Kalman meets Shannon

Ather Gattami

Ericsson Research
Färögatan 6, Stockholm, Sweden.
E-mail: ather.gattami@ericsson.com

Abstract: We consider the problem of communicating the state of a dynamical system via a Shannon Gaussian channel with a given power constraint and no feedback. The transmitter observes a possibly noisy measurement of the state. These measurements are then used to encode the message to be transmitted over a noisy Gaussian channel, where a power constraint is imposed on the transmitted message. The receiver, which acts as both a decoder and estimator, observes the noisy measurement of the channel output and makes an optimal estimate of the state of the dynamical system in the minimum mean square sense. Thus, we get a mixed problem of Shannon's source-channel coding problem and a sort of Kalman filtering problem. We show that optimal encoders and decoders are linear filters with a finite memory and we give explicitly the state space realization of the optimal filters. We also present the solution of the case where the transmitter has access to noisy measurements of the state where we derive a separation principle for this communication scheme. Finally, we give necessary and sufficient conditions for the existence of a stationary solution.

NOTATION

A_i	Denotes the ith row of the matrix A.
x^t	$x^t = (x(0), x(1), ..., x(t))$.
$\mathbb{L}$	The set of lower triangular matrices.
$\mathbf{B}$	Denotes the backward shift operator, $x(t-1) = \mathbf{B}x(t)$.
$\mathbf{E}\{\cdot\}$	$\mathbf{E}\{x\}$ denotes the expected value of the stochastic variable x.
$\mathbf{E}\{\cdot|\cdot\}$	$\mathbf{E}\{x|y\}$ denotes the expected value of the stochastic variable x given y.
$\mathbf{cov}$	$\mathbf{cov}(x,y) = \mathbf{E}\{xy^{\intercal}\}$.
$h(x)$	Denotes the entropy of x.
$h(x|y)$	Denotes the entropy of x given y.
$I(x;y)$	Denotes the mutual information between x and y.
$\mathcal{N}(m,V)$	Denotes the set of Gaussian variables with mean m and covariance V.

1. INTRODUCTION

1.1 Background

Shannon [1948, 1949] considered the problem of reliable communication of a one-dimensional source over a one-dimensional Gaussian channel. In particular, Shannon considered the following coding-decoding setting for an analog Gaussian channel:

$$\inf_{\substack{f:\mathbb{R}\to\mathbb{R} \\ g:\mathbb{R}\to\mathbb{R} \\ \mathbf{E}|g(x)|^2 \le P}} \mathbf{E}|x - f(g(x)+n)|^2$$
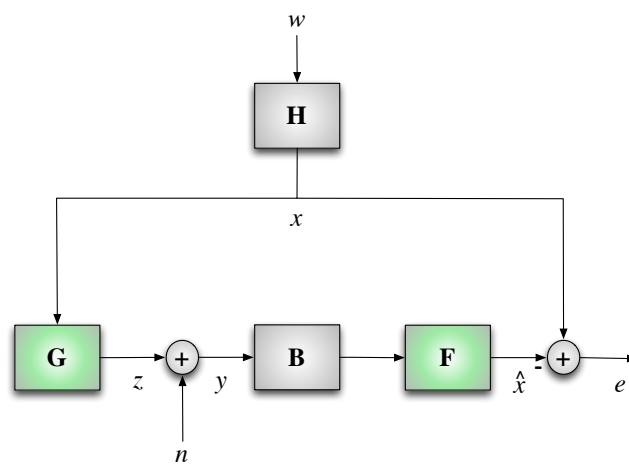

Fig. 1. A simple model of filtering problem over a Gaussian communications channel with Gaussian noise $n \sim \mathcal{N}(0,N)$ and delay given by the backward shift operator $\mathbf{B}$. The samples of the signal z are power limited with $\mathbf{E}|z(t)|^2 \le P$.

where $x \sim \mathcal{N}(0,X)$ and $n \sim \mathcal{N}(0,N)$. Shannon showed that the infimum can be attained by using linear encoder and decoder g and f, respectively .

However, the multi-input/multi-output problems or the problem when the message x to be estimated is given by a linear dynamical system driven by process noise are still open. For instance, the latter problem arises in video-streaming over a wireless channel. A video stream consists of highly correlated information described by a

dynamical system due to the correlation between the sequential picture frames.

More specifically, consider the block-diagram in Fig. 1. We have the process noise given by w, which is assumed to be Gaussian white noise, and the state is given by $x = \mathbf{H}w$ where $\mathbf{H}$ is a causal linear operator/filter.

The precoder is given by the causal operator $\mathbf{G}$, not necessarily linear. The encoded signal $z = \mathbf{G}x$ is then transmitted over a Gaussian channel with white noise given by n. Typically, one has power constraints on the transmitted signal $z(t)$, that is $\mathbf{E}|z(t)|^2 \leq P$, for some positive real number P. At the other end, the message received is $y(t) = z(t) + n(t)$, for $t = 0, ..., T-1$, and is delayed with one time step by the backward shift operator $\mathbf{B}$. Finally, the causal operator $\mathbf{F}$ is the decoder, designed to reconstruct the state x by $\hat{x} = \mathbf{F}\mathbf{B}y$, to minimize the mean squared error $\mathbf{E}|e|^2 = \mathbf{E}|x - \hat{x}|^2$.

For the case where $\mathbf{G}$ is a fixed linear operator, the optimal filter $\mathbf{F}$ is well known to be given by the optimal Kalman filter, which is a linear operator. However, if $\mathbf{G}$ is a precoder to be co-designed together with $\mathbf{F}$, we get a nonconvex problem. To this date, it's not known if the linear filters are optimal, and whether the order of the linear optimal filters is finite.

1.2 Previous work

Kalman [1960] made a fundamental contribution to optimal control and filtering of linear dynamical systems by deriving recursive state space solutions. The model considered by Kalman assumes linear measurements of the state, possibly partial and corrupted by noise. The role of a communication channel *with feedback* and its affect on stability was studied in Tatikonda et al. [2004] and necessary conditions for stability were given. Fundamental limitations of performance were studied in Martins and Dahleh [2008]. In Charalambous and Farhadi [2008], the capacity was matched to achieve a certain distortion, which required a time-varying power constraint that cannot be fixed beforehand. The problem of *linear* communication and filtering over a noisy channel for the stationary case has been considered in Johannesson et al. [2010] where it was shown that this problem can be transformed to a convex optimization problem that grows with the size of the time horizon. However, the order of the linear optimal filters obtained from Johannesson et al. [2010] is infinite

1.3 Contribution

We consider the linear dynamical system $\mathbf{H}$ given by
$$x(t+1) = ax(t) + bw(t), \qquad x(0) = 0, \quad 0 \leq t \leq T-1.$$
The main contributions of this paper are to show that the optimal filters $\mathbf{F}$ and $\mathbf{G}$ in the communication scheme without feedback, as described in Fig. 1, are linear and have a finite memory independent of the size of the time horizon. We also show explicitly that the state space realizations of the optimal filters are given by

$$\mathbf{G} : \begin{cases} s(t+1) = as(t) + K(t)\check{x}(t) \\ \quad \check{x}(t) = -s(t) + x(t) \\ \quad z(t) = \dfrac{\sqrt{P}}{\sigma_t}\check{x}(t) \end{cases}$$

$$\mathbf{F} : \qquad \hat{x}(t+1) = a\hat{x}(t) + K(t)y(t)$$

with $\sigma_t^2 = \mathbf{E}|\check{x}(t)|^2$, $K(t) = a\sigma_t\sqrt{P}(P+N)^{-1}$, $\check{x}(0) = 0$, and $s(0) = 0$. The interpretation of the state space equations is the following. $s(t) = \mathbf{E}\{\hat{x}(t)|x^t\}$ is the estimate at the transmitter of the estimate $\hat{x}(t)$ at the decoder. The transmitter's estimate of $e(t)$ is $\check{x}(t) = \mathbf{E}\{e(t)|x^t\} = x(t) - s(t)$. This estimate is then transmitted over the Gaussian channel, in order to supply the decoder with the innovations(the incremental information the decoder needs to correct its estimate of $x(t)$).

2. PROBLEM FORMULATION

2.1 Optimal Filtering over Noisy Communication Channel

Let $\mathbf{H}$ be a first order linear time invariant dynamical system with state-space realization

$$x(t+1) = ax(t) + bw(t), \qquad x(0) = 0, \quad 0 \leq t \leq T-1, \tag{1}$$

where $a, b \in \mathbb{R}$ and w is assumed to be white Gaussian noise with $w(t) \sim \mathcal{N}(0,1)$ for all $0 \leq t \leq T-1$.

The precoder is a map $\mathbf{G} : x^t \mapsto z(t)$, where z is the signal transmitted over the Gaussian channel. We have a power constraint on the transmitted signal $z(t)$ given by $\mathbf{E}|z(t)|^2 \leq P$.

The measurements at the decoder are given by $y(0) := 0$ and
$$y(t) = z(t) + n(t), \qquad \text{for } t \geq 1,$$
where n is a Gaussian white noise process with $n(t) \sim \mathcal{N}(0,N)$. The decoder is a map $\mathbf{F} : y^{t-1} \mapsto \hat{x}(t)$.

The objective is to design causal precoder and decoder maps $\mathbf{G} : x^t \mapsto z(t)$ and $\mathbf{F} : y^{t-1} \mapsto \hat{x}(t)$, respectively, such that the average of the mean squared error is minimized:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}|x(t) - \hat{x}(t)|^2 \to \min.$$

The precoder and decoder maps can be equivalently written as a causal dynamical system according to

$$\begin{aligned} z(t) &= g_t(x^t) \\ \hat{x}(t) &= f_t(y^{t-1}), \end{aligned} \tag{2}$$

where g_t is the precoder and f_t is the decoder.

Now we may formalize our first problem statement:

Problem 1. Consider the linear dynamical system
$$x(t+1) = ax(t) + bw(t), \qquad x(0) = 0, \quad 0 \leq t \leq T-1,$$
where $a, b \in \mathbb{R}$ and $w(t) \sim \mathcal{N}(0,1)$ for $0 \leq t \leq T-1$. Let n be a Gaussian white noise process independent of w, with $n(t) \sim \mathcal{N}(0,N)$. Find an optimal precoder and decoder pair (2) such that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}|x(t) - \hat{x}(t)|^2 \to \min,$$

where $y(0) = 0$ and $y(t) = z(t) + n(t)$, for $t \geq 1$.

2.2 Linear optimal precoder/decoder design

The linear filter $\mathbf{H}$ has the following Toeplitz matrix representation over the time $t = 1, ..., T$:

$$
\begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(T) \end{bmatrix} = \begin{bmatrix} b & 0 & 0 & \cdots & 0 \\ ab & b & 0 & \cdots & 0 \\ a^2b & ab & b & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a^Tb & a^{T-1}b & a^{T-2}b & \cdots & b \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ w(2) \\ \vdots \\ w(T-1) \end{bmatrix} \quad (3)
$$

Let the precoder $\mathbf{G}$ be a causal linear filter that maps x to z:

$$
\begin{bmatrix} z(1) \\ z(2) \\ z(3) \\ \vdots \\ z(T) \end{bmatrix} = \begin{bmatrix} G_{11} & 0 & 0 & \cdots & 0 \\ G_{21} & G_{22} & 0 & \cdots & 0 \\ G_{31} & G_{32} & G_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{T1} & G_{T2} & G_{T3} & \cdots & G_{TT} \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(T) \end{bmatrix} \quad (4)
$$

The precoder is subject to a power constraint on its output signal $z = \mathbf{G}x$ given by $\mathbf{E}|z(t)|^2 \leq P$, for $t = 0, ..., T$.

The decoder $\mathbf{F}$ is a causal linear filter that observes the delayed measurements with $y(0) := 0$ and $y(t) = z(t) + n(t)$ for $t \geq 1$. It has the following linear operator representation:

$$
\begin{bmatrix} \hat{x}(1) \\ \hat{x}(2) \\ \hat{x}(3) \\ \vdots \\ \hat{x}(T) \end{bmatrix} = \begin{bmatrix} F_{11} & 0 & 0 & \cdots & 0 \\ F_{21} & F_{22} & 0 & \cdots & 0 \\ F_{31} & F_{32} & F_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{T1} & F_{T2} & F_{T3} & \cdots & F_{TT} \end{bmatrix} \begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(T-1) \end{bmatrix} \quad (5)
$$

The output $\hat{x} = \mathbf{FB}y$ is the optimal estimate of x in the sense that the average of the mean squared error, is minimized:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbf{E}|x(t) - \hat{x}(t)|^2.
$$

Now let

$$
H = \begin{bmatrix} b & 0 & 0 & \cdots & 0 \\ ab & b & 0 & \cdots & 0 \\ a^2b & ab & b & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a^Tb & a^{T-1}b & a^{T-2}b & \cdots & b \end{bmatrix}, \quad (6)
$$

$$
G = \begin{bmatrix} G_{11} & 0 & 0 & \cdots & 0 \\ G_{21} & G_{22} & 0 & \cdots & 0 \\ G_{31} & G_{32} & G_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{T1} & G_{T2} & G_{T3} & \cdots & G_{TT} \end{bmatrix}, \quad (7)
$$

$$
F = \begin{bmatrix} F_{11} & 0 & 0 & \cdots & 0 \\ F_{21} & F_{22} & 0 & \cdots & 0 \\ F_{31} & F_{32} & F_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_{T1} & F_{T2} & F_{T3} & \cdots & F_{TT} \end{bmatrix}, \quad (8)
$$

$$
x = \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(T) \end{bmatrix}, \quad w = \begin{bmatrix} w(0) \\ w(1) \\ w(2) \\ \vdots \\ w(T-1) \end{bmatrix}, \quad z = \begin{bmatrix} z(1) \\ z(2) \\ z(3) \\ \vdots \\ z(T) \end{bmatrix},
$$

$$
\hat{x} = \begin{bmatrix} \hat{x}(1) \\ \hat{x}(2) \\ \hat{x}(3) \\ \vdots \\ \hat{x}(T) \end{bmatrix}, \quad n = \begin{bmatrix} n(0) \\ n(1) \\ n(2) \\ \vdots \\ n(T-1) \end{bmatrix}, \quad y = \begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ y(T-1) \end{bmatrix}.
$$

Then,

$$
x = Hw, \qquad z = GHw, \qquad y = GHw + n, \qquad \hat{x} = Fy,
$$

$$
\sum_{t=1}^{T} \mathbf{E}|x(t) - \hat{x}(t)|^2 = \mathbf{E}|x - \hat{x}|^2 = \mathbf{E}|Hw - Fy|^2
$$

After some algebra, the least mean square error for a linear precoder and decoder will be given by

$$
\inf_{\substack{G,F \in \mathbb{L} \\ G_t H H^* G_t^* \leq P}} \mathbf{E}|Hw - F(GHw + n)|^2 \quad (9)
$$

Note that the optimization problem above is inherently non-convex, since we have a coupling term between G and F in the quadratic objective function.

3. MAIN RESULTS

3.1 Finite-Horizon Filtering problem

The first result of this paper presents the structure of the optimal precoder and decoder:

Theorem 1. The optimal communication scheme to Problem 1 is given by

$$
\begin{aligned}
\hat{x}(t) &= \mathbf{E}\{x(t)|y^{t-1}\} \\
\tilde{x}(t) &= x(t) - \hat{x}(t) \\
\check{x}(t) &= \mathbf{E}\{\tilde{x}(t)|x^t\} \quad (10) \\
z(t) &= \frac{\sqrt{P}}{\sigma_t}\check{x}(t),
\end{aligned}
$$

where $\sigma_t^2 = \mathbf{E}|\check{x}(t)|^2$, for $t = 1, ..., T$.

Proof. See the Appendix. ∎

Theorem 2. The state space realization of the optimal communication scheme solution of Problem 1 is given by

$$
\begin{aligned}
\hat{x}(t+1) &= a\hat{x}(t) + K(t)y(t) \\
s(t+1) &= as(t) + K(t)\check{x}(t) \\
\check{x}(t) &= x(t) - s(t) \quad (11) \\
z(t) &= \frac{\sqrt{P}}{\sigma_t}\check{x}(t),
\end{aligned}
$$

where $s(0) = 0$, $V_{ss}(0) = V_{sx}(0) = V_{xx}(0) = 0$,

$$
K(t) = a\sigma_t\sqrt{P}(P + N)^{-1} \quad (12)
$$

$$
\sigma_t^2 = V_{xx}(t) - 2V_{sx}(t) + V_{ss}(t) \quad (13)
$$

$$
\begin{aligned}
&\begin{bmatrix} V_{ss}(t+1) & V_{sx}(t+1) \\ V_{xs}(t+1) & V_{xx}(t+1) \end{bmatrix} = \\
&\begin{bmatrix} a - K(t) & K(t) \\ 0 & a \end{bmatrix} \begin{bmatrix} V_{ss}(t) & V_{sx}(t) \\ V_{xs}(t) & V_{xx}(t) \end{bmatrix} \begin{bmatrix} a - K(t) & K(t) \\ 0 & a \end{bmatrix}^{\mathsf{T}} + \begin{bmatrix} 0 & 0 \\ 0 & b^2 \end{bmatrix}.
\end{aligned} \quad (14)
$$

Proof. See the Appendix. ∎

3.2 Separation Principle for Optimal Communication

Consider the linear system

$$x(t + 1) = ax(t) + bw(t)$$
$$\gamma(t) = cx(k) + dv(t)$$

for $0 \leq t \leq T - 1$, with $x(0) = 0$ and $v(t)$ is Gaussian white that might or might not be correlated with $w(t)$. We assume now that the transmitter doesn't have access to the state $x(t)$ but $\gamma(t)$ instead. The optimal transmission scheme is for the transmitter to find the best estimate of $x(t)$ based on γ^t, namely $\breve{x}(t) = \mathbf{E}\{x(t)|\gamma^t\}$.

To see this, let

$$\xi(t) = x(t) - \breve{x}(t)$$

be the estimation error. It's well known that $\breve{x}(t)$ is given by the Kalman filter

$$\breve{x}(t + 1) = a\breve{x}(t) + L(t)(c\xi(t) + dv(t)) \tag{15}$$

where $L(t)$ are the optimal Kalman filter gains for $t = 0, ..., T$ (see, e. g., Åström [1970]). We also know that γ^t and $\xi(t)$ are uncorrelated(see Proposition 3 in the Appendix). This implies in turn that y^t and $\xi(t)$ are uncorrelated. Hence, the averaged estimation error of the decoder is equal to

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}|x(t) - \hat{x}(t)|^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{E}|\breve{x}(t) - \hat{x}(t)|^2 + \mathbf{E}|\xi(t)|^2\right).$$

Obviously, the decoder can't do much about the error covariance $\mathbf{E}|\xi(t)|^2$. The decoder $\mathbf{F}$ minimizes the averaged estimation error above if and only if it minimizes the averaged estimation error of $\breve{x}(t)$. Thus, we have transformed the output measurement problem to a state measurement problem at the encoder $\mathbf{G}$, where the measured states are the states of the dynamical system given by (15).

3.3 Stationarity

In this section, we will present conditions under which a stationary solution exists to Problem 1 (that is a solution as $T \to \infty$). Let $\tilde{x}(t) = x(t) - \hat{x}(t)$ be the estimation error of $x(t)$ and consider the state space equations (11) of the optimal estimate. After some algebra, we get the state space equations for the estimation error (see the proof of Theorem 2 in the Appendix):

$$\tilde{x}(t + 1) = \left(a - K(t)\frac{\sqrt{P}}{\sigma_t}\right)\tilde{x}(t) - K(t)n(t) + bw(t)$$
$$+ K(t)\frac{\sqrt{P}}{\sigma_t}\bar{x}(t)$$

with

$$\bar{x}(t + 1) = a\bar{x}(t) - K(t)n(t).$$

Obviously, the state $\bar{x}(t)$ can be stationary if and only if $|a| < 1$. In addition, in order for $\tilde{x}(t)$ to be stationary, we must have

$$1 > \left|a - K(t)\frac{\sqrt{P}}{\sigma_t}\right| = \left|a - a \cdot \frac{N}{P + N}\right| = \frac{P}{P + N}|a|,$$

where we have used (41) in the first equality above. Clearly, the inequality is always fulfilled for $|a| < 1$. We conclude the result above:

Theorem *3.* Problem 1 has a solution as $T \to \infty$ if and only if $|a| < 1$.

4. CONCLUSIONS

We considered the problem of optimal encoder/decoder filter design over a Shannon Gaussian channel to estimate the state of a linear dynamical system. We showed that optimal encoders and decoders are linear filters with a finite memory and we give explicitly the state space realization of the optimal filters. We also presented the solution of the case where the transmitter has access to noisy measurements of the state. We derived a separation principle for this communication scheme. Necessary and sufficient conditions for the existence of a stationary solution where also given. The results were presented for first order dynamical systems to simplify the presentation of the paper but could be generalized to systems of arbitrary order and will be presented elsewhere because of space constraints.

REFERENCES

K. J. Åström. *Stochastic Control Theory*. Academic Press, 1970.

C. D. Charalambous and A. Farhadi. Lqg optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels. *Automatica*, 12(44):3181–3188, 2008.

T. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

R. G. Gallager. *Information theory and reliable communication*. Wiley, New York, 1968. ISBN 0471290483.

A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2012.

E. Johannesson, A. Rantzer, B. Bernhardsson, and A. Ghulchak. Encoder and decoder design for signal estimation. In *American Control Conference*, Baltimore, Maryland, USA, June 2010.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. of the ASME-Journal of Basic Engineering*, 82:35–45, 1960.

N.C. Martins and M.A. Dahleh. Feedback control in the presence of noisy channels: Fundamental limitations of performance. *Automatic Control, IEEE Transactions on*, 53(7):1604–1615, 2008.

C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423 and 623–656, 1948.

C. E. Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949.

A. N. Shiryaev. *Probability*. Springer, 1996.

S. Tatikonda, A. Sahai, and S. Mitter. Stochastic linear control over a communication channel. *IEEE Trans. on Automatic Control*, 49(9):1549–1561, 2004.

APPENDIX

Preliminaries

Definition *1.* The entropy of a real-valued stochastic variable X with probability distribution $p(x)$ is defined as

$$h(X) = -\int_{-\infty}^{\infty} p(x)\log_2 p(x)dx$$

Definition 2. For two real valued stochastic variables X and Y, the conditional entropy of X given Y is defined as
$$h(X|Y) = h(X,Y) - h(Y).$$

Definition 3. The mutual information between X and Y is defined as
$$I(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X).$$

Proposition 1. (Entropy Power Inequality). If X and Y are independent random variables, then
$$2^{2h(X+Y)} \geq 2^{2h(X)} + 2^{2h(Y)}$$
with equality if X and Y are Gaussian stochastic variables.

Proof. See Cover and Thomas [2006], p. 674 - 675. ∎

Definition 4. Random variables X, Y, Z are said to form a Markov chain in that order if the conditional distribution of Z depends only on Y and conditionally independent of X. This is denoted by $X \to Y \to Z$.

Proposition 2. (Data-Processing Inequality). If
$$X \to Y \to Z,$$
then
$$I(X;Z) \leq I(Y;Z).$$

Proof. See Cover and Thomas [2006], p. 34-35. ∎

Proposition 3. Let X and Y be two stochastic variables. The optimal solution to the optimization problem
$$\inf_{f(\cdot)} \mathbf{E}|X - f(Y)|^2$$
is unique and given by the expectation of X given Y
$$f_\star(Y) = \mathbf{E}\{X|Y\}.$$
Furthermore, $f_\star(Y)$ and $X - f_\star(Y)$ are uncorrelated.

Proof. Consult (Shiryaev [1996], p. 237). ∎

Proposition 4. Consider the stochastic variables X and Y, and let the estimation error of X based on Y be given by
$$\tilde{X} = X - \mathbf{E}\{X|Y\}.$$
Then,
$$\frac{1}{2} \log_2 \det\left(2\pi e \mathbf{E}\{\tilde{X}\tilde{X}^\intercal\}\right) \geq h(X|Y) = h(\tilde{X}) \quad (16)$$
with equality if and only if X and Y are jointly Gaussian.

Proof. Consult Gamal and Kim [2012], p. 21. ∎

Proof of Theorem 1

Suppose that $\mathbf{E}g_t(x^t) = \alpha_t$ where α_k, $k = 0,...,t$, are deterministic real numbers independent of x^t and are known at the encoder g_t and decoder f_t. Note that $y(t) = g_t(x^t) + n(t)$. The estimate of $x(t+1)$ based on $y(k)$, $k = 0,...,t$, is the same as the estimate of $x(t+1)$ based on $y(k) - \alpha_k$ for $k = 0,...,t$ since α_k is deterministic and known at the decoder. But it means that we can replace $g_t(x^t)$ with $g'_t(x^t) = g(x^t) - \alpha_t$, and $g'_t(x^t)$ satisfies both $\mathbf{E}g'_t(x^t) = 0$ and the power constraint $\mathbf{E}|g'_t(x^t)|^2 \leq P$ since
$$\mathbf{E}|g'_t(x^t)|^2 = \mathbf{E}|g_t(x^t) - \alpha|^2$$
$$= \mathbf{E}|g_t(x^t)|^2 - \alpha^2$$
$$= P - \alpha^2 \leq P.$$
Thus, without loss of generality, we may restrict the encoders g to the set $\{g \mid \mathbf{E}g(x^t) = 0\}$.

Consider time step $t = 0$, where the encoder has access to $x(0) = 0$. Obviously, $g_0(x(0)) = 0$. The decoder's output at $t = 1$ is $f_1(y(0)) = 0$, which gives $\tilde{x}(1) = x(1)$. Note that $\tilde{x}(1)$ is Gaussian, and both g_0 and f_1 are linear in their arguments. We will show by induction that the optimal g_{t-1} and f_t are linear and that $\tilde{x}(t)$ is Gaussian and for all integers $t \geq 0$.

Suppose that g_{t-1} and f_t are linear. Then, $\tilde{x}(t)$ and x^t are jointly Gaussian, for $t = 1,...,k$. Let $\hat{x}(t|t) = f'_t(y^t)$ be the optimal estimate of $x(t)$ based on y^t and let $\tilde{x}(t|t) = x(t) - \hat{x}(t|t)$, for $t = 0,...,T$. We have that $f'_t(y^t) = \mathbf{E}\{x(t)|y^t\}$ according to Proposition 3. Now we have that
$$\hat{x}(t|t) = \mathbf{E}\{x(t)|y^t\}$$
$$= \mathbf{E}\{(\hat{x}(t) + \tilde{x}(t)|y^t\} = \hat{x}(t) + \mathbf{E}\{\tilde{x}(t)|y^t\}, \quad (17)$$
$$\tilde{x}(t+1) = x(t+1) - \hat{x}(t+1)$$
$$= ax(t) + bw(t) - a\hat{x}(t|t) \quad (18)$$
$$= a\tilde{x}(t|t) + bw(t)$$
We see that minimizing $\mathbf{E}|\tilde{x}(t+1)|^2$ is equivalent to minimizing the mean square error of
$$\tilde{x}(t|t) = \tilde{x}(t) - \mathbf{E}\{\tilde{x}(t)|y^t\}$$
at the decoder. Now introduce
$$\check{x}(t) := \mathbf{E}\{\tilde{x}(t)|x^t\}$$
and
$$\bar{x}(t) := \tilde{x}(t) - \mathbf{E}\{\tilde{x}(t)|x^t\}.$$
Then, $\check{x}(t)$ is independent of $\bar{x}(t)$. Note that $\check{x}(t)$ is a linear function of x^t, since $\tilde{x}(t)$ and x^t are jointly Gaussian by the induction hypothesis. The Markov chain
$$\check{x}(t) \to x^t \to g_t(x^t) \to y(t) = g_t(x^t) + n(t),$$
together with Proposition 2, gives
$$I(\check{x}(t);y(t)) \leq I(g_t(x^t);y(t)). \quad (19)$$
The Shannon capacity of a Gaussian channel gives an upper bound for the mutual information between the transmitted message $z(t) = g_t(x^t)$ and received message $y(t)$ (see Gallager [1968]):
$$I(g_t(x^t);y(t)) \leq \frac{1}{2} \log_2 \left(1 + \frac{P}{N}\right). \quad (20)$$
Combining (19)-(20), we get
$$2^{-2I(\check{x}(t);y(t))} \geq \frac{N}{P+N} \quad (21)$$
with equality if $\check{x}(t)$ and $y(t)$ are mutually Gaussian and $g_t(x^t) = \frac{\sqrt{P}}{\sigma_t}\check{x}(t)$. From the definition of mutual information, we have that
$$h(\check{x}(t)|y(t)) = h(\check{x}(t)) - I(\check{x}(t);y(t)). \quad (22)$$
Now we get
$$2\pi e\mathbf{E}\{|\tilde{x}(t|t)|^2\} \geq 2^{2h(\tilde{x}(t)|y^t)} \quad (23)$$
$$= 2^{2h(\tilde{x}(t)|y(t))} \quad (24)$$
$$= 2^{2h(\check{x}(t)+\bar{x}(t)|y(t))} \quad (25)$$
$$\geq 2^{2h(\check{x}(t)|y(t))} + 2^{2h(\bar{x}(t)|y(t))} \quad (26)$$
$$= 2^{2h(\check{x}(t))-2I(\check{x}(t);y(t))} + 2^{2h(\bar{x}(t))} \quad (27)$$
$$\geq \frac{N}{P+N}2^{2h(\check{x}(t))} + 2^{2h(\bar{x}(t))} \quad (28)$$

where (23) follows from Proposition 4(with equality if $\tilde{x}(t)$ and y^t are jointly Gaussian), (24) follows from the fact that $\tilde{x}(t)$ is independent of y^{t-1}, (26) follows from the entropy power inequality(Proposition 1), (27) follows from equation (22) and the fact that $\tilde{x}(t)$ is independent of x^t(and thus independent of $y(t) = g_t(x^t) + n(t)$), and (28) follows from equation (21). Furthermore, equality holds in (23)-(28) if

$$z(t) = g_t(x^t) = \frac{\sqrt{P}}{\sigma_t}\check{x}(t)$$

with $\sigma_t^2 = \mathbf{E}|\check{x}(t)|^2$. This completes the proof.

Proof of Theorem 2

Let $\hat{x}(t) = \mathbf{E}\{x(t)|y^{t-1}\}$, $\tilde{x}(t) = x(t) - \hat{x}(t)$, $\hat{x}(t|t) = \mathbf{E}\{x(t)|y^t\}$, and $\tilde{x}(t|t) = x(t) - \hat{x}(t|t)$.

Then,

$$\begin{aligned}\hat{x}(t+1) &= a\hat{x}(t|t) \\ &= a\mathbf{E}\{\hat{x}(t) + \tilde{x}(t)|y^t\} \quad (29) \\ &= a\hat{x}(t) + a\mathbf{E}\{\tilde{x}(t)|y(t)\}\end{aligned}$$

and

$$\tilde{x}(t+1) = a\tilde{x}(t) - a\mathbf{E}\{\tilde{x}(t)|y(t)\} + bw(t). \quad (30)$$

According to Theorem 1, the optimal signal z is given by

$$\check{x}(t) = \mathbf{E}\{\tilde{x}(t)|x^t\}$$

$$z(t) = \frac{\sqrt{P}}{\sigma_t}\check{x}(t)$$

with $\sigma_t^2 = \mathbf{E}|\check{x}(t)|^2$. Now recall that $y(t) = z(t) + n(t)$, $\check{x}(t) = \mathbf{E}\{\tilde{x}(t)|x^t\}$ and $\bar{x}(t) = \tilde{x}(t) - \check{x}(t)$ and $\bar{x}(t)$ is orthogonal to x^t and hence to $y(t)$. Since $\check{x}(t)$ and $y(t)$ are jointly Gaussian, $\mathbf{E}\{\tilde{x}(t)|y(t)\}$ is a linear function of $y(t)$ given by

$$\begin{aligned}\mathbf{E}\{\tilde{x}(t)|y(t)\} &= \mathbf{E}\{\check{x}(t) + \bar{x}(t)|y(t)\} \\ &= \mathbf{E}\{\check{x}(t)|y(t)\} + \mathbf{E}\{\bar{x}(t)|y(t)\} \\ &= \mathbf{E}\{\check{x}(t)|y(t)\} \\ &= \mathbf{cov}(\check{x}(t), y(t))[\mathbf{cov}(y(t), y(t))]^{-1}y(t) \\ &= \kappa(t)y(t)\end{aligned}$$
$$(31)$$

with

$$\kappa(t) = \sigma_t\sqrt{P}(P+N)^{-1}. \quad (32)$$

Then, (29)-(31) imply

$$\begin{aligned}\hat{x}(t+1) &= a\hat{x}(t) + a\kappa(t)y(t) \\ &= a\hat{x}(t) + a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\check{x}(t) + a\kappa(t)n(t),\end{aligned} \quad (33)$$

$$\begin{aligned}\tilde{x}(t+1) &= a\tilde{x}(t) - a\kappa(t)y(t) + bw(t) \\ &= a\tilde{x}(t) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\check{x}(t) - a\kappa(t)n(t) + bw(t) \\ &= a\tilde{x}(t) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\tilde{x}(t) + a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\bar{x}(t) \\ &\quad - a\kappa(t)n(t) + bw(t).\end{aligned}$$
$$(34)$$

We will now show that

$$\check{x}(t+1) = a\check{x}(t) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\check{x}(t) + bw(t) \quad (35)$$

and

$$\bar{x}(t+1) = a\bar{x}(t) - a\kappa(t)n(t). \quad (36)$$

First we note that $\bar{x}$ as defined in (36) depends only on the channel noise n and is therefore independent of x^t and $\check{x}$. Also, we have that

$$\begin{aligned}\tilde{x}(t+1) &= \check{x}(t+1) + \bar{x}(t+1) \\ &= a\check{x}(t) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\check{x}(t) + bw(t) \\ &\quad - (a\bar{x}(t) - a\kappa(t)n(t)) \\ &= a(\check{x}(t) + \bar{x}(t)) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}(\tilde{x}(t) - \bar{x}(t)) \\ &\quad - a\kappa(t)n(t) + bw(t) \\ &= a\tilde{x}(t) - a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\tilde{x}(t) + a\kappa(t)\frac{\sqrt{P}}{\sigma_t}\bar{x}(t) \\ &\quad - a\kappa(t)n(t) + bw(t),\end{aligned}$$

which is exactly the expression given by (34). This establishes (35)-(36). Now we have

$$\begin{aligned}x(t) &= \mathbf{E}\{x(t)|x^t\} \\ &= \mathbf{E}\{\hat{x}(t) + \tilde{x}(t)|x^t\} \\ &= \mathbf{E}\{\hat{x}(t)|x^t\} + \mathbf{E}\{\tilde{x}(t)|x^t\} \\ &= \mathbf{E}\{\hat{x}(t)|x^t\} + \check{x}(t).\end{aligned}$$
$$(37)$$

From equation (33), we see that

$$\mathbf{E}\{\hat{x}(t)|x^t\} = s(t) \quad (38)$$

where

$$s(t+1) = as(t) + a\kappa(t)\check{x}(t+1), \quad (39)$$

since the noise signal n is independent of x^t. Finally, combining (37) - (39) gives

$$\check{x}(t) = x(t) - s(t). \quad (40)$$

Now set

$$K(t) = a\kappa(t) \quad (41)$$

Then,

$$\begin{bmatrix} s(t+1) \\ x(t+1) \end{bmatrix} = \begin{bmatrix} a - K(t) & K(t) \\ 0 & a \end{bmatrix}\begin{bmatrix} s(t) \\ x(t) \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix}w(t).$$

Introduce the covariance matrix

$$\begin{bmatrix} V_{ss}(t) & V_{sx}(t) \\ V_{xs}(t) & V_{xx}(t) \end{bmatrix} = \mathbf{E}\begin{bmatrix} s(t) \\ x(t) \end{bmatrix}\begin{bmatrix} s(t) \\ x(t) \end{bmatrix}^{\mathsf{T}}.$$

Since $w(t)$ is uncorrelated with $x(t)$ and $s(t)$, we get

$$\begin{aligned}&\begin{bmatrix} V_{ss}(t+1) & V_{sx}(t+1) \\ V_{xs}(t+1) & V_{xx}(t+1) \end{bmatrix} = \mathbf{E}\begin{bmatrix} s(t+1) \\ x(t+1) \end{bmatrix}\begin{bmatrix} s(t+1) \\ x(t+1) \end{bmatrix}^{\mathsf{T}} \\ &= \mathbf{E}\left\{\left(\begin{bmatrix} a-K(t) & K(t) \\ 0 & a \end{bmatrix}\begin{bmatrix} s(t) \\ x(t) \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix}w(t)\right) \times \right. \\ &\quad \left.\left(\begin{bmatrix} a-K(t) & K(t) \\ 0 & a \end{bmatrix}\begin{bmatrix} s(t) \\ x(t) \end{bmatrix} + \begin{bmatrix} 0 \\ b \end{bmatrix}w(t)\right)^{\mathsf{T}}\right\} \\ &= \begin{bmatrix} a-K(t) & K(t) \\ 0 & a \end{bmatrix}\begin{bmatrix} V_{ss}(t) & V_{sx}(t) \\ V_{xs}(t) & V_{xx}(t) \end{bmatrix}\begin{bmatrix} a-K(t) & K(t) \\ 0 & a \end{bmatrix}^{\mathsf{T}} \\ &\quad + \begin{bmatrix} 0 & 0 \\ 0 & b^2 \end{bmatrix}.\end{aligned}$$
$$(42)$$

Thus,

$$\begin{aligned}\sigma_t^2 = \mathbf{E}|\check{x}(t)|^2 &= \mathbf{E}|x(t) - s(t)|^2 \\ &= V_{xx}(t) - 2V_{sx}(t) + V_{xx}(t).\end{aligned} \quad (43)$$

Putting together (32), (33), (39) - (43) completes the proof.