^g + V_0, \quad i = 1,\ldots,n. \quad \text{(IV.202)}$$

Clearly, the dimension of the process $\{S_i^g \stackrel{\triangle}{=} B^{g,i-1} : i = 0,\ldots,n\}$ increases with time $i = 0,1,\ldots,n$. The following is stated as a conjecture.

Conjecture 18: Based on (IV.200), (IV.202) the optimal randomized strategy of the G-CM-A defined by (IV.96) can be found by using the method of Remark 11 or by restricting attention to stationary ergodic processes and applying Cholesky decomposition, and power spectral densities.

V. Feedback Capacity of G-CM-B & The Infinite Horizon LQG Theory of Directed Information

In this section, the per unit time limiting version of G-CM-B is investigated, and the characterization of feedback capacity is derived, irrespectively of whether the eigenvalues of the channel matrix C, that is, $spec(C)$ lie in the open disc of the unit circle in $\mathbb{C}$. Specifically, the characterizations of FTFI capacity given in Section III are applied to Gaussian channel models (G-CMs) of Definition 5, to obtain the following.

(a) Characterizations of feedback capacity for multiple input multiple output (MIMO) G-CMs, via the per unit time limit of the characterizations of FTFI capacity of MIMO G-CMs;

(b) relations between infinite horizon LQG stochastic optimal control theory, linear stochastic feedback controlled systems, feedback capacity and capacity without feedback.

A. Feedback Capacity of G-CM-B.1 & Infinite Horizon LQG Theory

Consider first, the G-CM-B.1, i.e., (IV.118), (IV.119). The extension to the general model G-CM-B.J, can be treated as discussed in Section IV-C. The next theorem establishes a hidden connection between, infinite horizon per unit time LQG stochastic optimal control theory, directed information stability (see (VI.240), (VI.241)), and optimal transmission rates. Moreover, through the computation of the feedback capacity, a separation principle is established, between the role of deterministic part of the randomized strategy to stabilize unstable channels, and the role of its random part to transmit new information.

Theorem 19 (Feedback Capacity of TI-G-CM-B.1): Consider the time-invariant version of G-CM-B.1, i.e., (IV.118), (IV.119), under Assumption B, called TI-G-CM.B.1, defined by

$$B_i = C\ B_{i-1} + D\ A_i + V_i, \quad B_{-1} = b_{-1}, \\ K_{V_i} = K_V \in S_{++}^{p\times p}, \quad i = 0, \ldots, n, \tag{V.203}$$

$$\frac{1}{n+1}\mathbf{E}\Big\{ \sum_{i=0}^{n} \Big(\langle A_i, RA_i\rangle + \langle B_{i-1}, QB_{i-1}\rangle \Big) \Big\} \leq \kappa, \\ R \in \mathbb{S}_{++}^{q\times q},\ Q \in \mathbb{S}_{+}^{p\times p}. \tag{V.204}$$

for some $\kappa \in [\kappa_{min}, \infty)$.

Assume the following conditions hold (see Appendix for definitions and implications).

$$i)\ \text{the pair } (C, D) \text{ is stabilizable} \tag{V.205}$$

$$ii)\ \text{the pair } (G, C) \text{ is detectable, where } Q = G^T G, \\ G \in \mathbb{S}_{+}^{p\times p}. \tag{V.206}$$

Moreover, assume the set of channel input conditional distributions is restricted to time-invariant distributions, i.e., $\{\pi_i^g(da_i|b_{i-1}) = \pi^{g,\infty}(da_i|b_{i-1}) : i = 0, \ldots, n\}$.

Then the following hold.

(a) Define

$$A_i^g \stackrel{\triangle}{=} U_i^g + Z_i, \quad U_i^g = g^{B.1}(B_{i-1}^g) \equiv \Gamma B_{i-1}^g, \quad i = 0, \ldots, n \tag{V.207}$$

where $\{U_i^g : i = 0, \ldots, n\}$ is the deterministic part of the randomized strategy and $\{Z_i : i = 0, \ldots, n\}$ is its random part. Then

$$B_i^g = CB_{i-1}^g + DU_i^g + DZ_i + V_i, \quad i = 0, \ldots, n. \tag{V.208}$$

Define

$$C_{A^n\to B^n}^{B.1}(\kappa) \\ \stackrel{\triangle}{=} \sup_{\left\{(g^{B.1}(\cdot), K_Z), i=0,\ldots,n\right\} \in \mathscr{C}_{[0,n]}^{B.1}(\kappa)} \sum_{i=0}^{n} H(B_i^g|B_{i-1}^g) \\ - H(V^n), \tag{V.209}$$

$$\mathscr{C}_{[0,n]}^{B.1}(\kappa) \\ \stackrel{\triangle}{=} \Big\{ g^{B.1} : \mathbb{R}^p \longmapsto \mathbb{R}^q, \quad u_i = g^{B.1}(b_{i-1}), \\ K_Z \in \mathbb{S}_{+}^{q\times q}, \quad i = 0, \ldots, n : \\ \frac{1}{n+1}\mathbf{E}^{g^{B.1}}\Big\{ \sum_{i=0}^{n} \Big[\langle A_i^g, RA_i^g\rangle + \langle B_{i-1}^g, QB_{i-1}^g\rangle \Big] \Big\} \leq \kappa \Big\} \tag{V.210}$$

$$C_{A^\infty\to B^\infty}^{B.1}(\kappa) \\ \stackrel{\triangle}{=} \liminf_{n\longrightarrow\infty} \frac{1}{n+1} C_{A^n\to B^n}^{FB,B.1}(\kappa) \tag{V.211}$$

and assume there exist an $(\{B_i^g : i = 0, \ldots, \}, g^{B.1}(\cdot), K_Z)$ such that the feasible set in (V.210) has an interior point.

Then $C_{A^\infty\to B^\infty}^{B.1}(\kappa)$ is the per unit time version of the characterization of FTFI capacity corresponding to (IV.154), that is,

$$C_{A^\infty\to B^\infty}^{B.1}(\kappa) \\ = C_{A^\infty\to B^\infty}^{G-B.1}(\kappa) \stackrel{\triangle}{=} \liminf_{n\longrightarrow\infty} \frac{1}{n+1} C_{A^n\to B^n}^{G-B.1}(\kappa) \tag{V.212}$$

$$= C_{A^\infty\to B^\infty}^{IL-G-B.1}(\kappa) \stackrel{\triangle}{=} \liminf_{n\longrightarrow\infty} \frac{1}{n+1} C_{A^n\to B^n}^{IL-G-B.1}(\kappa) \tag{V.213}$$

where $C_{A^n\to B^n}^{G-B.1}(\kappa)$ is defined by (IV.154).

(b) The pair $\Big(J^{B.1,*}, C^{B.1}(b)\Big)$, $J^{B.1,*} \in \mathbb{R}$, $C^{B.1} : \mathbb{R}^p \to \mathbb{R}$ satisfies the following dynamic programming equation (corresponding to $C_{A^\infty\to B^\infty}^{B.1}(\kappa)$).

$$J^{B.1,*} + C^{B.1}(b) \\ = \sup_{(u,K_Z)\in\mathbb{R}^q\times\mathbb{S}_{+}^{q\times q}} \Big\{ \frac{1}{2}\log\frac{|DK_ZD^T + K_V|}{|K_V|} \\ - tr\Big(sRK_Z\Big) + s\kappa - s\Big[\langle u, Ru\rangle + \langle b, Qb\rangle\Big] \\ + \mathbf{E}^{g^{B.1}}\Big\{ C^{B.1}(B_i^g)\Big| B_{i-1}^g = b \Big\} \Big\} \tag{V.214}$$

where $s \geq 0$ is found from the average transmission cost constraint.

(c) The optimal stationary policy $g^{B.1,\infty,*}(\cdot)$ and corresponding covariance matrix K of $\{B_i^* : i = 0, \ldots, n\}$ are given by the following equations.

$$g^{B.1,*} : \mathbb{R}^p \to \mathbb{R}^q, \quad \Gamma \in \mathbb{R}^{q\times p}, \quad P \in \mathbb{S}_{+}^{p\times p}, \tag{V.215}$$

$$g^{B.1,*}(b) = \Gamma^* b, \tag{V.216}$$

$$\Gamma^* = -H_{22}^{-1}H_{12}^T = -\Big(D^TPD + R\Big)^{-1}D^TPC, \tag{V.217}$$

$$H_{11} = C^TPC + Q, \quad H_{12} = C^TPD, \quad H_{22} = D^TPD + R, \tag{V.218}$$

$$P = H_{11} - H_{12}H_{22}^{-1}H_{12}^T \tag{V.219}$$

$$P = C^TPC + Q - C^TPD\Big(D^TPD + R\Big)^{-1}\Big(C^TPD\Big)^T, \tag{V.220}$$

$$K = \Big(C + D\Gamma^*\Big)K\Big(C + D\Gamma^*\Big)^T + DK_ZD^T + K_V, \tag{V.221}$$

$$spec\Big(C + D\Gamma^*\Big) \\ = spec\Big(C - D\Big(D^TPD + R\Big)^{-1}D^TPC\Big) \subset \mathbb{D}_o. \tag{V.222}$$

(d) The solution of the dynamic programming equation is given by

$$C^{B.1}(b) = -s\langle b, Pb\rangle, \tag{V.223}$$

$$J^{B.1,*} = \sup_{K_Z \in \mathbb{S}_+^{q\times q}} \Big\{ \frac{1}{2}\log \frac{|DK_ZD^T + K_V|}{|K_V|} + s\kappa - tr\Big(s\ RK_Z\Big) - tr\Big(sP\Big[DK_ZD^T + K_V\Big]\Big)\Big\} \tag{V.224}$$

$$g^{B.1,*}(b) = -\Big(D^TPD + R\Big)^{-1} D^TPC\ b. \tag{V.225}$$

(e) The optimal covariance K_Z^* and $s \geq 0$ are found from the optimization problem

$$\inf_{s\geq 0} J^{B.1,*} \quad \text{where } P \text{ is the solution of (V.220).} \tag{V.226}$$

The average transmission cost constraint evaluated on the optimal strategy is given by

$$\mathbf{E}^{g^{B.1,*}}\Big\{\langle g^{B.1,*}(B^*), Rg^{B.1,*}(B^*)\rangle + \langle B^*, QB^*\rangle\Big\} + tr\Big(RK_Z^*\Big) \leq \kappa \tag{V.227}$$

where the expectation is with respect to the invariant distribution $\mathbf{P}_B^{g^{B.1,*}}(db)$ of the optimal output process $\{B_i^* : i = 0, \ldots, n\}$ corresponding to $(g^{B.1,*}(\cdot), K_Z^*)$.

(f) $J^{B.1,*}\Big|_{s=s^*} = C^{B.1}_{A^\infty \to B^\infty}(\kappa)$, where s^* is the Lagrange multiplier found from (V.226) or the average constraint via (V.227).

(g) The information density and the constraint evaluated at the optimal stationary strategy are information stable, (see (VI.240), (VI.241) for precise definition). Specifically, for any initial distribution $\mathbf{P}_{B_{-1}}(db_{-1}) = \mu(db_{-1}) \in \mathscr{M}(\mathbb{R}^p)$, the following hold.

$$C^{B.1}_{A^\infty \to B^\infty}(\kappa) \equiv J(\pi^{g,\infty,*}, \mu) = J^{B.1,*}\Big|_{s=s^*}, \quad \forall \mu(\cdot) \in \mathscr{M}(\mathbb{R}^p), \tag{V.228}$$

$$J^0(\pi^{g,\infty,*}, \mu) = J^{B.1,*}\Big|_{s=s^*}, \quad \mathbf{P}_\mu^{\pi^{g,\infty,*}} - a.s., \quad \forall \mu(\cdot) \in \mathscr{M}(\mathbb{R}^p) \tag{V.229}$$

where

$$J^0(\pi^{g,\infty,*}, \mu) \triangleq \sup_{\mathring{\mathscr{P}}^{\infty,B.1}_{[0,\infty]}(\kappa)} \liminf_{n\to\infty} \frac{1}{n+1}\Big\{ \sum_{i=0}^{n} \log\Big(\frac{dQ_i(\cdot|B_{i-1}, A_i)}{dv_i^{\pi^{g,\infty}}(\cdot|B_{i-1})}(B_i)\Big)\Big\}, \tag{V.230}$$

$$\mathring{\mathscr{P}}^{\infty,B.1}_{[0,\infty]}(\kappa) \triangleq \Big\{\pi^{g,\infty}(da_i|b_{i-1}), i = 0, 1, \ldots, n : \limsup_{n\to\infty} \frac{1}{n+1}\sum_{i=0}^{n}\Big(\langle A_i^g, RA_i^g\rangle + \langle B_{i-1}^g, QB_{i-1}^g\rangle\Big) \leq \kappa\Big\}. \tag{V.231}$$

Proof: (a) This follows as in Theorem 14. (b)-(e) By the stabilizability and detectability conditions, i), ii) the dynamic programming equation (V.214) holds (see [37, Ch. 8, Sec. 5] or [5, Ch. 6]). By repeating the derivation of Theorem 14, if necessary, (c)-(d) are obtained. (f), (g) These follow from the fact that $N(0, K_B)$ is the unique invariant Gaussian distribution of (V.208), corresponding to the stabilizing optimal policy (V.225) (i.e., (V.222) holds), the ergodic properties of LQG stochastic optimal control theory [4], [5], and asymptotic pathwise optimality of the per unit time limit of directed information density (by applying [38, Th. 5.7.9]). A derivation based on information stability is given in [1, Th. 4.1]. □

Remark 20 (Comments on Theorem 19):

(a) Theorem 19 gives sufficient conditions in terms of detectability and stabilizability, i.e., (V.205), (V.206), for existence of feedback capacity, irrespectively of the initial distribution of B_{-1} and whether the eigenvalues of channel matrix C are stable or unstable, that is, whether they lie in the open unit disc of complex numbers, $spec(C) \subset \mathbb{D}_o \triangleq \{c \in \mathbb{C} : |c| < 1\}$ or outside $spec(C) \subset \mathbb{D}_o^c \triangleq \{c \in \mathbb{C} : |c| \geq 1\}$. In fact, without any assumptions on stationarity and ergodicity, the above theorem demonstrates that feedback capacity (i.e., the supremum of all achievable rates) depends on the á priori assumptions on the channel model coefficients, $\{C, D, R, Q, K_V\}$, because these determine whether the conditions of stabilizability of the pair (C, D) and detectability of the pair (G, C), i.e., (V.205), (V.206) are satisfied.

(b) Whether feedback capacity exists at all, for nonstationary and nonergodic processes, is directly related to these conditions of stabilizability and detectability, and the structure of the matrix $Q \geq 0$ entering the transmission cost function, plays a significant role, on the uniqueness of nonnegative stabilizing solutions to algebraic Riccati equations, and the ability of the optimal feedback strategy $\{g^{B.1,*}(b_{i-1}) : i = 0, \ldots, \}$ given by (V.225) to stabilize even unstable channels, that is, when the eigenvalues of channel matrix C do not lie on the open unit disc of complex numbers.

Appendix E, Theorem 27 and Theorem 28, summarize the implications of stabilizability and detectability on the optimal capacity achieving channel input distribution, and the corresponding ergodic properties of the optimal joint process $\{(A_i^{g,*}, B_i^{g,*}) : i = 0, \ldots, \}$ and the output process $\{B_i^{g,*} : i = 0, \ldots, \}$, via properties of algebraic and discrete-time recursive Lyapunov equations and Riccati equations.

(c) To apply Theorem 28 to the feedback capacity of Theorem 19, in order to determine the properties of solutions to the algebraic Riccati equation), the following substitutions are invoked.

$$A \longmapsto C^T, \quad C^T \longmapsto D, \quad BK_WB^T \triangleq GG^T \longmapsto Q \triangleq G^TG, \quad NK_VN^T \longmapsto R. \tag{V.232}$$

The following implications hold.

(i) *If the pair (C, D) is stabilizable and the pair (G, C) is detectable, then by Theorem 28, (a), (d), the deterministic part of the optimal feedback strategy ensures stability,*

thus establishing validity of (V.222), irrespectively of the eigenvalues of channel matrix C. By Appendix E, if the pair $\big(C, D\big)$ is controllable then it is stabilizable and if the pair $\big(G, C\big)$ is observable then it is detectable.

(ii) *If the conditions in (i) hold, and in addition $(C, K_V^{\frac{1}{2}})$, $K_V \stackrel{\triangle}{=} K_V^{\frac{1}{2}} K_V^{\frac{1}{2},T}$ is a controllable pair, which is satisfied because K_V is full rank, then by Theorem 27, (d), the Lyapunov matrix equation (V.221) has a unique positive definite solution $K \succ 0$, which implies the channel output process $\{B_i^{g,*} : i = 0, \ldots, \}$ has a unique invariant distribution.*

The next example further illustrates the importance of stabilizability and detectability conditions, in determining feedback capacity, and the role of zero matrix $Q = 0$ versus $Q \neq 0$.

Example 21 (Consequences of Theorem 19): Consider the feedback capacity given in Theorem 19.

(a) Scalar with $p = q = 1, R = 1, Q = 0$. This is discussed in Section I-C. Specifically, (I.34)-(I.40), are obtained from the expressions of Theorem 19. The same solution is obtained independently in Example 13, without using the Riccati equation. This example demonstrates, once again, that whether feedback increases capacity depends on the channel parameters and transmission cost parameters $\{C, D, R, Q, K_V\}$.

(b) MIMO with $Q = 0$: Since $Q = 0$, the algebraic Riccati equation (V.220) reduces to the following matrix equation.

$$P = C^T PC - C^T PD\Big(D^T PD + R\Big)^{-1}\Big(C^T PD\Big)^T$$

$$\Longrightarrow P = 0 \quad \textit{i.e., the zero matrix is one solution.} \tag{V.233}$$

It is shown next, that feedback capacity $C_{A^\infty \to B^\infty}^{B.1}(\kappa) \equiv J^{B.1,*}\Big|_{s=s^*}$ depends on whether the eigenvalues lie inside the unit disc of the space of complex numbers $\mathbb{D}_o$, and whether feedback increases capacity is determined from the solutions of the algebraic Riccati equation.

(i) Case 1 (MIMO Stable Channel, $spec\big(C\big) \subset \mathbb{D}_o$). Since $spec\big(C\big) \subset \mathbb{D}_o$ then $\big(G, C\big), Q \stackrel{\triangle}{=} G^T G$ is detectable even though, $G = 0$, because by Definition 26, there exists an matrix L such that $spec(C - LG) \subset \mathbb{D}_o$, i.e., take $L = 0$. Similarly, (C, D) is stabilizable. By invoking Theorem 28, (with substitutions (V.232)), then the Riccati matrix equation (V.233) with $P \succeq 0$ has at most one solution, and hence $P = 0$ is the only solution. Substituting $P = 0$ into the Lyapunov equation (V.221) and (V.223)-(V.225) the following are obtained.

$$\Gamma^* = 0, \quad C^{B.1}(b) = 0,$$

$$J^{B.1,*} = \sup_{K_Z \in \mathbb{S}_+^{q\times q}} \Big\{\frac{1}{2}\log\frac{|DK_ZD^T + K_V|}{|K_V|} + s\kappa - tr\Big(s\ RK_Z\Big)\Big\}, \tag{V.234}$$

$$K = CKC^T + DK_ZD^T + K_V. \tag{V.235}$$

Recall that K is the covariance of the channel output process $\{B_i^* : i = 0, \ldots, ..\}$. By Theorem 27, (d), if K_V is full rank or the analog of 2) holds, then 1), 2) imply $K \succ 0$. Further, by Theorem 27, (b), $K \succ 0$ is the unique solution of (V.235), and hence the channel output process $\{B_i^* : i = 0, \ldots, \}$ has a unique invariant distribution.

Finally, by (V.234) and the fact that "s" correspond to the Lagrange multiplier of the transmission cost constraint, then the following holds.

$$C_{A^\infty \to B^\infty}^{B.1}(\kappa) = \sup_{K_Z \in \mathbb{S}_+^{q\times q}: tr\left(RK_Z\right) \le \kappa} \frac{1}{2}\log\frac{|DK_ZD^T + K_V|}{|K_V|} = C_{A^\infty \to B^\infty}^{noFB,B.1}(\kappa). \tag{V.236}$$

where $C_{A^\infty \to B^\infty}^{noFB,B.1}(\kappa)$ is the capacity of (V.203), (V.204) (with $Q = 0$) without feedback.

Moreover, it can be shown that the capacity achieving channel input distribution without feedback is asymptotically stationary (even if $\{Z_i : i = 0, 1 \ldots\}$ is not restricted to a stationary process, and satisfies conditional independence

$$P_{A_i|A^{i-1}}^*(da_i|a^{i-1}) = P_{A_i}^*(da_i), \quad i = 0, 1 \ldots, \tag{V.237}$$

The above discussion generalizes the scalar example discussed in Section I-C, (I.34)-(I.40), to MIMO channels.

(ii) Case 2 (MIMO Unstable Channel, $spec\big(C\big) \in \mathbb{D}_o^c \stackrel{\triangle}{=} \big\{c \in \mathbb{C} : |c| \ge 1\big\}$ and $spec\big(C\big)$ not on unit circle.) For unstable channels, $\big(G, C\big), Q \stackrel{\triangle}{=} G^T G$ is not detectable (i.e., since $Q = 0$ and C is unstable), hence condition (V.206) is violated. However, even if detectability is violated, by Theorem 27, (d) if the pair $(C, K_V^{\frac{1}{2}})$ is controllable and Lyapunov equation (V.221) has a positive definite solution $K \succ 0$, then (V.222) holds, that is, the feedback optimal strategy is stabilizing, i.e., $spec\Big(C + D\Gamma^*\Big) \in \mathbb{D}_o$, and by Theorem 28, (e), then the matrix Riccati equation has a unique solution $P \succ 0$. This is often called the maximal and stabilizing solution of the matrix Riccati equation.

The above example illustrates the link between LQG stochastic optimal control theory, and feedback capacity of G-LCM-B.1.

B. Feedback Capacity of G-CM-B.J, G-CM-A & Infinite Horizon LQG Theory

Consider the G-CM-B.J defined in Section IV-C. Then Theorem 19 is easily generalized to G-CM-B.J; this is left to the reader.

Remark 22 (Generalizations):

(a) It is possible to derive analogous results for the time-invariant version of the G-CM-A and G-CM.B.J, by invoking the formulation in Section IV-C, Section IV-E.

(b) The material of this section, illustrate that for MIMO TI-G-CMs, by invoking stochastic optimal control theory, then the characterizations of feedback capacity can be computed. Moreover, detectability and stabilizability are sufficient conditions, for the optimal channel input distribution to induce, asymptotically, an invariant distribution for the joint process $\{(A_i^g, B_i^g) : i = 0, 1, \ldots, \}$ and its marginals such that that feedback capacity is independent of the initial distribution $\mu(db_{-1})$. This illustrates the direct connection between

ergodic LQG stochastic optimal control theory and feedback capacity.

VI. Relations Between Characterizations of FTFI Capacity and Coding Theorems

In this section the importance of the characterizations of FTFI capacity is discussed, with respect to the converse and the direct part of the channel coding theorems. Sufficient conditions are identified so that the per unit time limits of the characterizations of FTFI capacity for classes A, B and C channels, corresponds to feedback capacity, independently of the material of Section V. It is noted that coding theorem based on the material of Section V are found in [1, Th. 4.1].

Consider the following definition of a code.

Definition 23 (Achievable Rates of Codes With Feedback): Given a channel distribution of Class A or B and a transmission cost function of Class A or B, an $\{(n, M_n, \epsilon_n, b^{-1}) : n = 0, 1, \ldots\}$ *code with feedback consists of the following.*

(a) A set of uniformly distributed messages $\mathscr{M}_n \stackrel{\triangle}{=} \{1, \ldots, M_n\}$ *and a set of encoding maps, mapping source messages into channel inputs of block length* $(n+1)$*, defined by*

$$\begin{aligned}\mathscr{E}^{FB}_{[0,n]}(\kappa) &\stackrel{\triangle}{=} \Big\{ g_i : \mathscr{M}_n \times \mathbb{B}^{i-1} \longmapsto \mathbb{A}_i, \quad a_0 = g_0(w, b^{-1}), \\ &\quad a_1 = g_i(w, b^0), \ldots, a_n = g_n(w, b^{i-1}), \quad w \in \mathscr{M}_n : \\ &\quad \frac{1}{n+1}\mathbf{E}^g_{b^{-1}}\Big(c_{0,n}(A^n, B^n)\Big) \leq \kappa \Big\}. \end{aligned} \tag{VI.238}$$

The codeword for any $w \in \mathscr{M}_n$ *is* $u_w \in \mathbb{A}^n$, $u_w = (g_0(w, b^{-1}), g_1(w, b^0), \ldots, g_n(w, b^{n-1}))$*, and* $\mathscr{C}_n = (u_1, u_2, \ldots, u_{M_n})$ *is the code for the message set* $\mathscr{M}_n$*.*

(b) Decoder measurable mappings $d_{0,n}(b^{-1}, \cdot) : \mathbb{B}^n_0 \longmapsto \mathscr{M}_n$, $Y^n_0 = d_{0,n}(b^{-1}, B^n_0)$*, such that the average probability of decoding error satisfies*

$$\begin{aligned}\mathbf{P}^{(n)}_e(b^{-1}) &\stackrel{\triangle}{=} \frac{1}{M_n}\sum_{w \in \mathscr{M}_n} \mathbb{P}^g\Big\{d_{0,n}(b^{-1}, B^n_0) \neq w | B^{-1} = b^{-1}, W = w\Big\} \\ &\equiv \mathbb{P}^g\Big\{d_{0,n}(B^{-1}, B^n_0) \neq W | B^{-1} = b^{-1}\Big\} \leq \epsilon_n \in [0, 1) \end{aligned} \tag{VI.239}$$

where $r_n \stackrel{\triangle}{=} \frac{1}{n+1}\log M_n$ *is the coding rate or transmission rate (and the messages are uniformly distributed over* $\mathscr{M}_n$*).*

A rate R *is said to be an achievable rate, if there exists a code sequence satisfying* $\lim_{n\longrightarrow\infty}\epsilon_n = 0$ *and* $\liminf_{n\longrightarrow\infty}\frac{1}{n+1}\log M_n \geq R$*. The feedback capacity is defined by* $C \stackrel{\triangle}{=} \sup\{R : R \text{ is achievable}\}$*, where* $R \equiv R(b^{-1})$ *may depend on the initial data* b^{-1}*.*

Note that in the above definition, the code depends on the initial data $B^{-1} = b^{-1}$, which is known to the encoder and decoder. Knowledge of $B^{-1} = b^{-1}$ at the decoder can be relaxed. However, whether the rate depends on the initial data, is often determined by establishing that in the limit, as $n \longrightarrow \infty$, the per unit time of the FTFI capacity is indeed the supremum of all achievable rates, and that the optimal channel input conditional distribution induces a channel output process, having a unique invariant distribution. Theorem 19 establishes sufficient conditions in terms of detectability and stabilizability.

Direct and converse coding theorems are derived in [2], [3], [28], [29], and [39], with respect to the above definition of a code or variants of it, under different assumptions. These are can be separated into those which treat Gaussian channels with memory, and those which treat channels with finite alphabet input and output spaces. The coding theorems in [2], [3], [28], [29], and [39] are directly applicable to channels of Class A or B and transmission cost functions of Class A or B, provided, the assumptions based on which these are derived, are adopted, or they are modified to account for additional generalities. For example, the coding theorems derived by Cover and Pombra [2] for scalar nonstationary nonergodic AGN channels with memory, are directly applicable to the G-CM-A presented in Section IV-A. For finite alphabet spaces $\{\mathbb{A}_i = \mathbb{A}, \mathbb{B}_i = \mathbb{B} : i = 0, \ldots, n\}$, the coding theorems derived by Kim [39], for the class of stationary channels with feedback, are directly applicable to NCM-A and NCM-B (without transmission cost), given in Definition 5, and they can be extended to include transmission cost constraints. The coding theorem derived by Chen and Berger [16] for the UMCO (i.e., $\{\mathbf{P}_{B_i|B_{i-1},A_i} : i = 0, \ldots, n\}$) with finite alphabet spaces, is directly applicable, while a transmission cost function of Class B with $K = 1$ can be easily incorporated. The various coding theorems derived by Kim *et al.* [3] and Permuter *et al.* [29] for finite alphabet spaces (without transmission cost constraints), under the assumption of time-invariant deterministic feedback, are directly applicable to channel distributions of Class A or B, and since their method is based on irreducibility of the channel distribution, they also extend to problems with time-invariant transmission cost functions of Class A or B.

Most converse coding theorems found in the literature, do not address the fundamental questions whether i) an optimal channel input conditional distribution corresponding to the characterization of FTFI capacity exists, and ii) its per unit time limit exists and it is finite. For channels defined on countable, continuous or abstract alphabet spaces, it is well-known that Shannon's information measures, such as, entropy, relative entropy, mutual information, and conditional mutual information, are not necessarily continuous with respect to strong topologies (see [40] for various examples). For such spaces, conditions for existence of optimal channel input distributions corresponding to the characterization of FTFI capacity are given in [31, Th. 8 and 17], using the topology of weak convergence of probability distributions. Moreover, to ensure tightness of upper bounds on any achievable rate, expressed in terms of information theoretic measures, it is necessary to identify the information structures of optimal channel input distributions corresponding to the characterization of FTFI, i.e., as obtained in Theorem 1.

For the direct part of the coding theorem, in addition to the conditions of the converse coding theorem, it is sufficient to identify conditions for information stability in the

sense of Pinsker [32]. Information stability implies that the asymptotic equipartition property (AEP) of directed information holds, from which the direct part of the coding theorem follows by standard arguments. To show information stability for general channels defined on abstract alphabet spaces is often a challenging task. However, such conditions can be identified via the ergodic theory of Markov decision [38, Th. 5.7.9], by showing asymptotic pathwise optimality of the per unit time limit of directed information density. For any channel distribution of Class A, Class B, and transmission cost of Class A or B, with corresponding information density and transmission cost, evaluated at the optimal channel input distributions, $\{\pi^*(da_i|b^{i-1}) : i = 0, 1, \ldots, n\} \in \overline{\mathscr{P}}^{A}_{[0,\infty]} \bigcap \mathscr{P}_{[0,n]}(\kappa)$, $\overset{\circ}{\mathscr{P}}{}^{B.J}_{[0,\infty]} \bigcap \mathscr{P}_{[0,n]}(\kappa)$, respectively, such results can be obtained by combining [31, Th. 8 and 17] and [38, Th. 5.7.9].

For the Gaussian channels considered in Section IV and Section V, the coding theorems in [1, Th. 4.1], illustrate the various connections to ergodic theory, for unstable channels. Specifically, Theorem 19 describes sufficient conditions for validity of both the direct and the converse parts of the coding theorems, using the ergodic theory of linear-quadratic Gaussian stochastic optimal control problems, in terms of detectability and stabilizability (see [1, Th. 4.1]. for details).

The coding theorem stated below, is generic, in the sense that sufficient conditions are imposed, to ensure both the converse part and direct part of coding theorem hold.

Theorem 24 (Coding Theorem): Consider any channel distribution and transmission cost function of Class A or B, with corresponding characterizations of FTFI capacity denoted by $C_{A^n \to B^n}(\kappa)$, and channel input conditional distributions denoted by $\{\pi_i(da_i|\mathscr{I}_i^P) : i = 0, 1, \ldots, n\} \in \mathscr{P}_{[0,n]}(\kappa)$, where $\{\mathscr{I}_i^P : i = 0, \ldots, n\}$ is the information structure of channel input distributions.

Suppose the following two conditions hold.

i) Conditional independence (II.56) holds;

ii) there exists an optimal channel input conditional distribution $\{\pi_i^*(da_i|\mathscr{I}_i^P) : i = 0, 1, \ldots, n\} \in \mathscr{P}_{[0,n]}(\kappa)$, which achieves the characterization of the FTFI capacity, and its per unit time limit exists (if not replace it by lim inf*) it is finite, and independent of initial distribution $B^{-1} \sim \mu(db^{-1})$, denoted by $C_{A^\infty \to B^\infty}(\kappa)$.*

Define the following.

For $\{\pi_i^*(da_i|\mathscr{I}_i^P) : i = 0, 1, \ldots, n\} \in \mathscr{P}_{[0,n]}(\kappa)$ (assuming condition ii)), the directed information density is called stable, if $\forall \varepsilon > 0$ and $\forall \delta > 0$ there exists an integer $n_0(\varepsilon, \delta) > 0$ such that

$$\mathbf{P}^{\pi^*}_{b^{-1}}\Big\{(A^n, B^n) \in \mathbb{A}^n \times \mathbb{B}^n : \frac{1}{n+1}\Big|\mathbf{E}^{\pi^*}_{\mu}\big\{\mathbf{i}^{\pi^*}(A^n, B^n)\big\} - \mathbf{i}^{\pi^*}(A^n, B^n)\Big| > \varepsilon\Big\} < \delta, \quad \forall n > n_0(\varepsilon, \delta), \quad \forall b^{-1} \in \mathbb{B}^{-1} \tag{VI.240}$$

where for channel distribution of Class A, and transmission cost of Class A or B, the directed information density is $\mathbf{i}^{\pi^*}(A^n, B^n) \stackrel{\triangle}{=} \sum_{i=0}^n \log\Big(\frac{\mathbf{P}(\cdot|B^{i-1}, A_i)}{\mathbf{P}^{\pi^*}(\cdot|B^{i-1})}(B_i)\Big)$, $i = 0, \ldots, n$ (and similarly for Channels of Class B).

For $\{\pi_i^*(da_i|\mathscr{I}_i^P) : i = 0, 1, \ldots, n\} \in \mathscr{P}_{[0,n]}(\kappa)$ (assuming condition ii)), the transmission cost is called stable, if $\forall \varepsilon > 0$ and $\forall \delta > 0$ there exists an integer $n_0(\varepsilon, \delta) > 0$ such that

$$\mathbf{P}^{\pi^*}_{b^{-1}}\Big\{(A^n, B^n) \in \mathbb{A}^n \times \mathbb{B}^n : \frac{1}{n+1}\Big|\mathbf{E}^{\pi^*}_{\mu}\big\{c_{0,n}(A^n, B^n)\big\} - c_{0,n}(A^n, B^n)\Big| > \varepsilon\Big\} < \delta, \quad \forall n > n_0(\varepsilon, \delta), \quad \forall b^{-1} \in \mathbb{B}^{-1}. \tag{VI.241}$$

Then the following hold.

(a) (Converse) If conditions i), ii) hold, then any achievable rate R of codes with feedback given in Definition 23, satisfies the following inequalities.

$$\begin{aligned} R &\le \liminf_{n\to\infty} \frac{1}{n+1}\log M_n \\ &\le \liminf_{n\to\infty} \sup_{\{g_i(\cdot,\cdot):i=0,\ldots,n\}\in\mathscr{C}^{FB}_{[0,n]}(\kappa)} \Big\{ \frac{1}{n+1}\sum_{i=0}^{n} I(A_i; B_i|B^{i-1})\Big\} \qquad \text{(VI.242)} \\ &\le \liminf_{n\to\infty} \sup_{\{\pi(a_i|\mathscr{I}_i^P):i=0,1,\ldots,n\}\in\mathscr{P}_{[0,n]}(\kappa)} \Big\{ \frac{1}{n+1}\sum_{i=0}^{n} I(A_i; B_i|B^{i-1})\Big\} \\ &\equiv C_{A^\infty \to B^\infty}(\kappa) \qquad \text{(VI.243)} \end{aligned}$$

(b) (Direct) If conditions i), ii) hold and in addition

iii) the directed information density is stable,

iv) the transmission cost is stable,

v) For each n the FTFI capacity $C_{A^n \to B^n}(\kappa)$ is continuous in $\kappa \in (\kappa_{min}, \infty)$,

then any rate $R < C_{A^\infty \to B^\infty}(\kappa)$ is achievable.

Proof: (a) Condition i) implies the well-known data processing inequality, while condition ii) implies existence of the optimal channel input distribution and finiteness of the corresponding characterizations of the FTFI capacity and its per unit time limit. Hence, the statements of inequalities follow by applying Fano's inequality. If conditions i) and ii) hold, the derivation is also found in many references, (i.e., [2], [3], [28], [29], [39]).

(b) This is standard, because conditions ii)-v) are sufficient to ensure the AEP holds, and hence standard random coding arguments hold (i.e., following Ihara [7], by replacing the information density of mutual information by that of directed information). □

It is noted that alternative achievability theorems can be obtained by combining the achievability theorem derived by Permuter *et al.* [29], which is based on bounding the error of Maximum Likelihood (ML) decoding, and the characterizations of FTFI capacity and feedback capacity.

The following remark summarizes the connection between Theorem 24 (achievability) and Theorem 19.

Remark 25 (Achievability Based on LQG Theory [1, Th. 4.1]):

(a) For the TI-G-CM-B.1, Theorem 19 gives sufficient conditions, in terms of the channel variables $\{C, D, R, Q, K_V\}$, expressed in terms of detectability and

stabilizability, for $J^{B.1,*}\Big|_{s=s^*} = C^{B.1}_{A^\infty \to B^\infty}(\kappa)$, defined by (V.225), to correspond to Feedback Capacity, for any initial distribution $\mu(db_{-1})$, irrespectively, of whether the channel is stable or unstable.

(b) For the TI-G-CM-B.J, similarly to Theorem 19, sufficient conditions can be obtained, for the corresponding solution of the dynamic programming, denoted by $J^{B.J,*}\Big|_{s=s^*} = C^{B.J}_{A^\infty \to B^\infty}(\kappa)$ to correspond to Feedback Capacity.

(c) For Multidimensional Gaussian sources to be encoded and transmitted over any one of the channels, G-CM-A, G-CM-B.1, G-CM-B.J, coding strategies can be constructed, which achieve the corresponding characterizations of the FTFI capacity, and Feedback capacity.

Finally, it is noted for general Gaussian channels with past dependence on channel inputs and channel outputs then a treatment analogous to the one of this paper can be carried out, although it is more involved [11].

VII. Conclusion

The information structures of optimal channel input conditional distributions derived in [1] are applied to derive alternative characterizations of FTFI capacity, based on randomized information lossless strategies, driven by independent RVs. Their per unit time limiting versions are analyzed, without imposing á priori assumptions, which rule out the dual role of such strategies, to achieve the FTFI capacity characterizations and feedback capacity, to control the channel output process and to transmit new information through the channel.

The characterizations of FTFI capacity and feedback capacity are investigated for application examples of MIMO Gaussian channel models (G-CMs) with memory. In these application examples, the randomized strategies decompose into a deterministic part, which corresponds to the control process, and a random part, which corresponds to an innovations process. Via this decomposition a separation principle is established. The deterministic control part is shown to be directly related to the role of optimal control strategies of linear-quadratic-Gaussian control theory, to control output processes, and, in general, to the feedback control theory of linear stochastic systems; the random or innovations part is shown to be directly related to role of encoders to achieve capacity, by transmitting new information over the channel. Moreover, whether feedback capacity exists, and feedback increases capacity is shown to be directly linked to the role of the deterministic part of randomized strategies to control the channel output process.

Appendix

A. Proof of Theorem 6.

(a) By Theorem 1, (1), the optimal channel input distributions belong to $\overline{\mathscr{P}}^A_{[0,n]} = \{\pi_i(da_i|b^{i-1}) \equiv \mathbf{P}_{A_i|B^{i-1}}(a_i|b^{i-1}) : i = 0, 1, \ldots, n\}$, and satisfy the transmission cost constraint. An application of Lemma 3 implies CON(a.1) holds. Moreover, by assumption (III.73), the property of optimal channel input distribution, and by virtue of (III.80), CON(a.2), also holds. Clearly, ii) and iii) imply the processes $\{U_i : i = 0, \ldots, n\}$ and $\{V_i : i = 0, \ldots, n\}$ are independent. (III.81)-(III.83) are direct consequences, because for each channel input distribution $\overline{\mathscr{P}}^A_{[0,n]}$ there exists a randomized strategy $\{\overline{e}^A_i(\cdot,\cdot) : i = 0, 1, \ldots, n\}$ which realizes it.

(b) The first part follows from Remark 4, i.e., the quantize representation of distributions, as follows. Set $z_i \stackrel{\triangle}{=} G(u) = \mathbf{I}[F_{Z_i}](u_i), i = 0, \ldots, n$. Then from (a) it follows that $\overline{e}^A_i(b^{i-1}, u_i) = \overline{e}^A_i(b^{i-1}, G(u_i))$. Hence, for each $\overline{e}^A_i(\cdot,\cdot)$ there exists another function $e^A_i(\cdot,\cdot)$ driven by z_i such that $a_i = e^A_i(b^{i-1}, z_i), i = 0, \ldots, n$. Since $\{U_i : i = 0, \ldots, n\}$ are independent then $Z_i = G(U_i)$ are independent, and (a) holds as claimed. Moreover, (III.85) is obtained because, the distributions are expressed in terms of the randomized strategy $\{e^A_i(\cdot,\cdot) : i = 0, \ldots, n\}$ and $\{Z_i : i = 0, \ldots, n\}$. Next, it shown that (III.89) holds, for the restricted class of randomized strategies $\mathscr{E}^{IL-A}_{[0,n]}(\kappa)$ defined by (III.88). Recall that for channels of Class A, $I(A^n \to B^n) = \sum_{i=0}^n I(A^i; B_i|B^{i-1}) = \sum_{i=0}^n I(A_i; B_i|B^{i-1}), i = 0, \ldots, n$, as defined by (II.58), without the supremum. For any $\{e^A_i(\cdot,\cdot) : i = 0, 1, \ldots, n\} \in \mathscr{E}^{IL-A}_{[0,n]}(\kappa)$, and for a fixed b^{i-1}, by the bijective property of the map $e^A_i(b^{i-1}, \cdot)$ and the measurability of its inverse, for $i = 0, \ldots, n$, then the following sequence of identities hold (see Pinsker [32, Th. 3.7.1], and Corollary following it, or Ihara [7, Th. 1.6.3, (I.3)]).

$$\begin{aligned} I(A_i; B_i|B^{i-1} = b^{i-1}) &\stackrel{(a)}{=} I(A_i, Z_i; B_i|B^{i-1} = b^{i-1}) \\ &\stackrel{(b)}{=} I(A_i; B_i|B^{i-1} = b^{i-1}, Z_i) + I(Z_i; B_i|B^{i-1} = b^{i-1}) \\ &\stackrel{(c)}{=} I(Z_i; B_i|B^{i-1} = b^{i-1}), \quad \forall b^{i-1},\ i = 0, \ldots, n \end{aligned} \tag{A.244}$$

where (a) holds because for a fixed b^{i-1}, then $a_i = e^A_i(b^{i-1}, \cdot)$ uniquely defines z_i, (b) is due to the chain rule of mutual information, (c) is due to $I(A_i; B_i|B^{i-1} = b^{i-1}, Z_i) = 0$, which follows from $\{e^A_i(\cdot,\cdot) : i = 0, \ldots, n\} \in \mathscr{E}^{IL-A}_{[0,n]}(\kappa)$ implies $\mathbf{P}_{B_i|B^{i-1},A_i,Z_i} = \mathbf{P}_{B_i|B^{i-1},Z_i} = 0$, i.e., $a_i = e^A_i(b^{i-1}, z_i)$, $i = 0, \ldots, n$. Moreover, by using

$$I(A_i; B_i|B^{i-1}) = \int I(A_i; B_i|B^{i-1} = b^{i-1}) \mathbf{P}_{B^{i-1}}(db^{i-1})$$

then from (A.244) it follows that

$$\begin{aligned} I(A_i; B_i|B^{i-1}) &= I(A_i, Z_i; B_i|B^{i-1}) \\ &= I(A_i; B_i|B^{i-1}, Z_i) + I(Z_i; B_i|B^{i-1}) \\ &= I(Z_i; B_i|B^{i-1}), \quad i = 0, \ldots, n \\ &\quad \text{if } \{e^A_i(\cdot,\cdot) : i = 0, \ldots, n\} \in \mathscr{E}^{IL-A}_{[0,n]}(\kappa). \end{aligned} \tag{A.245}$$

The above identities establish (III.89), (III.90), where the supremum is taken over all information lossless randomized strategies $\mathscr{E}^{IL-A}_{[0,n]}(\kappa)$ and $\{\mathbf{P}_{Z_i} : i = 0, \ldots, n\}$. This completes the prove.

B. Proof of Theorem 8.

(a) This is obtained by utilizing the information structure of the optimal channel input distribution $\{\pi_i(da_i|b_{i-J}^{i-1}) \equiv \mathbf{P}_{A_i|B_{i-J}^{i-1}}(a_i|b_{i-J}^{i-1}) : i = 0, 1, \ldots\}$, and Lemma 3. (b) The rest of the derivation follows from that of Theorem 6.

C. Proof of Theorem 10.

(a) By Assumption A, the channel distribution is conditionally Gaussian, given by

$$\mathbb{P}\Big\{B_i \le b_i \Big| B^{i-1} = b^{i-1}, A^i = a^i\Big\} = \mathbb{P}\Big\{V_i \le b_i - \sum_{j=0}^{i-1} C_{i,j} b_j - D_{i,i}\, a_i)\Big\}, \tag{A.246}$$

$$\sim N\Big((\sum_{j=0}^{i-1} C_{i,j} b_j + D_{i,i} a_i), K_{V_i}\Big), \quad i = 0, 1, \ldots, n \tag{A.247}$$

that is, the conditional mean of B_i is linear in $\{b^{j-1}, a_i\}$ and the conditional covariance is constant. The conditional probability distribution of $\{B_i : i = 0, \ldots, n\}$ is given by

$$\mathbb{P}\Big\{B_i \le b_i \Big| B^{i-1} = b^{i-1}\Big\} = \int_{\mathbb{A}_i} \mathbb{P}\Big\{V_i \le b_i - \sum_{j=0}^{i-1} C_{i,j} b_j - D_{i,i}\, a_i\Big\} \pi_i(da_i|b^{i-1}), \quad i = 0, 1, \ldots, n. \tag{A.248}$$

In view Assumption A, and properties of conditional entropy, then $H(B_i|B^{i-1}, A_i) = H(V_i|B^{i-1}, A_i) = H(V_i)$, $i = 0, \ldots, n$, and directed information is given by

$$I(A^n \to B^n) = \sum_{i=0}^{n} \Big\{H(B_i|B^{i-1}) - H(B_i|B^{i-1}, A_i)\Big\} = \sum_{i=0}^{n} H(B_i|B^{i-1}) - \sum_{i=0}^{n} H(V_i). \tag{A.249}$$

Hence, the characterization of FTFI Feedback Capacity is given by the following expression.

$$C_{A^n \to B^n}^{A}(\kappa) \stackrel{\triangle}{=} \sup_{\substack{\{\pi_i(da_i|b^{i-1}), i=0,\ldots,n: \\ \frac{1}{n+1}\sum_{i=0}^{n} \mathbf{E}\{\langle A_i, R_{i,i} A_i\rangle + \langle B^{i-1}, Q_i(i-1)B^{i-1}\rangle\} \le \kappa\}}} \Big\{ H(B^n) - H(V^n)\Big\}. \tag{A.250}$$

By the entropy maximizing property of the Gaussian distribution the right hand side of (A.250) is bounded above by the inequality $H(B^n) \le H(B^{g,n})$, where $B^{g,n} \stackrel{\triangle}{=} \{B_i^g : i = 0, 1, \ldots, n\}$ is jointly Gaussian distributed, and the average transmission cost constraint is satisfied. Suppose the channel input distribution is conditionally Gaussian, denoted by $\{\pi_i^g(da_i|b^{i-1}) \equiv \mathbf{P}^g_{A_i|B^{i-1}}(a_i|b^{i-1}) : i = 0, 1, \ldots, n\}$, with conditional mean which is a linear combination of $\{B_i : i = 0, \ldots, n-1\}$, and conditional covariance which is non-random, i.e., independent of the channel output process $\{B^{i-1} : i = 0, \ldots, n\}$. Then for such conditionally Gaussian distributions there exists an orthogonal realization $A_i = \sum_{j=0}^{i-1} \Gamma_{i,j} B_j + Z_i, i = 0, \ldots, n$, where $\Gamma_{i,j}$ are non-random, and $\{Z_i : i = 0, \ldots, n\}$ is independent Gaussian, satisfying (IV.109), (IV.110) (these follow from Assumption A and the information structure of the maximizing channel input distribution, $\{\mathbf{P}_{A_i|A^{i-1},B^{i-1}}(a_i|a^{i-1}, b^{i-1}) = \mathbf{P}^g_{A_i|B^{i-1}}(a_i|b^{i-1}) : i = 0, 1, \ldots, n\}$ or CON(a.2)). This implies the joint process is jointly Gaussian, i.e., $\{(A_i, B_i) \equiv (A_i^g, B_i^g) : i = 0, \ldots, n\}$, hence the upper bound $H(B^n) \le H(B^{g,n})$ holds with equality.

(b) This follows from (a). However, it can be established via the alternative characterization given in Theorem 6, (III.83), as follows. Since any candidate of the optimal channel input distribution is $\{\mathbf{P}_{A_i|B^{i-1}}(a_i|b^{i-1}) : i = 0, 1, \ldots, n\}$, by Theorem 6 there exists a measurable function $e_i^A : \mathbb{B}^{i-1} \times \mathbb{Z}_i \to \mathbb{A}_i, \mathbb{Z}_i, a_i = e_i^A(b^{i-1}, z_i), i = 0, 1, \ldots, n$ such that

$$\mathbf{P}_{A_i|B^{i-1}}(da_i|b^{i-1}) = \mathbf{P}_{Z_i}\big(z_i : e_i^A(b^{i-1}, z_i) \in da_i\big), \quad i = 0, 1, \ldots, n. \tag{A.251}$$

Substituting the randomized strategy into the channel model (IV.96), then

$$B_i = \sum_{j=0}^{i-1} C_{i,j} B_j + D_{i,i}\, e_i^A(B^{i-1}, Z_i) + V_i, \quad i = 1, \ldots, n, \tag{A.252}$$

$$\mathscr{E}_{[0,n]}^{A}(\kappa) \stackrel{\triangle}{=} \Big\{e_i^A(B^{i-1}, Z_i), i = 0, \ldots, n : \text{ Theorem 6, CON(a.2) holds with } \{U_i : i = 0, \ldots, n\} \text{ replaced by } \{Z_i : i = 0, \ldots, n\}, \frac{1}{n+1}\mathbf{E}^{e^A}\Big(\sum_{i=0}^{n}\big\{\langle e_i^A(B^{i-1}, Z_i), R_{i,i} e_i^A(B^{i-1}, Z_i)\rangle + \langle B^{i-1}, Q_i(i-1)B^{i-1}\rangle\big\}\Big) \le \kappa\Big\}. \tag{A.253}$$

By the entropy maximizing property of the Gaussian distribution the right hand side of (A.249) (with $\{a_i = e_i^A(b^{i-1}, z_i) : i = 0, \ldots, n\}$) is bounded above by the inequality[11] $H^{e^A}(B^n) \le H^{e^A}(B^{g,n})$, where $B^{g,n} \stackrel{\triangle}{=} \{B_i^g : i = 0, 1, \ldots, n\}$ is jointly Gaussian distributed. The upper bound is achieved if $e_i^A(b^{i-1}, z_i)$ is a linear combination of (b^{i-1}, z_i) for $i = 0, \ldots, n$, $\{Z_i : i = 0, \ldots, n\}$ is independent Gaussian satisfying (IV.109), (IV.110) and the average transmission cost constraint is satisfied. Hence, $A^n = A^{g,n} \stackrel{\triangle}{=} \{A_i^g : i = 0, 1, \ldots, n\}, B^n = B^{g,n}$ are jointly Gaussian distributed. Thus, the alternative characterization of the FTFI capacity is given by (IV.102)-(IV.110) are obtained. Note that (IV.109), (IV.110) follows from Assumption A and the information structure of the maximizing channel input distribution, $\{\mathbf{P}_{A_i|A^{i-1},B^{i-1}}(a_i|a^{i-1}, b^{i-1}) = \mathbf{P}^g_{A_i|B^{i-1}}(a_i|b^{i-1}) : i = 0, 1, \ldots, n\}$ or CON(a.2).

[11] The superscript indicates the distribution depends on the strategy $\{e_i^A(\cdot) : i = 0, \ldots, n\}$.

D. Proof of Theorem 14

(a) (IV.154), (IV.156), follow directly from the re-formulation of the problem.

(b) Clearly, (IV.157) is the cost-to-go for (IV.154).

(c) The dynamic programming recursions follow directly from (IV.157), and these are generalizations of classical dynamic programming solutions [5, Ch. 5], [4, Ch. 7]. It should be noted that, in general, the cost-to-go (IV.157) can be computed in two steps; in the first step the cost-to-go (IV.157) is defined without the optimization over $\{K_{Z_i} : i = 0, \ldots, n\}$, which implies $C_i^{B.1}(b_{i-1})$ is replaced by $C_i^{B.1}(b_{i-1}; K_{Z_j}, j = i, \ldots, n)$, while in the second step the optimization $C_0^{B.1}(b_{-1}; K_{Z_0}, j = 0, \ldots, n)$ is carried out over $K_{Z_j}, j = 0, \ldots, n$.

(d)-(e) The derivation is based on solving the dynamic programming equations, as done for LQG stochastic optimal control problems [5], with some modifications to account for the fact that the strategies are randomized (instead of deterministic). An alternative shorter derivation is given in Remark 15, (a). Let $C_n^{B.1}(b_{n-1}) = -s\langle b_{n-1}, Q_{n,n-1}b_{n-1}\rangle + r(n)$, $P(n) = Q_{n,n-1}$, and $r(n)$ given by (IV.170). It can be verified this is indeed the solution at the last stage of the dynamic programming recursions, i.e., (IV.158), and that $g_n^{B.1,*}(b_{n-1}) = 0$. Then $P(n) = P^T(n) \succeq 0$. Suppose for $j = i+1, i+2, \ldots, n$, $P(j) = P^T(j) \succeq 0$, $C_j^{B.1}(b_{j-1}) = -s\langle b_{j-1}, P(j)b_{j-1}\rangle + r(j)$. It will be shown that $P(i) = P^T(i) \succeq 0$, $C_i^{B.1}(b_{i-1}) = -s\langle b_{i-1}, P(i)b_{i-1}\rangle + r(i)$, as stated in (d), (e).

The following calculations follow directly from Assumptions B (i.e., $\mathbf{E}^{g^{B.1}}\{Z_i | B_{i-1}\} = 0, \mathbf{E}^{g^{B.1}}\{V_i | B_{i-1}\} = 0$, and Z_i independent of V_i).

$$-s\Big\{\langle u_i, R_{i,i}u_i\rangle + \langle b_{i-1}, Q_{i,i-1}b_{i-1}\rangle\Big\} + \mathbf{E}^{g^{B.1}}\Big\{C_{i+1}^{B.1}(B_i^g)\Big|B_{i-1}^g = b_{i-1}\Big\} \tag{A.254}$$

$$= -s\Big\{\langle u_i, R_{i,i}u_i\rangle + \langle b_{i-1}, Q_{i,i-1}b_{i-1}\rangle\Big\} + \mathbf{E}^{g^{B.1}}\Big\{C_{i+1}^{B.1}(C_{i,i-1}B_{i-1}^g + D_{i,i}U_i^g + D_{i,i}Z_i + V_i)\Big|B_{i-1}^g = b_{i-1}\Big\} \tag{A.255}$$

$$= -s\begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix}^T \begin{bmatrix} Q_{i,i-1} & 0 \\ 0 & R_{i,i} \end{bmatrix} \begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix} + r(i+1) - s\mathbf{E}^{g^{B.1}}\Big\{\langle C_{i,i-1}B_{i-1}^g + D_{i,i}U_i^g + D_{i,i}Z_i + V_i, P(i+1)\Big(C_{i,i-1}B_{i-1}^g + D_{i,i}U_i^g + D_{i,i}Z_i + V_i\Big)\rangle\Big|B_{i-1} = b_{i-1}\Big\} \tag{A.256}$$

$$= -s\begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix}^T \cdot \begin{bmatrix} C_{i,i-1}^T P(i+1)C_{i,i-1} + Q_{i,i-1} & C_{i,i-1}^T P(i+1)D_{i,i} \\ D_{i,i}^T P(i+1)C_{i,i-1} & D_{i,i}^T P(i+1)D_{i,i} + R_{i,i} \end{bmatrix} \cdot \begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix} + r(i+1) - tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\Big]\Big) \tag{A.257}$$

$$= -s\begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix}^T \begin{bmatrix} H_{11}(i) & H_{12}(i) \\ H_{12}^T(i) & H_{22}(i) \end{bmatrix} \begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix} + r(i+1) - tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\Big]\Big). \tag{A.258}$$

Note that

$$H_{11}(i) = H_{11}^T(i) \succeq 0, \tag{A.259}$$

$$H_{22}(i) = H_{22}^T(i) = D_{i,i}P(i+1)D_{i,i} + R_{i,i} \succeq R_{i,i} \succ 0. \tag{A.260}$$

By the induction hypothesis and $R_{i,i} \in S_{++}^{q\times q}$, $Q_{i,i-1} \in S_+^{p\times p}$, the following hold.

$$\sup_{(u_i, K_{Z_i})\in\mathbb{R}^q \times S_+^{q\times q}} \Big\{\frac{1}{2}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(sR_{i,i}K_{Z_i}\Big) - s\Big[\langle u_i, R_{i,i}u_i\rangle + \langle b_{i-1}, Q_{i,i-1}b_{i-1}\rangle\Big] + \mathbf{E}^{g^{B.1}}\Big\{C_{i+1}^{B.1}(B_i^g)\Big|B_{i-1}^g = b_{i-1}\Big\}\Big\}$$

$$= \sup_{(u_i, K_{Z_i})\in\mathbb{R}^q \times S_+^{q\times q}} \Big\{\frac{1}{2}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(sR_{i,i}K_{Z_i}\Big) - tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\Big]\Big) - s\begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix}^T \begin{bmatrix} H_{11}(i) & H_{12}(i) \\ H_{12}^T(i) & H_{22}(i) \end{bmatrix} \begin{bmatrix} b_{i-1} \\ u_i \end{bmatrix} + r(i+1)\Big\} \tag{A.261}$$

$$= \sup_{K_{Z_i}\in S_+^{q\times q}} \sup_{u_i\in\mathbb{R}^q} \Big\{ -s\begin{bmatrix} b_{i-1} \\ u_i + H_{22}^{-1}(i)H_{12}^T(i)b_{i-1} \end{bmatrix}^T \cdot \begin{bmatrix} H_{11}(i) - H_{12}(i)H_{22}^{-1}(i)H_{12}^T(i) & 0 \\ 0 & H_{22}(i) \end{bmatrix} \cdot \begin{bmatrix} b_{i-1} \\ u_i + H_{22}^{-1}(i)H_{12}^T(i)b_{i-1} \end{bmatrix} + \frac{1}{2}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(sR_{i,i}K_{Z_i}\Big) - tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\Big]\Big) + r(i+1)\Big\} \tag{A.262}$$

$$= \sup_{K_{Z_i}\in S_+^{q\times q}} \Big\{ -s\langle b_{i-1}, [H_{11}(i) - H_{12}(i)H_{22}^{-1}(i)H_{12}^T(i)]b_{i-1}\rangle + \frac{1}{2}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}|}{|K_{V_i}|} - tr\Big(sR_{i,i}K_{Z_i}\Big) - tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T + K_{V_i}\Big]\Big) + r(i+1)\Big\} \tag{A.263}$$

(because $H_{22}(i) \succ 0$, and the optimal control is $u_i = -H_{22}^{-1}(i)H_{12}^T(i)b_{i-1}$),

$$= -s\langle b_{i-1}, [H_{11}(i) - H_{12}(i)H_{22}^{-1}(i)H_{12}^T(i)]b_{i-1}\rangle$$

$$+\sup_{K_{Z_i}\in S_+^{q\times q}}\Big\{\frac{1}{2}\log\frac{|D_{i,i}K_{Z_i}D_{i,i}^T+K_{V_i}|}{|K_{V_i}|}-tr\Big(sR_{i,i}K_{Z_i}\Big)$$
$$-tr\Big(sP(i+1)\Big[D_{i,i}K_{Z_i}D_{i,i}^T+K_{V_i}\Big]\Big)+r(i+1)\Big\} \tag{A.264}$$
$$=-s\langle b_{i-1},P(i)b_{i-1}\rangle+r(i),$$
$$\text{if } (IV.163),\ (IV.169)\ \text{hold.} \tag{A.265}$$

Note that $P(\cdot)\stackrel{\triangle}{=}H_{11}(\cdot)-H_{12}(\cdot)H_{22}^{-1}(i)H_{12}^T(\cdot)$ is precisely (IV.163), and it is not affected by $\{K_{Z_i}:i=0,\ldots,n\}$, that is, the deterministic part of the strategy is separated from the random part of the strategy. For each $i=0,1,\ldots,n-1$, define $g_i^{B.1,*}(\cdot)$ by

$$g_i^{B.1,*}(b_{i-1})\stackrel{\triangle}{=}-H_{22}^{-1}(i)H_{12}^T(i)b_{i-1}$$
$$=-\Big[D_{i,i}P(i+1)D_{i,i}+R_{i,i}\Big]^{-1}$$
$$\times D_{i,i}^TP(i+1)C_{i,i-1}b_{i-1}. \tag{A.266}$$

Then

$$C_i^{B.1}(b_{i-1})$$
$$=-s\langle b_{i-1},P(i)b_{i-1}\rangle+r(i),\quad(\text{since } P(\cdot)\text{ does not depend on } K_{Z_i},\text{ then})$$
$$=-s\langle b_{i-1},P(i)b_{i-1}\rangle+s(n+1)\kappa$$
$$+\sup_{K_{Z_j}\in S_+^{q\times q}:j=i,\ldots,n}\Big\{\frac{1}{2}\sum_{j=i}^{n}\log\frac{|D_{j,j}K_{Z_j}D_{j,j}^T+K_{V_j}|}{|K_{V_j}|}$$
$$-\sum_{j=i}^{n}tr\Big(s\,R_{j,j}K_{Z_j}+sP(j+1)\Big[D_{j,j}K_{Z_j}D_{j,j}^T+K_{V_j}\Big]\Big)\Big\} \tag{A.267}$$

where the supremum in (A.267) follows from (A.264). Finally, since $P(\cdot)\stackrel{\triangle}{=}H_{11}(\cdot)-H_{12}(\cdot)H_{22}^{-1}(i)H_{12}^T(\cdot)$, then the Riccati difference equation (IV.164) is obtained.

(f), (g) follow from the constraint and expression of cost-to-go.

(h) By definition, the minimum cost for zero information rate corresponds to $K_{Z_i}=0, i=0,\ldots,n$, hence the statement. This completes the prove.

E. Lyapunov & Riccati Equations of Gaussian Linear Stochastic Systems and LQG Theory

In this section, some of the basic concepts of linear systems are introduced, and fundamental theorems relating Lyapunov stability and Riccati equations to stability of linear systems are given. These are found in [4], [5], and [41].

The open unit disc of the space of complex number $\mathbb{C}$, is defined by $\mathbb{D}_o\stackrel{\triangle}{=}\{c\in\mathbb{C}:|c|<1\}$. The Spectrum of a matrix $A\in\mathbb{R}^{q\times q}$ (the set of all its eigenvalues), is denoted by $spec(A)\subset\mathbb{C}$. A matrix $A\in\mathbb{R}^{q\times q}$ is called exponentially stable if all its eigenvalues are within the open unit disc, that is, $spec(A)\subset\mathbb{D}_o$.

Consider the time-invariant representation of a finite-dimensional Gaussian system described by the following equations.

$$X_{i+1}=AX_i+BW_i,\quad X_0=x_0,\quad i=1,\ldots,n, \tag{A.268}$$
$$Y_i=CX_i+NV_i,\quad i=0,\ldots,n \tag{A.269}$$
$$X_0\in\mathbb{R}^q,\ X_0\sim N(0,K_{X_0}),\quad W_i\sim N(0,K_W),\ W_i\in\mathbb{R}^k,$$
$$Y_i\in\mathbb{R}^p,\quad V_i\sim N(0,K_V),\ V_i\in\mathbb{R}^m,\ i=0,\ldots,n$$
$$\text{and } \{(W_i,V_i):i=0,\ldots,n\}\text{ mutual independent} \tag{A.270}$$
$$\text{and independent of } X_0. \tag{A.271}$$

Answers to questions of convergence of covariance matrices and existence of invariant distribution of the joint process $\{(X_i,Y_i):i=0,\ldots,n\}$ (and its marginals), and convergence of conditional covariances and existence of conditional invariant distribution of minimum mean-square error estimates of $\{X_i:i=0,\ldots,n\}$ from data $\{Y_i:i=0,1,\ldots,\}$, governed by the Kalman filter recursions, are directly related to certain properties of the matrices $\{A,B,C,N,K_W,K_V\}$. These are defined below.

Definition 26 (Stabilizablity and Detectability): Let $(A,B,C)\in\mathbb{R}^{q\times q}\times\mathbb{R}^{q\times k}\times\mathbb{R}^{p\times q}$.

(a) The pair of matrices (A,B) *is called stabilizable if there exists an matrix* $K\in\mathbb{R}^{k\times q}$ *such that the eigenvalues of* $A-BK$ *lie in* $\mathbb{D}_o$ *(i.e.,* $spec\big(A-BK\big)\subset\mathbb{D}_o$*).*

(b) The pair of matrices (C,A) *is called detectable if* (A^T,C^T) *is stabilizable, i.e., there exists an matrix* $L\in\mathbb{R}^{q\times p}$ *such that the eigenvalues of* $A-LC$ *lie in* $\mathbb{D}_o$ *(i.e.,* $spec\big(A-LC\big)\subset\mathbb{D}_o$*).*

(c) The pair of matrices (A,B) *is called controllable if*

$$rank\Big(\mathscr{C}\Big)=q,\quad \mathscr{C}\stackrel{\triangle}{=}\big[\,B\ AB\ \ldots\ A^{q-1}B\,\big]\in\mathbb{R}^{q\times qk}. \tag{A.272}$$

(d) The pair of matrices (C,A) *is called observable if* (A^T,C^T) *is controllable, i.e., if*

$$rank\Big(\mathscr{O}\Big)=q,\quad \mathscr{O}\stackrel{\triangle}{=}\begin{bmatrix}C\\CA\\\vdots\\CA^{q-1}\end{bmatrix}\in\mathbb{R}^{pq\times q}. \tag{A.273}$$

Note that (A,B) controllable pair implies (A,B) stabilizable pair, and (C,A) observable pair implies (C,A) detectable pair.

The next theorem relates covariances of time-invariant finite dimensional Gaussian systems (A.268)-(A.271) to Lyapunov equations. It is borrowed from [5].

Theorem 27 (Properties of Lyapunov Equations [5]): Consider the covariance function $\Sigma:\{0,1,\ldots,n\}\to\mathbb{R}^{q\times q}$ *of the process* $\{X_i:i=0,\ldots,n\}$ *satisfying the recursion*

$$\Sigma_i=A\Sigma_{i-1}A^T+BK_WB^T,\quad \Sigma_0=given,\quad i=1,\ldots,n. \tag{A.274}$$

Consider the discrete Lyapunov equation for $\Sigma\in\mathbb{R}^{q\times q}$*:*

$$\Sigma=A\Sigma A^T+BK_WB^T. \tag{A.275}$$

The following hold.

(a) *If A is an exponentially stable matrix then $\lim_{i\longrightarrow\infty}\Sigma_i = \Sigma$ exists and Σ is a solution of the equation (A.275) (irrespectively of initial condition).*

(b) *If A is an exponentially stable matrix then (A.275) has an unique solution, which satisfies $\Sigma = \Sigma^T \succeq 0$.*

(c) *Let $r \in \{1, 2, \ldots, \}$, $G \in \mathbb{R}^{q\times r}$ be such that $BK_WB^T = GG^T$. Assume that $\{A, G\}$ is a stabilizable pair and there exists a $\Sigma \in \mathbb{R}^{q\times q}$ which satisfies*

$$\Sigma = A\Sigma A^T + BK_WB^T, \quad \text{and } \Sigma = \Sigma^T \succeq 0. \tag{A.276}$$

Then A is an exponentially stable matrix.

(d) *Let $\Sigma \in \mathbb{R}^{q\times q}$ be a solution of (A.275). Any two of the following three statements implies the third:*

1) *A is an exponentially stable matrix ($spec(A) \subset \mathbb{D}_o$);*
2) *$\big(A, G\big)$ is a controllable pair ($Rank(\mathscr{C}) = q$);*
3) *$\Sigma \succ 0$.*

Note that if the initial condition of (A.274) is set to $\Sigma_0 = \Sigma$, where Σ is a solution of (A.275), then $\Sigma_i = \Sigma, i = 1, 2, \ldots, n$, that is, the solution of the discrete recursion (A.274) is stationary.

Consider the problem of estimating $\{X_i : i = 0, \ldots, \}$ from $\{Y_i : i = 0, 1, \ldots, \}$, for the time-invariant finite dimensional Gaussian system (A.268)-(A.271), with respect to the following criterion.

$$\inf_{g_i(\cdot):i=0,\ldots,n} \mathbf{E}\Big\{||X_i - g_i(Y^{i-1})||^2_{\mathbb{R}^q}\Big\}, \quad \text{where } g_i(\cdot) \text{ is a measurable function of } y^{i-1}, \quad i = 0, \ldots, n. \tag{A.277}$$

Then the optimal estimator exists, it is unique, and it is given by the conditional expectation

$$g_i^*(y^{i-1}) = \mathbf{E}\big\{X_i|y^{i-1}\big\} = \int x\mathbf{P}(dx|y^{i-1}), \quad i = 0, \ldots, n.$$

The conditional distribution $\big\{\mathbf{P}(dx|y^{i-1}) : i = 0, \ldots, n\big\}$ is finite dimensional, and it is described by only two statistics, the conditional mean and the conditional covariance, defined by

$$\widehat{X}_{i|i-1} \stackrel{\triangle}{=} \mathbf{E}\Big\{X_i|Y^{i-1}\Big\},$$
$$Q_{i|i-1} \stackrel{\triangle}{=} \mathbf{E}\Big\{\Big(X_i - \widehat{X}_{i|i-1}\Big)\Big(X_i - \widehat{X}_{i|i-1}\Big)^T\Big|Y^{i-1}\Big\}, \quad i = 0, \ldots, n.$$

The conditional covariance is independent of the data and it is equal to the unconditional covariance,

$$Q_{i|i-1} = \mathbf{E}\Big\{\Big(X_i - \widehat{X}_{i|i-1}\Big)\Big(X_i - \widehat{X}_{i|i-1}\Big)^T\Big\} \quad i = 0, \ldots, n.$$

Moreover, $\big\{\widehat{X}_{i|i-1} : i = 0, \ldots, n\big\}$ satisfies a recursive equation known as the Kalman-filter equation, and $\big\{Q_{i|i-1} : i = 0, \ldots, n\big\}$ satisfies a recursive equation, known as the filtering Riccati difference matrix equation.

The properties of the Kalman-filter, such as, the convergence of the covariance (of the error) and the existence of invariant conditional distribution are determined from the properties of Riccati difference and algebraic equations.

The following theorem is borrowed from [5]; it summarizes properties of matrix Riccati difference and algebraic equations.

Theorem 28 (Properties of Riccati Equations [5]): Assume $NK_VN^T \succ 0$. Let $f : \mathbb{R}^{q\times q} \to \mathbb{R}^{q\times q}, G \in \mathbb{R}^{q\times q}$ be defined by

$$f(Q) \stackrel{\triangle}{=} AQA^T + BK_WB^T - AQC^T\big[CQC^T + NK_VN^T\big]^{-1}\big(AQC^T\big)^T, \quad GG^T \stackrel{\triangle}{=} BK_WB^T. \tag{A.278}$$

Let $F : \mathbb{R}^{q\times q} \to \mathbb{R}^{q\times p}, Q \mapsto F(Q)$, and $A : \mathbb{R}^{q\times q} \to \mathbb{R}^{q\times q}, A \mapsto A(Q)$ be defined by

$$F(Q) \stackrel{\triangle}{=} AQC^T\big[CQC^T + NK_VN^T\big]^{-1}, \quad A(Q) = A - F(Q)C \tag{A.279}$$

Define the discrete-time Riccati recursion for $Q : \{0, 1, \ldots, n\} \to \mathbb{R}^{q\times q}$ by

$$Q_{i+1} = f(Q_i), \quad Q_0 = given, \ i = 1, \ldots, n. \tag{A.280}$$

and the algebraic Riccati equation for the matrix $Q \in \mathbb{R}^{q\times q}$:

$$Q = f(Q). \tag{A.281}$$

The following hold.

(a) If (C, A) is a detectable pair and (A, G) is a stabilizable pair, then there exists a positive semidefinite solution $Q \in \mathbb{R}^{q\times q}$ to the algebraic Riccati equation

$$Q = f(Q), \quad Q = Q^T \succeq 0. \tag{A.282}$$

(b) If (A, G) is a stabilizable pair then the algebraic Riccati equation (A.282) has at most one solution.

(c) Under the assumptions of (a) the limit $\lim_{n\longrightarrow\infty} Q_i = Q$ exists and Q is the positive semidefinite solution of the algebraic Riccati equation (A.281).

(d) If (A, G) is a stabilizable pair and if there exists a positive semidefinite solution Q to the algebraic Riccati equation (A.281), then $spec\big(A(Q)\big) \subset \mathbb{D}_o$.

(e) Consider the algebraic Riccati equation for $Q \in \mathbb{R}^{q\times q}$ given by (A.281), with the conditions that $CQC^T + NK_VN^T \succ 0$ and $spec\big(A(Q)\big) \subset \mathbb{D}_o$ (but without the condition that $Q = Q^T \succeq 0$). The algebraic Riccati equation with these conditions has at most one solution $Q \in \mathbb{R}^{q\times q}$.

(f) Assume (A, G) is a controllable pair and that there exists a $Q \in \mathbb{R}^{q\times q}$ such that $Q = f(Q)$ and $Q = Q^T \succeq 0$. Then $Q \succ 0$.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous reviewers for many helpful comments and suggestions.

REFERENCES

[1] C. K. Kourtellaris and C. D. Charalambous, "Information structures of capacity achieving distributions for feedback channels with memory and transmission cost: Stochastic optimal control & variational equalities," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 4962–4992, Jul. 2018.

[2] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inf. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.

[3] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.

[4] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 986.
[5] J. H. van Schuppen, *Mathematical Control and System Theory of Discrete-Time Stochastic Systems*. Lecture Notes, 2010.
[6] P. M. Ebert, "The capacity of the Gaussian channel with feedback," *Bell Syst. Tech. J.*, vol. 49, no. 8, pp. 1705–1712, Oct. 1970.
[7] S. Ihara, *Information Theory for Continuous Systems*. Singapore: World Scientific, 1993.
[8] S. Butman, "Linear feedback rate bounds for regressive channels," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 3, pp. 363–366, May 1976.
[9] 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.
[10] J. P. M. 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.
[11] C. D. Charalambous, C. Kourtellaris, I. Tzortzis, and S. Loyka, "The decentralized structures of capacity achieving distributions of channels with memory and feedback," in *Proc. Int. Zurich Seminar Inf. Commun. (IZS)*, Feb. 2018, pp. 84–88.
[12] 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.
[13] 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, Jun. 2014.
[14] H. H. Permuter, H. Asnani, and T. Weissman, "Capacity of a post channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6041–6057, Oct. 2014.
[15] S. Yang, A. Kavčić, and S. Tatikonda, "Feedback capacity of finite-state machine channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.
[16] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–798, Mar. 2005.
[17] P. A. Stavrou, C. D. Charalambous, and C. K. Kourtellaris, "Sequential necessary and sufficient conditions for capacity achieving distributions of channels with memory and feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7095–7115, Nov. 2017.
[18] C. K. Kourtellaris and C. D. Charalambous, "Capacity of binary state symmetric channel with and without feedback and transmission cost," in *Proc. IEEE Inf. Theory Workshop (ITW)*, May 2015, pp. 1–5.
[19] C. K. Kourtellaris, C. D. Charalambous, and J. Boutros, "Nonanticipative transmission for sources and channels with memory," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 521–525.
[20] R. T. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.
[21] R. E. Blahut, *Principles and Practice of Information Theory* (Electrical and Computer Engineering). Reading, MA, USA: Addison-Wesley, 1987.
[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[23] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
[24] G. Kramer, "Directed information for channels with feedback," Ph.D. dissertation, ETH Zurich, Zürich, Switzerland, Dec. 1998.
[25] T. S. Han, *Information-Spectrum Methods in Information Theory*, 2nd ed. Berlin, Germany: Springer-Verlag, 2003.
[26] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[27] S. C. Tatikonda, "Control under communication constraints," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, USA, 2000.
[28] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
[29] 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.
[30] A. El Gamal and H. Y. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, Dec. 2011.
[31] C. D. Charalambous and P. A. Stavrou, "Directed information on abstract spaces: Properties and variational equalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6019–6052, Nov. 2016.
[32] M. Pinsker, *Information and Information Stability of Random Variables and Processes*, A. Feinstein, Ed. San Francisco, CA, USA: Holden-Day, 1964.
[33] J. L. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl. (ISITA)*, Nov. 1990, pp. 303–305.
[34] D. G. Luenberger, *Optimization by Vector Space Methods*. New York, NY, USA: Wiley, 1969.
[35] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*. Berlin, Germany: Springer-Verlag, 1979.
[36] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.
[37] E. Teletar, "Capacity of multi-antenna Gaussian channels," *Trans. Emerg. Telecommun. Technol.*, vol. 10, no. 6, pp. 585–595, 1999.
[38] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria* (Applications of Mathematics Stochastic Modelling and Applied Probability), vol. 1. Berlin, Germany: Springer-Verlag, 1996.
[39] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory.*, vol. 54, no. 4, pp. 1488–1499, Apr. 2008.
[40] S.-W. Ho and R. W. Yeung, "On the discontinuity of the Shannon information measures," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5362–5374, Dec. 2009.
[41] P. E. Caines, *Linear Stochastic Systems* (Wiley Series in Probability and Statistics). New York, NY, USA: Wiley, 1988.

Charalambos D. Charalambous received his B.S., M.E., and Ph.D. in 1987,1988, and 1992, respectively, all from the Department of Electrical Engineering, Old Dominion University, Virginia, USA. In 2003 he joined the Department of Electrical and Computer Engineering, University of Cyprus. He was an Associate Professor at University of Ottawa, from 1999 to 2003. He served on the faculty of McGill University, Department of Electrical and Computer Engineering, as a non-tenure faculty member, from 1995 to 1999. From 1993 to 1995 he was a post-doctoral fellow at Idaho State University. He is currently an associate editor of the journals, *Mathematics of Control, Signals, and Systems*, *Systems and Control Letters*, and *Nonlinear Analysis Hybrid Systems*. In the past he served as an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and IEEE COMMUNICATIONS LETTERS. Charalambous' research spans, Stochastic dynamical decision and control systems, information theory, optimization subject to ambiguity, stochastic dynamic games, decentralized decision systems.

Christos K. Kourtellaris received his Diploma in Electrical and Computer Engineering from Aristotle University of Thessaloniki in 2005, his M.Sc. degree in communications and signal processing in 2006 from Bristol University, and his Ph.D. degree in 2014 from University of Cyprus. He served as post-doctoral researcher at the department of Electrical and Computer Engineering at Texas A&M University, Qatar, and currently he is a post-doctoral researcher at the Department of Electrical and Computer Engineering at University of Cyprus. His research focuses on information theory, coding theory, control for communication applications, stochastic control, communication networks and game theory.

Sergey Loyka was born in Minsk, Belarus. He received the Ph.D. degree in Radio Engineering from the Belorussian State University of Informatics and Radioelectronics (BSUIR), Minsk, Belarus in 1995 and the M.S. degree with honors from Minsk Radioengineering Institute, Minsk, Belarus in 1992. Since 2001 he has been a faculty member at the School of Electrical Engineering and Computer Science, University of Ottawa, Canada. Prior to that, he was a research fellow in the Laboratory of Communications and Integrated Microelectronics (LACIME) of Ecole de Technologie Superieure, Montreal, Canada; a senior scientist at the Electromagnetic Compatibility Laboratory of BSUIR, Belarus; an invited scientist at the Laboratory of Electromagnetism and Acoustic (LEMA), Swiss Federal Institute of Technology, Lausanne, Switzerland. His research areas are wireless communications and networks and, in particular, MIMO systems and security aspects of such systems, in which he has published extensively. He received a number of awards from the URSI, the IEEE, the Swiss, Belarus and former USSR governments, and the Soros Foundation.