Shannon meets Wiener II: On MMSE estimation in successive decoding schemes

G. David Forney, Jr.
MIT
Cambridge, MA 02139 USA
forneyd@comcast.net

Abstract

We continue to discuss why MMSE estimation arises in coding schemes that approach the capacity of linear Gaussian channels. Here we consider schemes that involve successive decoding, such as decision-feedback equalization or successive cancellation.

"Everything should be made as simple as possible, but not simpler."— A. Einstein.

1 Introduction

The occurrence of minimum-mean-squared-error (MMSE) linear estimation filters in constructive coding schemes that approach information-theoretic limits of linear Gaussian channels has been repeatedly observed, and justified by various arguments. For example, in an earlier paper [5] we showed the necessity of the MMSE estimation factor in the capacity-approaching lattice coding scheme of Erez and Zamir [3] for the classic additive white Gaussian noise (AWGN) channel.

In particular, MMSE decision-feedback equalizer (MMSE-DFE) filters have been used in coding schemes that approach the capacity of linear Gaussian intersymbol interference (ISI) channels [1], and generalized MMSE-DFE (MMSE-GDFE) filters have been used in coding schemes that approach the capacity region of multiple-input, multiple-output (MIMO) linear Gaussian channels [2]. These successive decoding schemes combine "analog" discrete-time linear MMSE estimation with the essentially "digital" assumption of ideal decision feedback (perfect prior decisions).

The fact that MMSE filters allow information-theoretic limits to be approached in successive decoding scenarios is widely understood, and has been proved in various ways. Our aim here is to provide the simplest and most transparent justification possible. Some principal features of our approach are:

- As in [2, 8], we use a geometric Hilbert space formulation;
- Our results are based mainly on the sufficiency property of MMSE estimators, with information-theoretic results mostly as corollaries;
- Proofs of almost all results are given. All proofs are brief and straightforward.

In developing this approach, we have benefited from our earlier work with Cioffi *et al.* [1, 2] and from the insightful development of Guess and Varanasi [6, 8]. We would also like to acknowledge helpful comments on earlier drafts of this paper by G. Caire, J. Cioffi, U. Erez, T. Guess, S. Shamai and G. Wornell.

1.1 Hilbert spaces of jointly Gaussian random variables

All random variables in this note will be finite-variance, zero-mean, proper (circularly symmetric) complex Gaussian random variables. Random variables will be denoted by capital letters such as X. If the variance σ^2 of X is nonzero, then X has a probability density function (pdf)

$$p_X(x) = \frac{1}{\pi\sigma^2} \exp -\frac{|x|^2}{\sigma^2},$$

and thus its differential entropy is $h(X) = \mathsf{E}[-\log p_X(x)] = \log \pi e \sigma^2$. If the variance of X is zero, then X is the deterministic zero variable 0.

Sets of such random variables will be denoted by a script letter such as $\mathcal{X} = \{X_i\}$. In this paper, we will consider only finite sets of random variables. A particular application may involve a finite set of such sets such as $\{\mathcal{X}, \mathcal{Y}, \mathcal{Z}\}$.

Whenever we have a set of Gaussian variables, their statistics will be assumed to be jointly Gaussian. A set of variables is jointly Gaussian if they can all be expressed as linear combinations of a common set of independent Gaussian random variables. It follows that any set of linear combinations of jointly Gaussian random variables is jointly Gaussian.

The set of all complex linear combinations of a given finite set $\mathcal{X}$ of finite-variance, zero-mean, proper jointly Gaussian complex random variables is evidently a complex vector space $\mathcal{G}$. Every element of $\mathcal{G}$ is a finite-variance, zero-mean, proper complex Gaussian random variable, and every subset of $\mathcal{G}$ is jointly Gaussian. The zero vector of $\mathcal{G}$ is the unique zero variable 0. The dimension of $\mathcal{G}$ is at most the size $|\mathcal{X}|$ of $\mathcal{X}$.

It is well known that if an inner product is defined on $\mathcal{G}$ as the cross-correlation $\langle X, Y \rangle = \mathsf{E}[XY^*]$, then $\mathcal{G}$ becomes a Hilbert space (a complete inner product space), a subspace of the Hilbert space $\mathcal{H}$ consisting of all finite-variance zero-mean complex random variables. The squared norm of $X \in \mathcal{G}$ is then its variance, $||X||^2 = \langle X, X \rangle = \mathsf{E}[|X|^2]$. Variances are real, finite and strictly non-negative; *i.e.,* if $X \in \mathcal{G}$ has zero variance, $||X||^2 = 0$, then X must be the deterministic zero variable, $X = 0$.

If $\mathcal{G}$ is generated by $\mathcal{X}$, then all inner products between elements of $\mathcal{G}$ are determined by the inner product (autocorrelation) matrix $R_{xx} = \{\langle X, X' \rangle \mid X, X' \in \mathcal{X}\}$ (the Gram matrix of $\mathcal{X}$). In other words, the matrix R_{xx} completely determines the geometry of $\mathcal{G}$. Since all subsets of variables in $\mathcal{G}$ are jointly Gaussian, the joint statistics of any such subset of $\mathcal{G}$ are completely determined by their second-order statistics, and thus by R_{xx}.

A subset $\mathcal{Y} \subset \mathcal{G}$ is called *linearly dependent* if there is some linear combination of the elements of $\mathcal{Y}$ that is equal to the zero variable 0, and linearly independent otherwise. We will see that a subset $\mathcal{Y} \subset \mathcal{G}$ is linearly independent if and only if its autocorrelation matrix R_{yy} has full rank.

Two random variables are *orthogonal* if their inner product is zero; *i.e.,* if they are uncorrelated. If two jointly Gaussian variables are orthogonal, then they are statistically independent. The only variable in $\mathcal{G}$ that is orthogonal to itself (*i.e.,* satisfies $\langle X, X \rangle = 0$) is the zero variable 0. If $\langle X, Y \rangle = 0$, then the *Pythagorean theorem* holds:

$$||X + Y||^2 = ||X||^2 + ||Y||^2.$$

Given any subset $\mathcal{Y} \subset \mathcal{G}$, the *closure* $\overline{\mathcal{Y}}$ of $\mathcal{Y}$, or the *subspace generated by* $\mathcal{Y}$, is the set of all linear combinations of elements of $\mathcal{Y}$. Also, the set of all $X \in \mathcal{G}$ that are orthogonal to all elements of $\mathcal{Y}$ is a subspace of $\mathcal{G}$, called the *orthogonal subspace* $\mathcal{Y}^\perp \subseteq \mathcal{G}$. Since 0 is the only element of $\mathcal{G}$ that is orthogonal to itself, the only common element of $\overline{\mathcal{Y}}$ and $\mathcal{Y}^\perp$ is 0.

1.2 The projection theorem

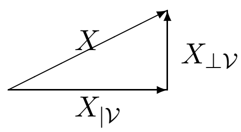
The key geometric property of the Hilbert space $\mathcal{G}$ is the **projection theorem**: if $\mathcal{V}$ and $\mathcal{V}^\perp$ are orthogonal subspaces of $\mathcal{G}$, then there exists a *unique* $X_{|\mathcal{V}} \in \mathcal{V}$ and $X_{\perp\mathcal{V}} \in \mathcal{V}^\perp$ such that $X = X_{|\mathcal{V}} + X_{\perp\mathcal{V}}$. $X_{|\mathcal{V}}$ and $X_{\perp\mathcal{V}}$ are called the *projections* of X onto $\mathcal{V}$ and $\mathcal{V}^\perp$, respectively.

A explicit formula for a projection $X_{|\mathcal{V}}$ such that $X - X_{|\mathcal{V}} \in \mathcal{V}^\perp$ will be given below. Uniqueness is the most important part of the projection theorem, and may be proved as follows: if $X = Y + Z$ and also $X = Y' + Z'$, where $Y, Y' \in \mathcal{V}$ and $Z, Z' \in \mathcal{V}^\perp$, then

$$0 = ||X - X||^2 = ||Y - Y'||^2 + ||Z - Z'||^2,$$

where the Pythagorean theorem applies since $Y - Y' \in \mathcal{V}$ and $Z - Z' \in \mathcal{V}^\perp$. Since norms are non-negative, this implies $||Y - Y'||^2 = ||Z - Z'||^2 = 0$, which implies $Y = Y'$ and $Z = Z'$.

The projection theorem is illustrated by the little "Pythagorean" diagram below. Since $X_{|\mathcal{V}}$ and $X_{\perp\mathcal{V}}$ are orthogonal, we have $||X||^2 = ||X_{|\mathcal{V}}||^2 + ||X_{\perp\mathcal{V}}||^2$.



If $\overline{\mathcal{Y}}$ is a subspace that is generated by a set of variables $\mathcal{Y}$, then with mild abuse of notation we will write $X_{|\mathcal{Y}}$ and $X_{\perp\mathcal{Y}}$ rather than $X_{|\overline{\mathcal{Y}}}$ and $X_{\perp\overline{\mathcal{Y}}}$.

1.3 Innovations representations

Let $\mathcal{X} \subset \mathcal{G}$ be a finite subset of elements of $\mathcal{G}$, and let $\overline{\mathcal{X}} \subseteq \mathcal{G}$ be the subspace of $\mathcal{G}$ generated by $\mathcal{X}$. An orthogonal basis for $\overline{\mathcal{X}}$ may then be found by a recursive (Gram-Schmidt) decomposition, as follows.

Denote the elements of the generator set $\mathcal{X}$ by $X_1, X_2, \ldots$, and let $\overline{\mathcal{X}_1^{i-1}}$ denote the subspace of $\mathcal{G}$ generated by $\mathcal{X}_1^{i-1} = \{X_1, X_2, \ldots, X_{i-1}\}$. To initialize, set $i = 1$ and $\mathcal{X}_1^0 = \emptyset$. For the ith recursion, using the projection theorem, write X_i uniquely as

$$X_i = (X_i)_{|\mathcal{X}_1^{i-1}} + (X_i)_{\perp\mathcal{X}_1^{i-1}}.$$

We have $(X_i)_{\perp\mathcal{X}_1^{i-1}} = 0$ if and only if $X_i \in \overline{\mathcal{X}_1^{i-1}}$. In this case $\overline{\mathcal{X}_1^{i}} = \overline{\mathcal{X}_1^{i-1}}$, so we can delete X_i from the generator set $\mathcal{X}$ without affecting $\overline{\mathcal{X}}$. Otherwise, we can take the "innovation" variable $E_i = (X_i)_{\perp\mathcal{X}_1^{i-1}} \neq 0$ as a replacement for X_i in the generator set; the space generated by $\mathcal{X}_1^{i-1}$ and E_i is still $\overline{\mathcal{X}_1^{i}}$, but E_i is orthogonal to $\overline{\mathcal{X}_1^{i-1}}$. By induction, the nonzero innovations variables up to E_i generate $\overline{\mathcal{X}_1^{i}}$ and are mutually orthogonal; *i.e.*, they form an orthogonal basis for $\overline{\mathcal{X}_1^{i}}$.

This recursive decomposition thus shows that:

- Any generator set $\mathcal{X}$ for a subspace $\overline{\mathcal{X}}$ contains a linearly independent generator set $\mathcal{X}' \subseteq \mathcal{X}$ that generates $\overline{\mathcal{X}}$. Therefore, without loss of generality, we may assume that any generator set $\mathcal{X}$ for $\overline{\mathcal{X}}$ is linearly independent.

- Given a linearly independent generator set $\mathcal{X} = \{X_1, X_2, \ldots\}$ for $\overline{\mathcal{X}}$, we can find an orthogonal basis $\mathcal{E} = \{E_1, E_2, \ldots\}$ for $\overline{\mathcal{X}}$, where $E_i = (X_i)_{\perp \mathcal{X}_1^{i-1}} = X_i - (X_i)_{|\mathcal{X}_1^{i-1}}$. Since $(X_i)_{|\mathcal{X}_1^{i-1}}$ is a linear combination of $X_1, X_2, \ldots, X_{i-1}$, if we write $\mathcal{X}$ and $\mathcal{E}$ as column vectors, then we have
$$\mathcal{E} = L^{-1}\mathcal{X},$$
where L^{-1} is a monic (*i.e.,* having ones on the diagonal) lower triangular matrix. Since L^{-1} is square and has a monic lower triangular inverse L, we may write alternatively
$$\mathcal{X} = L\mathcal{E}.$$

We conclude that a finite set of random variables $\mathcal{X}$ is jointly Gaussian if and only if $\mathcal{X}$ can be written as a monic lower triangular ("causal") linear transformation $\mathcal{X} = L\mathcal{E}$ of an orthogonal innovations sequence $\mathcal{E}$. All innovations variables are nonzero (*i.e.,* $\mathcal{E}$ is linearly independent) if and only if $\mathcal{X}$ is linearly independent. This is called an *innovations representation* of $\mathcal{X}$.

Moreover, the expression $\mathcal{X} = L\mathcal{E}$ implies that the autocorrelation matrix of $\mathcal{X}$ is
$$R_{xx} = LR_{ee}L^* = LD^2L^*,$$
where L^* denotes the conjugate transpose of L (a monic upper triangular matrix), and R_{ee} is a non-negative real diagonal matrix D^2, because $\mathcal{E}$ is an orthogonal sequence. This is called a *Cholesky decomposition* of R_{xx}; the diagonal elements $(d_i)^2 = ||E_i||^2$ of D^2 are called the *Cholesky factors* of R_{xx}. The Cholesky factors are all nonzero, and thus R_{ee} and R_{xx} have full rank, if and only if $\mathcal{X}$ is linearly independent. In general, the rank of R_{xx} is the number of nonzero innovations variables E_i in the innovations representation $\mathcal{X} = L\mathcal{E}$.

1.4 Differential entropy

To find the differential entropy $h(\mathcal{X})$ of a linearly independent set $\mathcal{X}$ of N jointly Gaussian random variables, we first recall that the differential entropy of a complex Gaussian variable X with variance $||X||^2 > 0$ is $h(X) = \log \pi e ||X||^2$. Then, by the innovations representation of a linearly independent generator set $\mathcal{X}$, we write
$$\mathcal{X} = \{X_1, f_2(X_1) + (X_2)_{\perp X_1}, f_3(\mathcal{X}_1^2) + (X_3)_{\perp \mathcal{X}_1^2}, \ldots\}$$
where $f_i(\mathcal{X}_1^{i-1}) = (X_i)_{|\mathcal{X}_1^{i-1}}$ is a linear function of $\mathcal{X}_1^i$, and $(X_i)_{\perp \mathcal{X}_1^{i-1}}$ is a sequence of orthogonal Gaussian variables with variances $(d_i)^2 = ||(X_i)_{\perp \mathcal{X}_1^{i-1}}||^2 > 0$. Now if two variables X, Y are independent and $f(X)$ is any function of X, then, by the chain rule of differential entropy and the invariance of differential entropy under translation, we have
$$h(X, Y + f(X)) = h(X) + h(Y + f(X) \mid X) = h(X) + h(Y).$$
Therefore
$$h(\mathcal{X}) = \log \pi e (d_1)^2 + \log \pi e (d_2)^2 + \cdots = \log(\pi e)^N |R_{xx}|,$$
where $|R_{xx}| = \det R_{xx}$. (This can be recognized as a version of the chain rule $h(\mathcal{X}) = h(X_1) + h(X_2 \mid X_1) + \cdots$ for Gaussian variables.)

Thus the differential entropy per complex dimension is
$$\frac{h(\mathcal{X})}{N} = \log \pi e |R_{xx}|^{1/N},$$
where $|R_{xx}|^{1/N}$ is the geometric mean of the Cholesky factors (or eigenvalues) of R_{xx}. Note that this result is independent of the order in which we take the variables in $\mathcal{X}$.

1.5 Fundamentals of MMSE estimation theory

Suppose that X represents a random variable to be estimated and that $\mathcal{Y}$ represents a set of observed variables, where X and $\mathcal{Y}$ are jointly Gaussian. A *linear estimate* of X is a linear function of $\mathcal{Y}$; *i.e.,* a random variable V in the space $\overline{\mathcal{Y}}$. The *estimation error* is then $E = X - V$.

By the projection theorem, the projection $X_{|\mathcal{Y}} \in \overline{\mathcal{Y}}$ minimizes the estimation error variance $||X - V||^2$ over $V \in \overline{\mathcal{Y}}$, because, using the Pythagorean theorem and the fact that $X_{|\mathcal{Y}} - V \in \overline{\mathcal{Y}}$ while $X_{\perp\mathcal{Y}} \in \overline{\mathcal{Y}}^{\perp}$, we have

$$||X - V||^2 = ||X_{|\mathcal{Y}} + X_{\perp\mathcal{Y}} - V||^2 = ||X_{|\mathcal{Y}} - V||^2 + ||X_{\perp\mathcal{Y}}||^2 \geq ||X_{\perp\mathcal{Y}}||^2,$$

with equality if and only if $V = X_{|\mathcal{Y}}$. For this reason $X_{|\mathcal{Y}}$ is called the *minimum-mean-squared error* (MMSE) *linear estimate* of X given $\mathcal{Y}$, and $X_{\perp\mathcal{Y}}$ is called the *MMSE estimation error.* Moreover, the *orthogonality principle* holds: $V \in \overline{\mathcal{Y}}$ is the MMSE linear estimate of X given $\mathcal{Y}$ if and only if $X - V$ is orthogonal to $\overline{\mathcal{Y}}$.

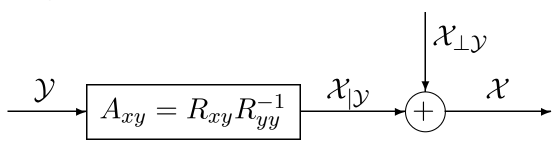
Similarly, if $\mathcal{X} \subseteq \mathcal{G}$ is a set of random variables, then by the orthogonality principle the set $\mathcal{V} \in \overline{\mathcal{Y}}$ is the corresponding set $\mathcal{X}_{|\mathcal{Y}}$ of MMSE linear estimates of $\mathcal{X}$ given $\mathcal{Y}$ if and only if $\langle \mathcal{X} - \mathcal{V}, \mathcal{Y} \rangle = 0$, or $\langle \mathcal{V}, \mathcal{Y} \rangle = \langle \mathcal{X}, \mathcal{Y} \rangle$. Writing $\mathcal{V}$ as a set of linear combinations of the elements of $\mathcal{Y}$ in matrix form, namely $\mathcal{V} = A_{xy}\mathcal{Y}$, and defining R_{xy} as the cross-correlation matrix $\langle \mathcal{X}, \mathcal{Y} \rangle$ and R_{yy} as the autocorrelation matrix $\langle \mathcal{Y}, \mathcal{Y} \rangle$, we obtain a unique solution

$$A_{xy} = R_{xy} R_{yy}^{-1},$$

where without loss of generality we assume that R_{yy} is invertible; *i.e.,* that $\mathcal{Y}$ is a linearly independent generator set for $\overline{\mathcal{Y}}$. In short, an explicit formula for the projection of $\mathcal{X}$ onto $\overline{\mathcal{Y}}$ is

$$\mathcal{X}_{|\mathcal{Y}} = R_{xy} R_{yy}^{-1} \mathcal{Y}.$$

The expression $\mathcal{X} = A_{xy}\mathcal{Y} + \mathcal{X}_{\perp\mathcal{Y}}$ shows that $\mathcal{X}$ may be regarded as the sum of a linear estimate derived from $\mathcal{Y}$ and an independent error (innovations) variable $\mathcal{E} = \mathcal{X}_{\perp\mathcal{Y}}$. This decomposition is illustrated in the block diagram below.



Since $\mathcal{X}_{\perp\mathcal{Y}}$ has zero mean and is independent of $\mathcal{Y}$, we have $\mathsf{E}[\mathcal{X} \mid \mathcal{Y}] = \mathcal{X}_{|\mathcal{Y}}$; *i.e.,* the MMSE linear estimate $\mathcal{X}_{|\mathcal{Y}}$ is the *conditional mean* of $\mathcal{X}$ given $\mathcal{Y}$. Indeed, this decomposition shows that the conditional distribution of $\mathcal{X}$ given $\mathcal{Y}$ is Gaussian with mean $\mathcal{X}_{|\mathcal{Y}}$ and autocorrelation matrix $R_{ee} = R_{xx} - R_{xy} R_{yy}^{-1} R_{yx}$, by Pythagoras. Thus $\mathcal{X}_{|\mathcal{Y}}$ is evidently the *unconstrained* MMSE estimate of $\mathcal{X}$ given $\mathcal{Y}$; *i.e.,* our earlier restriction to a linear estimate is no real restriction.

Moreover, this block diagram implies that the MMSE estimate $\mathcal{X}_{|\mathcal{Y}}$ is a *sufficient statistic* for estimation of $\mathcal{X}$ from $\mathcal{Y}$, since $\mathcal{Y} - \mathcal{X}_{|\mathcal{Y}} - \mathcal{X}$ is evidently a Markov chain; *i.e.,* $\mathcal{Y}$ and $\mathcal{X}$ are conditionally independent given $\mathcal{X}_{|\mathcal{Y}}$. We call this the **sufficiency property** of the MMSE estimate. This implies that $\mathcal{X}$ can be estimated as well from the projection $\mathcal{X}_{|\mathcal{Y}}$ as from $\mathcal{Y}$, so there is no loss of estimation optimality if we first reduce $\mathcal{Y}$ to $\mathcal{X}_{|\mathcal{Y}}$.

Actually, $\mathcal{X}_{|\mathcal{Y}}$ is a *minimal* sufficient statistic; *i.e.,* $\mathcal{X}_{|\mathcal{Y}}$ is a function of every other sufficient statistic $f(\mathcal{Y})$. This follows from the fact that the conditional distribution of $\mathcal{X}$ given $f(\mathcal{Y})$ must be the same as the conditional distribution given $\mathcal{Y}$, which implies that the conditional mean $\mathcal{X}_{|\mathcal{Y}}$ can be determined from $f(\mathcal{Y})$.

1.6 A bit of information theory

By the sufficiency property, the MMSE estimate $\mathcal{X}_{|\mathcal{Y}}$ is a function of $\mathcal{Y}$ that satisfies the data processing inequality of information theory with equality: $I(\mathcal{X};\mathcal{Y}) = I(\mathcal{X};\mathcal{X}_{|\mathcal{Y}})$. In other words, the reduction of $\mathcal{Y}$ to $\mathcal{X}_{|\mathcal{Y}}$ is *information-lossless*.

Moreover, since $\mathcal{X} = A_{xy}\mathcal{Y} + \mathcal{X}_{\perp\mathcal{Y}}$ is a linear Gaussian channel model with Gaussian input $\mathcal{Y}$, Gaussian output $\mathcal{X}$, and independent additive Gaussian noise $\mathcal{E} = \mathcal{X}_{\perp\mathcal{Y}}$, we have

$$I(\mathcal{X};\mathcal{Y}) = h(\mathcal{X}) - h(\mathcal{X} \mid \mathcal{Y}) = h(\mathcal{X}) - h(\mathcal{E}) = \log \frac{|R_{xx}|}{|R_{ee}|},$$

where we recall that the differential entropy of a set $\mathcal{X}$ of N complex Gaussian random variables with nonsingular autocorrelation matrix R_{xx} is $h(\mathcal{X}) = \log(\pi e)^N |R_{xx}|$. (We assume that R_{ee} is nonsingular, else $\{\mathcal{X}, \mathcal{Y}\}$ is linearly dependent, so at least one dimension of $\mathcal{X}$ may be determined precisely from $\mathcal{Y}$ and $I(\mathcal{X};\mathcal{Y}) = \infty$.)

1.7 Chain rule of MMSE estimation

Suppose that $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are jointly Gaussian sets of random variables and that we wish to estimate $\mathcal{X}$ based on $\mathcal{Y}$ and $\mathcal{Z}$. The MMSE estimate is then $\mathcal{X}_{|\mathcal{Y}\mathcal{Z}}$, the projection of $\mathcal{X}$ onto the subspace $\overline{\mathcal{Y}} + \overline{\mathcal{Z}}$ generated by the variables in both $\mathcal{Y}$ and $\mathcal{Z}$.

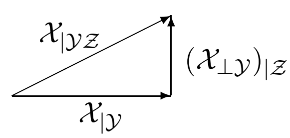
The subspace $\overline{\mathcal{Y}} + \overline{\mathcal{Z}}$ may be written as the sum of two orthogonal subspaces as follows:

$$\overline{\mathcal{Y}} + \overline{\mathcal{Z}} = \overline{\mathcal{Y}} + \left(\overline{\mathcal{Y}}^{\perp} \cap \overline{\mathcal{Z}}\right).$$

Correspondingly, we may write the projection $\mathcal{X}_{|\mathcal{Y}\mathcal{Z}}$ as the sum of two orthogonal projections as follows:

$$\mathcal{X}_{|\mathcal{Y}\mathcal{Z}} = \mathcal{X}_{|\mathcal{Y}} + (\mathcal{X}_{\perp\mathcal{Y}})_{|\mathcal{Z}}.$$

We call this the *chain rule of MMSE estimation*. It is illustrated below:



Generalizing, if we wish to estimate $\mathcal{X}$ based on a sequence $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots\}$ of random variables such that $\mathcal{X}$ and $\mathcal{Y}$ are jointly Gaussian, then the chain rule of MMSE estimation becomes

$$\mathcal{X}_{|\mathcal{Y}} = \mathcal{X}_{|\mathcal{Y}_1} + (\mathcal{X}_{\perp\mathcal{Y}_1})_{|\mathcal{Y}_2} + \cdots + (\mathcal{X}_{\perp\mathcal{Y}_1^{i-1}})_{|\mathcal{Y}_i} + \cdots,$$

where $\mathcal{Y}_1^{i-1} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_{i-1}\}$. The incremental estimate $(\mathcal{X}_{\perp\mathcal{Y}_1^{i-1}})_{|\mathcal{Y}_i}$ thus represents the "new information" given by the observation $\mathcal{Y}_i$ about $\mathcal{X}$, given the previous observations $\mathcal{Y}_1^{i-1}$.

The innovations representation may be seen as a special case of the chain rule of MMSE estimation. Indeed, if $\mathcal{X} = \{X_1, X_2, \ldots\}$ and we take $\mathcal{Y} = \mathcal{X}$, then $\mathcal{X}_{|\mathcal{X}} = \mathcal{X}$, and the "new information" sequence becomes

$$(\mathcal{X}_{\perp\mathcal{X}_1^{i-1}})_{|X_i} = \{0, \ldots, 0, (X_i)_{\perp\mathcal{X}_1^{i-1}}, \ldots\};$$

i.e., the first i components of $(\mathcal{X}_{\perp\mathcal{X}_1^{i-1}})_{|X_i}$ are $\{0, \ldots, 0, E_i\}$, where $E_i = (X_i)_{\perp\mathcal{X}_1^{i-1}}$ is the ith innovation variable of $\mathcal{X}$; the remaining components are evidently linearly dependent on E_i.

2 Successive decoding

Often it is natural or helpful to regard a set $\mathcal{X}$ of Gaussian random variables as a sequence of subsets, $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots\}$. For instance $\mathcal{X}_1, \mathcal{X}_2, \ldots$ might represent a discrete-time sequence, in which case the ordering naturally follows the time ordering; or, in a multi-user scenario, $\mathcal{X}_1, \mathcal{X}_2, \ldots$ might represent different users, in which case the ordering may be arbitrary. Thus the index set $\{1, 2, \ldots\}$ indicates an ordering, but is not necessarily a time index set.

Our aim will be to signal at a rate approaching the mutual information $I(\mathcal{X}; \mathcal{Y})$. As above, we may write

$$I(\mathcal{X}; \mathcal{Y}) = I(\mathcal{X}; \mathcal{X}_{|\mathcal{Y}}) = h(\mathcal{X}) - h(\mathcal{E}) = \log \frac{|R_{xx}|}{|R_{ee}|},$$

where $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ is the sequence of estimation error subsets $\mathcal{E}_i = (\mathcal{X}_i)_{\perp\mathcal{Y}}$.

We will consider a successive decoding scenario in which the subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots$ are detected sequentially from a set $\mathcal{Y}$ of observed variables. For each index i, we will aim to signal at a rate approaching the incremental rate

$$R_i = h(\mathcal{X}_i \mid \mathcal{X}_i^{i-1}) - h(\mathcal{E}_i \mid \mathcal{E}_i^{i-1}),$$

where $\mathcal{X}_i^{i-1} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{i-1}\}$ and $\mathcal{E}_i^{i-1} = (\mathcal{X}_i^{i-1})_{\perp\mathcal{Y}}$. By the chain rule of differential entropy, we will then approach a total rate of $\sum_i R_i = h(\mathcal{X}) - h(\mathcal{E}) = I(\mathcal{X}; \mathcal{Y})$.

For successive decoding, we will make the following critical assumption:

> **Ideal decision feedback assumption**: In the detection of the variable subset $\mathcal{X}_i$, the values of the previous variables $\mathcal{X}_i^{i-1}$ are known precisely.

The ideal decision feedback assumption is the decisive break between the classical analog estimation theory of Wiener *et al.* and the digital Shannon theory. If the $\mathcal{X}_i$ are continuous Gaussian variables, then in general it is nonsense to suppose that they can be estimated precisely (assuming that $\mathcal{X}$ and $\mathcal{Y}$ are not linearly dependent). On the other hand, if the $\mathcal{X}_i$ are codewords in some discrete code $\mathcal{C}$ whose words are chosen randomly according to the Gaussian statistics of $\mathcal{X}_i$ given $\mathcal{X}_i^{i-1}$, and if the length of $\mathcal{C}$ is large enough and the rate of $\mathcal{C}$ is less than the incremental rate R_i, then Shannon theory shows that the probability of not decoding $\mathcal{X}_i$ precisely given $\mathcal{Y}$ and $\mathcal{X}_1^{i-1}$ may be driven arbitrarily close to 0. So in a digital coding scenario, the ideal decision feedback assumption may be quite reasonable.

The MMSE estimate $(\mathcal{X}_i)_{|\mathcal{Y},\mathcal{X}_1^{i-1}}$ of $\mathcal{X}_i$ is a sufficient statistic for estimation of $\mathcal{X}_i$ given $\mathcal{Y}$ and $\mathcal{X}_1^{i-1}$. Moreover, since $\mathcal{X}_i = (\mathcal{X}_i)_{|\mathcal{Y}} + \mathcal{E}_i$, the subspace generated by $\mathcal{Y}$ and $\mathcal{E}_i$ is equal to the subspace generated by $\mathcal{Y}$ and $\mathcal{X}_i$, and similarly the subspace generated by $\mathcal{Y}$ and $\mathcal{E}_1^{i-1}$ is the subspace generated by $\mathcal{Y}$ and $\mathcal{X}_1^{i-1}$. Therefore we may alternatively write

$$(\mathcal{X}_i)_{|\mathcal{Y},\mathcal{X}_1^{i-1}} = (\mathcal{X}_i)_{|\mathcal{Y},\mathcal{E}_1^{i-1}} = (\mathcal{X}_i)_{|\mathcal{Y}} + ((\mathcal{X}_i)_{\perp\mathcal{Y}})_{|\mathcal{E}_1^{i-1}} = (\mathcal{X}_i)_{|\mathcal{Y}} + (\mathcal{E}_i)_{|\mathcal{E}_1^{i-1}},$$

where we have used the chain rule of MMSE estimation. The estimation error is $(\mathcal{E}_i)_{\perp\mathcal{E}_1^{i-1}}$. In short, $\mathcal{X}_i$ is the sum of three independent components: the MMSE estimate of $\mathcal{X}_i$ given $\mathcal{Y}$, the MMSE prediction of $\mathcal{E}_i$ given $\mathcal{E}_1^{i-1}$, and the estimation error $(\mathcal{E}_i)_{\perp\mathcal{E}_1^{i-1}}$. The differential entropy of the estimation error may thus be written in any of the following ways:

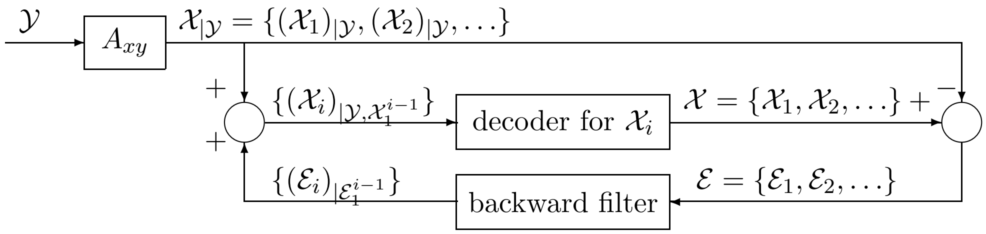
$$h((\mathcal{E}_i)_{\perp\mathcal{E}_i^{i-1}}) = h(\mathcal{X}_i \mid \mathcal{Y}, \mathcal{X}_i^{i-1}) = h(\mathcal{E}_i \mid \mathcal{E}_i^{i-1}).$$

We note therefore that $\sum_i R_i = I(\mathcal{X};\mathcal{Y})$ follows alternatively from the chain rule of mutual information, since

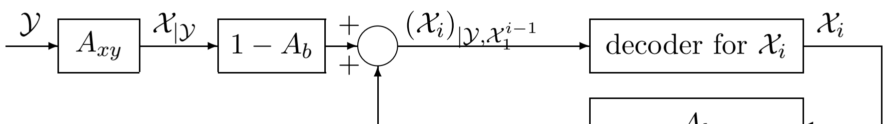
$$I(\mathcal{X}_i;\mathcal{Y} \mid \mathcal{X}_i^{i-1}) = h(\mathcal{X}_i \mid \mathcal{X}_i^{i-1}) - h(\mathcal{X}_i \mid \mathcal{Y}, \mathcal{X}_i^{i-1}) = h(\mathcal{X}_i \mid \mathcal{X}_i^{i-1}) - h(\mathcal{E}_i \mid \mathcal{E}_i^{i-1}) = R_i.$$

Successive decoding then works as follows. The sequence to be decoded is $\mathcal{X}_1, \mathcal{X}_2, \ldots$, and the observed sequence is $\mathcal{Y}$. We first reduce $\mathcal{Y}$ to the MMSE estimate $(\mathcal{X}_1)_{|\mathcal{Y}}$ and decode $\mathcal{X}_1$ from it, in the presence of the error $\mathcal{E}_1 = (\mathcal{X}_1)_{\perp\mathcal{Y}}$. If the decoding of $\mathcal{X}_1$ is correct, then we can compute $\mathcal{E}_1 = \mathcal{X}_1 - (\mathcal{X}_1)_{|\mathcal{Y}}$ and form the estimate $(\mathcal{E}_2)_{|\mathcal{E}_1}$, which we add to $(\mathcal{X}_2)_{|\mathcal{Y}}$ to form the input to a decoder for $\mathcal{X}_2$ with error $(\mathcal{E}_2)_{\perp\mathcal{E}_1}$, and so forth.

This "decision feedback" scheme is illustrated in the figure below. The "forward filter" A_{xy} is the MMSE estimator of the sequence $\mathcal{X}$ given $\mathcal{Y}$. The "backward filter" is the MMSE predictor of $\mathcal{E}_i$ given $\mathcal{E}_1^{i-1}$, where ideal decision feedback is assumed in computing the previous error $\mathcal{E}_1^{i-1}$.



This decision-feedback scheme is said to be in "noise-predictive" form, since the error sequence $\mathcal{E}$ is predicted by the causal backward filter. By linearity, we can put it into more standard decision-feedback form as shown below, where the backward filter is denoted by A_b:



Successive decoding thus breaks the joint detection of $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots\}$ into a series of "per-user" steps. This idea underlies classical decision-feedback schemes for sequential transmission on a single channel, and also successive interference cancellation schemes on multi-access channels.

Moreover, if we can achieve a small error probability with a code of rate close to R_i for each i, then we can achieve an aggregate rate close to $I(\mathcal{X};\mathcal{Y})$ with an error probability no greater than the sum of the component error probabilities, by the union bound. Again, this holds regardless of the ordering of the users.

In practice, achieving a rate approaching the mutual information will require very long codes. This is usually not an obstacle in a multi-access scenario. In the case of sequential transmission on a single channel which is not memoryless, it can be achieved in principle by interleaving beyond the memory length of the channel (for details, see [7]). Alternatively, if the channel is known at the transmitter, then interference may be effectively removed at the transmitter by various precoding or precancellation schemes (*e.g.,* [1, 4, 9]).

These schemes naturally extend to infinite jointly stationary and jointly Gaussian sequences $\mathcal{X} = \{\ldots, \mathcal{X}_0, \mathcal{X}_1, \ldots\}$ and $\mathcal{Y} = \{\ldots, \mathcal{Y}_0, \mathcal{Y}_1, \ldots\}$. The forward and backward filters shown above become time-invariant in the limit. Cholesky decompositions become multivariate spectral factorizations. Sequence mutual information quantities such as $I(\mathcal{X};\mathcal{Y})$ are replaced by information rates. For a full development, see Guess and Varanasi [8]. The point is that the conceptual basis of the development is essentially the same.

References

[1] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu and G. D. Forney, Jr., "MMSE decision-feedback equalizers and coding— Part I: Equalization results; Part II: Coding results," *IEEE Trans. Commun.*, vol. 43, pp. 2581–2604, Oct. 1995.

[2] J. M. Cioffi and G. D. Forney, Jr., "Generalized decision-feedback equalization for packet transmission with ISI and Gaussian noise," in *Communications, Computation, Control and Signal Processing* (A. Paulraj *et al.*, eds.), pp. 79–127.

[3] U. Erez and R. Zamir, "Lattice decoding can achieve $\frac{1}{2}\log(1 + \mathrm{SNR})$ on the AWGN channel," *Proc. Intl. Symp. Inform. Theory* (Washington, DC), p. 300, June 2001.

[4] M. V. Eyuboglu and G. D. Forney, Jr., "Trellis precoding: Combined coding, shaping and precoding for intersymbol interference channels," *IEEE Trans. Inform. Theory*, vol. 38, pp. 301–314, Mar. 1992.

[5] G. D. Forney, Jr., "On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener," *Proc. 2003 Allerton Conf.* (Monticello, IL), Oct. 2003.

[6] T. Guess and M. K. Varanasi, "An information-theoretic derivation of the MMSE decision-feedback equalizer," *Proc. 1998 Allerton Conf.* (Monticello, IL), Sept. 1998.

[7] T. Guess and M. K. Varanasi, "A new successively decodable coding technique for intersymbol interference channels," *Proc. Intl. Symp. Inform. Theory* (Sorrento, Italy), p. 102, June 2000.

[8] T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," submitted to *IEEE Trans. Inform. Theory*, Sept. 2002; revised Mar. 2004.

[9] S. Shamai (Shitz) and R. Laroia, "The intersymbol interference channel: Lower bounds on capacity and channel precoding loss," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1388–1404, Sept. 1996.