

Prediction and quantization

Prediction, especially linear prediction, can be incorporated into quantization in many ways.

- Prediction primer
 - One-step prediction
 - Optimal prediction
 - Optimal linear prediction
 - Optimal prediction with Gaussian vectors
 - General linear prediction
- Linear prediction and quantization
 - Predictive quantization

- Linear predictive coding (LPC)
- Code excited linear prediction (CELP)

Prediction Primer

A typical prediction problem is the following classic one-step prediction problem:

Fix m . Observe a data sequence $X^m = \{X_0, X_1, \dots, X_{m-1}\}$.

What is the *optimal* predictor of X_m of the form $\tilde{X}_m = p(X_0, \dots, X_{m-1})$?

Common definition of “optimal”: predictor is optimal if it minimizes the mean squared error (MSE) $E(\epsilon_m^2)$ where $\epsilon = X_m - \hat{X}_m$ is the prediction error.

Optimal (MMSE) prediction

Classic problem with classic answer, standard result (in most basic probability and statistics courses, see, e.g., Section 4.9, *Introduction to Statistical Signal Processing*, Gray and Davisson, Cambridge, (2004)): Optimal predictor is conditional expectation

$$p(X_0, \dots, X_{m-1}) = E(X_m | X_0, \dots, X_{m-1})$$

and MMSE is conditional variance

$$E(\epsilon_m^2) = \sigma_{X_m | X_0, \dots, X_{m-1}}^2$$

Important property of conditional expectation.

Problem: Generally difficult to compute, unless X^{m+1} Gaussian

Linear prediction

Difficulty of optimal (possibly nonlinear) predictor suggests constraining structure of prediction function to a simple form: linear prediction

What is the optimal linear predictor of the form

$$\tilde{X}_m = -\sum_{l=1}^m a_l X_{m-l} = -\sum_{l=0}^{m-1} a_{m-l} X_l ?$$

The sign is chosen for later convenience. In particular, *prediction error* is

$$\epsilon_m = X_m - \tilde{X}_m = X_m + \sum_{l=1}^m a_l X_{m-l} = \sum_{l=0}^m a_l X_{m-l}$$

if define $a_0 = 1$

Matrix version:

$$\epsilon_m = \sum_{l=0}^m a_l X_{m-l} = a^t X^{(m)}$$

where $X^{(m)} = (X_{m-1}, X_{m-2}, \dots, X_0)^t$, a time-reversed version of $X^m = (X_0, X_1, \dots, X_{m-1})^t \Rightarrow$

$$\begin{aligned} D_m(a) &= E(\epsilon_m^2) = E(a^t X^{(m)} X^{(m)t} a) \\ &= a^t E(X^{(m)} X^{(m)t}) a \\ &= a^t R^{(m+1)} a \end{aligned}$$

where $(m+1) \times (m+1)$ -dimensional matrix $R^{(m+1)} = \{r(m-i, m-j); i, j = 0, 1, \dots, m\}$

Optimal one-step linear predictor: Find $a_0 = 1, a_1, a_2, \dots, a_m$ that minimizes

$$\begin{aligned} D_m(a) &\triangleq E(\epsilon_m^2) \\ &= E\left[\left(\sum_{l=0}^m a_l X_{m-l}\right)^2\right] = E\left[\left(\sum_{l=0}^m a_{m-l} X_l\right)^2\right] \\ &= \sum_{l=0}^m \sum_{j=0}^m a_l a_j E(X_{m-l} X_{m-j}) = \sum_{l=0}^m \sum_{j=0}^m a_{m-l} a_{m-j} E(X_l X_j) \\ &\triangleq \sum_{l=0}^m \sum_{j=0}^m a_l a_j r(m-i, m-j) = \sum_{l=0}^m \sum_{j=0}^m a_{m-l} a_{m-j} r(i, j) \end{aligned}$$

quadratic forms

If $\{X_n\}$ drawn from stationary source, $R^{(m+1)} = R_{m+1}$ and $D_m(a) = a^t R_{m+1} a$.

Classic linear prediction (LP) problem:

Minimize quadratic form $D_m(a) = a^t R^{(m+1)} a$ over all $a : a_0 = 1$

$$D_m \triangleq \inf_{a: a_0=1} a^t R^{(m+1)} a$$

Optimal Prediction: Gaussian case

Assume random vector $X^{m+1} = (X_0, X_1, \dots, X_m)^t$ Gaussian

Then linear algebra \Rightarrow (e.g., Section 4.9, *Introduction to Statistical Signal Processing*)

$$\begin{aligned} E[X_m|X^m] &= (r(m, 0), r(m, 1), \dots, r(m, m-1))R_m^{-1}X^m \quad (85) \\ &\triangleq -(a_m, \dots, a_2, a_1)^t X^m \\ &= -\sum_{l=1}^m a_l X_{m-l} \end{aligned}$$

$$\text{MMSE} = |R_{m+1}| / |R_m| \triangleq \alpha_m \quad (86)$$

Optimal linear prediction

Implication of Gaussian case: Since $D_m(a)$ depends on pdf of X^{m+1} *only* through correlation $R^{(m+1)}$,

$$\underset{a: a_0=1}{\operatorname{argmin}} D_m(a) = \text{solution to (87)} \quad (88)$$

$$\min_{a: a_0=1} D_m(a) = \alpha_m \quad (89)$$

regardless of whether Gaussian or not!

i.e., $D_m(a) = a^t R^{(m+1)} a$ = mean squared linear prediction error for *any* pdf given R_{m+1} and a .

If X^{m+1} Gaussian, then can achieve α_m using a of (87), but even if not Gaussian, $D_m(a)$ depends only on R_{m+1} and a , so can also achieve $D_m(a) = \alpha_m$ in non Gaussian case

i.e., **optimal predictor is linear** with a determined by (85):

$$(a_m, \dots, a_2, a_1) = -(r(m, 0), r(m, 1), \dots, r(m, m-1))R_m^{-1}, \quad a_0 = 1 \quad (87)$$

optimal predictor for Gaussian = optimal linear predictor for Gaussian

Predictor and performance are determined entirely by correlation matrix R_{m+1} !

Cannot do better in non Gaussian case, i.e., cannot get $D_m(a) < \alpha_m$, since then in the Gaussian case with same autocorrelation linear predictor would outperform optimal predictor. #

\Rightarrow Gaussian optimal solution \Rightarrow general linear optimal solution:
 $\Rightarrow D_m = \alpha_m$

Moral: Gaussian assumption provides short cut proofs in nonGaussian problems — no calculus and get global optimality!

Efficient inversion to find a : Cholesky decomposition \implies

· *covariance method* in speech processing

If R_{m+1} Toeplitz, Levinson-Durbin algorithm \implies

· *autocorrelation method* in speech processing

Orthogonality principle and the normal equations

Common approach to prove optimal linear predictor: Calculus or orthogonality principle \Rightarrow normal equations (Wiener-Hopf, Yule-Walker): m linear equations in m unknowns.

For completeness sketch the derivation for the case of samples drawn from a stationary process.

Orthogonality Principle

Well known that the optimal linear estimator must cause the prediction error to be orthogonal to the observations, the *orthogonality principle*, i.e., $E(\epsilon_n X_{n-k}) = 0$, $k = 1, 2, \dots, m$

(See, e.g., Section 4.11 in *Introduction to Statistical Signal Processing*)

Normal equations are equivalent to the solution of (87) in the stationary case, where (87) becomes

$$(a_m, \dots, a_1)R_m = -(r(m), r(m-1), \dots, r(1))$$

Proof: ((87)) \Leftrightarrow

$$\sum_{i=0}^{m-1} a_{m-i} r(i, j) = -r(m-j); j = 0, 1, \dots, m-1$$

Changing variables to $l = m - i$

$$\sum_{l=1}^m a_l r(m-l, j) = -r(m-j); j = 0, 1, \dots, m-1$$

can express as

$$\begin{aligned} 0 &= E\left[\left(\sum_{l=0}^m X_{n-l} a_l\right) X_{n-k}\right] \\ &= \sum_{l=0}^m a_l E(X_{n-l} X_{n-k}) \\ &= \sum_{l=0}^m a_l R(k-l); k = 1, 2, \dots, m \end{aligned}$$

This set of m linear equations in m unknowns a_l ; $l = 1, \dots, m$ is known as the *normal equations* or *finite-memory Wiener-Hopf equation in discrete time* or the *Yule-Walker equations*

and $k = m - j$

$$\sum_{i=1}^m a_i r(m-l, m-k) = -r(k); k = 1, \dots, m$$

Since the process is assumed stationary, this is exactly the normal equations

$$\sum_{l=1}^m a_l r(k-l) = -r(k); k = 1, \dots, m$$

Aside: For completeness, show that the solution to normal equations indeed provides a global minimum to the LP optimization:

Suppose that $a = \{a_k; k = 0, \dots, m\}$ solves the normal equations and that $a_0 = 1$ and let $b = \{b_k; k = 0, \dots, m\}$ be any other set of linear prediction coefficients with $b_0 = 1$. MSE with b_l is

$$\begin{aligned}
D_m(b) &= E \left(\left| \sum_{l=0}^m b_l X_{n-l} \right|^2 \right) \\
&= E \left(\left| \sum_{l=0}^m a_l X_{n-l} + (b_l - a_l) X_{n-l} \right|^2 \right) \\
&= E \left(\left| \sum_{l=0}^m a_l X_{n-l} \right|^2 \right) + E \left(\left| \sum_{l=0}^m (b_l - a_l) X_{n-l} \right|^2 \right) \\
&\quad + 2E \left(\left(\sum_{l=0}^m a_l X_{n-l} \right) \left(\sum_{k=0}^m (b_k - a_k) X_{n-k} \right) \right) \\
&\geq D_m(a) + 2 \sum_{k=0}^m (b_k - a_k) \sum_{l=0}^m a_l E(X_{n-l} X_{n-k}) \\
&= D_m(a) + \sum_{k=0}^m (b_k - a_k) \sum_{l=0}^m a_l R(k-l) = D_m(a)
\end{aligned}$$

Estimating correlations

What if don't know R_m , observe long sequence of actual data X_0, X_1, \dots, X_{n-1} ? Under suitable conditions can estimate:

$$\begin{aligned}
\hat{r}_k &= \frac{1}{n-m} \sum_{l=m}^{n-1} X_l X_{l-|k|}; \hat{R}_{m+1} = \{\hat{r}_{i,j}; i, j = 0, 1, \dots, m\} \\
\bar{r}_{i,j} &= \frac{1}{n-m} \sum_{l=m}^{n-1} X_{l-i} X_{l-j}; \bar{R}_{m+1} = \{\bar{r}_{i,j}; i, j = 0, 1, \dots, m\}
\end{aligned}$$

and "plug in."

\hat{R}_m Toeplitz, \bar{R}_m not.

Under suitable assumptions, as $n \rightarrow \infty$, $\bar{R}_{m+1} \approx \hat{R}_{m+1} \approx R_{m+1}$

$$LP \Leftrightarrow \operatorname{argmin}_{a: a_0=1} a^t R_{m+1} a$$

where used the facts that $b_0 = a_0 = 1$ and

$$\sum_{l=0}^m a_l R(k-l) = 0; k = 1, 2, \dots, m$$

Processes and Filters

Extend vector prediction to one-step prediction of a random process.

The past m values of a stationary zero mean random process $X_{n-m}, \dots, X_{n-2}, X_{n-1}, X_{n-1}$ are observed and based on these observations, an estimate $\tilde{X}_n = p(\{X_l; l \leq n-1\})$ of X_n is formed

E.g., $\tilde{X}_n = X_{n-1}$ or

$$\tilde{X}_n = - \sum_{l=1}^m a_l X_{n-l}$$

a general finite-order linear predictor, or

$$\tilde{X}_n = - \sum_{l=1}^{\infty} a_l X_{n-l},$$

an infinite-order predictor.

When allowing the filter to have infinite order, A is required to be stable and have minimum phase (or minimum delay, i.e., all the roots of $\det A(z) = 0$ lie inside the unit circle in the z plane, which insures that A will have a causal and stable inverse.)

The prediction \tilde{X}_n can be considered to be produced by passing the signal X_n through an LTI filter with Kronecker delta response p_k where $p_l = -a_l$, $l = 1, 2, \dots, m$ and $p_l = 0$ otherwise. (*prediction filter*):

$$a_k = \delta_k - p_k$$

If Fourier transforms $X(f) = \sum_n X_n e^{-i2\pi n f}$, $E(f) = \sum_n \epsilon_n e^{-i2\pi n f}$ exist, then

$$E(f) = X(f)A(f)$$

or

$$X(f) = \frac{E(f)}{A(f)}$$

\Rightarrow If LTI filter $A(f)$ is invertible, X_n can be perfectly recovered from its prediction residuals:

$$X_n \rightarrow \boxed{A(f)} \xrightarrow[\text{residual, excitation}]{\epsilon_n = \sum_{k=0}^m a_k X_{n-k}} \boxed{1/A(f)} \rightarrow X_n = \epsilon_n - \sum_{l=1}^m a_l X_{n-l}$$

$$X_n \rightarrow \boxed{P(f)} \longrightarrow \tilde{X}_n = \sum_{k=0}^m p_k X_{n-k}$$

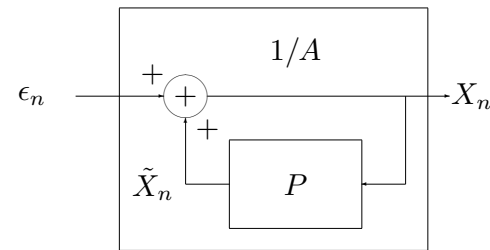
where $P(f) = \sum_{n=0}^m p_n e^{-i2\pi n f} = - \sum_{n=1}^m a_n e^{-i2\pi n f}$

Prediction error sequence can be viewed as filtering the input sequence by a filter with Kronecker delta response a_l ; $l = 0, 1, \dots, m$, the *prediction error filter*

$$X_n \rightarrow \boxed{A(f)} \longrightarrow \epsilon_n = \sum_{k=0}^m a_k X_{n-k}$$

prediction error filter or *inverse filter* $A(f) = \sum_{n=0}^m a_n e^{-i2\pi n f}$

By construction, $P(f) = 1 - A(f)$ so the picture becomes



$$\text{i.e., } E(f) = A(f)X(f) = (1 - P(f))X(f) \Rightarrow X(f) = E(f) + P(f)X(f)$$

This picture relating the input, prediction residuals, and predictions will later play an important role in motivating quantization schemes.

Quality of predictor is measured by the energy of the prediction error

$$\epsilon_n = X_n - \tilde{X}_n,$$

the mean-squared error (MSE)

$$D_m(a) = E(\epsilon_n^2) = E(|X_n - \tilde{X}_n|^2),$$

where m is the predictor order or memory and $a = (a_1, a_2, \dots, a_m)$ is the vector of linear prediction coefficients

an m th order linear predictor is optimum if it minimizes the MSE or, equivalently, maximizes the prediction gain defined by

$$10 \log_{10} \frac{\sigma_X^2}{D_m(a)}$$

on a dB scale.

which for the optimal predictor reduces to

$$D_m(a) = \sum_{n=0}^m a_n r(n)$$

Can also be stated in the frequency domain:

Define the power spectral density

$$S_\epsilon(f) = \mathcal{F}_f(R_\epsilon) \equiv \sum_n R_\epsilon(n) e^{-i2\pi n f}$$

Denote this optimum by $D_m = \inf_a D_m(a)$

Since process assumed stationary, $D_m(a)$ does not depend on n and optimal a is given as in the vector case by (87) and $D_m = \alpha_m$

The stationarity of the process and the vector results imply that

$$\begin{aligned} D_m(a) &= R_\epsilon(0) = a^t R_{m+1} a \\ &= \sum_{l=0}^m \sum_{n=0}^m a_l a_n r(l-n) \\ &= \sum_{l=0}^m a_l \underbrace{\left(\sum_{n=0}^m a_n r(n-l) \right)}_{=0 \text{ for } \ell=1,2,\dots,m} \end{aligned}$$

also have

$$\begin{aligned} D_m(a) &= \int_{-1/2}^{1/2} S_\epsilon(f) df \\ &= \int_{-1/2}^{1/2} |A(f)|^2 S(f) df \end{aligned}$$

where

$$A(f) = \mathcal{F}_f(a) = \sum_n a_n e^{-i2\pi n f}$$

is the transfer function of the prediction error filter

Infinite order one-step linear prediction

Since constraint set gets larger as m grows, D_m is a nondecreasing sequence and has a limit, which can be shown to be the minimal MSE using an infinite-order predictor

$$D_\infty \equiv \lim_{m \rightarrow \infty} D_m = \inf_m D_m$$

i.e., the limit of D_m is the minimal error using an infinite order filter.

The asymptotic theory of Toeplitz matrices (R_m is a Toeplitz matrix since $R_m(k, j)$ depends only on $k - j$) can be used to show that

$$\lim_{m \rightarrow \infty} D_m = e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S(f) df} = D_\infty$$

(See, e.g., *Toeplitz and Circulant Matrices: a Review, Foundations and Trends in Communications and Information Theory*, vol.2, no. 3, pp. 155-329, 2006, <http://ee.stanford.edu/~gray/toeplitz.pdf>.)

In the case of $m = \infty$, the orthogonality principal implies that $E(\epsilon_n X_{n-k}) = 0$, $k = 1, 2, \dots$ and hence that ϵ_n is uncorrelated with all previous X_n

Since $\epsilon_n = \sum_{k=0}^{\infty} a_k X_{n-k}$, this implies also that ϵ_n is uncorrelated with all previous ϵ_k

$$(E(\epsilon_n \epsilon_k) = 0 \text{ for } k < n)$$

This implies that ϵ_n is an uncorrelated process, i.e., is *discrete time white noise* with $S_\epsilon(f) = c$ for some constant c for all f .

Then

$$D_\infty = \int_{-1/2}^{1/2} S_\epsilon(f) df = c$$

so that

$$S_\epsilon(f) = |A(f)|^2 S(f) = D_\infty; \text{ all } f$$

or

$$S(f) = \frac{D_\infty}{|A(f)|^2}$$

which is the causal spectral factorization of the spectrum $S(f)$

(there is always an $A(f)$ of the desired form satisfying this relation provided that $D_\infty > 0$, i.e., the process is nondeterministic in the Wiener sense)

This points out

- the goal of optimal prediction is to produce a prediction error filter that *whitens* the error process, and
- optimal prediction is intimately connected with *modeling* a random process, the ideal prediction error filter is found from the causal spectral factorization of the original psd

The optimal finite order prediction also can be viewed as attempting to *whiten the prediction error as much as possible* given the limited number of coefficients.

This can be made quantitative with the idea of a *spectral flatness measure* of a process, defined for ϵ_n by

$$\gamma_\epsilon \triangleq \frac{e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_\epsilon(f) df}}{\int_{-\frac{1}{2}}^{\frac{1}{2}} S_\epsilon(f) df}$$

from Jensen's inequality,

$$\ln D_m(a) = \ln \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} S_\epsilon(f) df \right) \geq \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_\epsilon(f) df$$

with equality iff $S_\epsilon(f)$ is a constant, or

$$\gamma_\epsilon = \frac{e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_\epsilon(f) df}}{D_m(a)} \leq 1$$

with equality iff $S_\epsilon(f)$ is a constant, i.e., white.

The closer to 1 the spectral flatness measure, the more “white” the process can be considered

The expression for γ_ϵ can be simplified further using a classic result from orthogonal polynomials or from the theory of Toeplitz forms:

It turns out that regardless of $\{a_k; k = 0, \dots, m\}$ ($a_0 = 1$),

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S_\epsilon(f) df &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln |A(f)|^2 S(f) df \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S(f) df \\ &= \ln D_\infty \end{aligned}$$

or, equivalently, for any $\{a_k; k = 0, \dots\}$ with $a_0 = 1$,

$$\int \ln |A(f)|^2 df = 0$$

This result follows from the theory of orthogonal polynomials or the asymptotic theory of Toeplitz matrices.

Intuitively, this result says that the prediction error process has the same minimum MSE as does the original process.

To explain this result a bit further:

Suppose that the optimal one-step predictor error for a process with power spectral density $S(f)$ is given by $D_\infty = e^{\int_{-\frac{1}{2}}^{\frac{1}{2}} \ln S(f) df}$.

This optimization problem was shown to be the same as the following problem:

Find the causal and causally invertible filter with delta-response $a_0 = 1, a_1, a_2, a_3, \dots$ that produces the output with minimum energy when a process with power spectral density $S(f)$ is the input.

Suppose now that a is not necessarily the optimal (minimum energy) filter, but the output residual error process ϵ_n has power spectral density $S_\epsilon(f)$:

Suppose that ϵ_n is put into a causal and causally invertible filter, say B with $b_0 = 1$ with minimum output variance process δ_n which has energy $e^{\int_{-1/2}^{1/2} \ln S_\epsilon(f) df}$

Suppose that this is less than D_∞ . If that is the case, than the cascade of the two filters A and B is also causal (and causally invertible) and it has lead coefficient 1 in its delta-response (e.g., the convolution of the causal sequence a with the causal sequence b will have output at time 0 equal to $a_0 b_0 = 1$).

so that *minimizing the finite order prediction error over a is exactly equivalent to making the spectral flatness measure as close to 1 as possible*, i.e., whitening the prediction error as much as possible.

Thus the cascade AB gives an output energy less than the minimal energy over all such filters, which is not possible. Thus

$$e^{\int_{-1/2}^{1/2} \ln S_\epsilon(f) df} \geq e^{\int_{-1/2}^{1/2} \ln S(f) df}$$

A similar argument shows the inequality in the other direction, proving the claim that

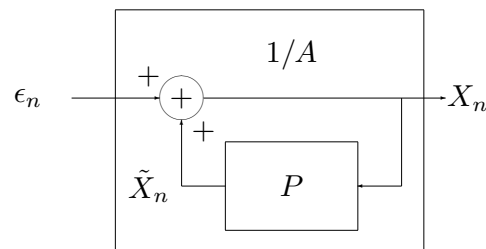
$$\int_{-1/2}^{1/2} \ln S_\epsilon(f) df = \int_{-1/2}^{1/2} \ln S(f) df$$

\Rightarrow for any a the spectral flatness measure is

$$\gamma_\epsilon = \frac{e^{\int_{-1/2}^{1/2} \ln S_\epsilon(f) df}}{D_m(a)} = \frac{D_\infty}{D_m(a)}$$

Modeling and Prediction

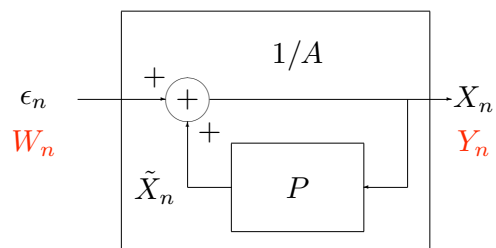
Recall the representation



where now know that ϵ_n should be as nearly white as possible (perfectly in the limit of large m)

Suggests that X_n can be modeled as a linear filter driven by approximately white noise

If replace true prediction residuals ϵ_n by a white noise process W_n , then the resulting output Y_n should have second order properties close to those of the true X_n and hence “sound like” X_n



If instead a white random process is used, then the output will be a random process with approximately the second order properties of the original X_n

1. View a sample sequence of the process X_n
2. Based on the observations, estimate the autocorrelation of the sequence out to lag m for an n th order model, e.g., given a sample sequence $\{x_n; n = 0, 1, \dots, L - 1\}$, form

$$\hat{R}(k) \equiv \frac{1}{L - k} \sum_{l=k}^{L-1} x_l x_{l-k}; \quad k = 0, 1, \dots, m$$

3. Find $a = (a_0, a_1, \dots, a_m)$ yielding $D_m(a) = D_m$ (this is usually done using the Levinson-Durbin algorithm and the autocorrelation)
4. The process is modeled as the output of a linear filter $1/A(f)$ with a white noise input of variance D_m , the optimal prediction error variance

Turns out that

$$R_Y(k) = R(k); \quad k = 0, 1, \dots, m$$

— correlation matching

This provides a paradigm for producing models of random processes which inherit the second order properties of the original process:

This modeling idea is used in several compression systems, including systems for waveform coding, parametric coding, and hybrids.

Resulting model is an m th order autoregressive model of the form

$$X_n = W_n + \sum_{k=1}^m X_{n-k}$$

where W_n is a white process with variance $\sigma_W^2 = \alpha_m$.

In speech processing, one typically performs the linear prediction analysis for a window of 50-100ms of speech to generate the filter parameters. Substituting a locally generated white process for the true prediction residuals should synthesize understandable speech.

Modeling by autoregressive sources

The LP theory provides an approach to constructing an autoregressive model (A_m, α_m) for observed data. Autoregressive models have played an important role in speech processing and other signal processing applications such as are found in geophysics and medical signal processing.

Briefly mention several other approaches that yield essentially the same results.

process will be close to the X_n process in some sense, in particular to have as similar an autocorrelation or psd as possible

Suppose have a distortion measure $d(S, S_Y)$ (or $d(R, R_Y)$) that measures the distortion or cost of approximating an original spectrum S (or autocorrelation R) by an approximating spectrum S_Y (or autocorrelation R_Y), then the best model in the class \mathcal{A}_m of all m th order autoregressive processes for an original process could be chosen by choosing the one which minimized $d(S, S_Y)$

Minimum Distortion Model Selection

The selection of an optimal predictor can also be viewed as a “minimum distortion” modeling problem.

Suppose that the original process $\{X_n\}$ has an autocorrelation function R and psd S

It is desired to model this process by an m th order autoregressive process $\{Y_n\}$, i.e., a process formed by passing white noise z_n with variance σ_Z^2 through a linear filter of the form

$$Y_n = Z_n - \sum_{k=1}^m a_k Y_{n-k}$$

the goal is to choose the regression coefficients a_k so that the $\{Y_n\}$

One such distortion measure was proposed by Itakura and Saito and now bears their name:

$$d(S, S_Y) = \int_{-1/2}^{1/2} \left(\frac{S(f)}{S_Y(f)} - \ln \frac{S(f)}{S_Y(f)} - 1 \right) df$$

The particular form has many information theoretic and other interpretations (related to relative entropy, minimum discrimination information)

It can be shown using linear systems theory that the distortion can be expressed in the time domain for any process in \mathcal{A}_m with variance σ_Y^2 and regression coefficients a_1, \dots, a_m as

$$d(S, S_Y) = \frac{a^t R_{m+1} a}{\sigma_Y^2} - \ln \frac{D_m}{\sigma_Y^2} - 1$$

where as usual $a = (1, a_1, \dots, a_m)^t$

Maximum Entropy View

Suppose we have estimate \hat{R}_m of correlations to lag m of stationary random process X_n .

What m -step Markov random process maximizes the Shannon differential entropy rate:

$$h(X) = \lim_{n \rightarrow \infty} \frac{1}{n} h(X^n)$$

where

$$h(X^n) = - \int f_{X^n}(x^n) \log f_{X^n}(x^n) dx^n$$

Since assume Markov, $h(X) = h(X_m | X^m)$.

Note: No Gaussian **assumption**, stated as a *variational problem*.

Predictive quantization

An application of linear prediction to quantization: predictive quantization, DPCM

Instead of scalar quantization of $\{X_n\}$, use linear prediction based on past to predict next step, form prediction error or residual signal, then quantize residual signal.

Decoder reconstructs estimate of X_n from quantized residuals.

Heuristic motivation: prediction error residuals should have less information, provide more efficient quantization

Problem: Decoder has only quantized residuals, not true past residuals.

Answer: If n and R_n are fixed, then largest differential entropy is (surprise!) obtained by Gaussian density as

$$h(X^n) = \frac{1}{2} \log(2\pi e)^n |R_n|$$

⇒ MAXDET problem, which has a long history and large literature.

As $n \rightarrow \infty$, wish to maximize $h(X_m | X^m)$, accomplished by a Gaussian density with variance

$$\sigma_{X_m | X^m}^2 = \frac{|R_{m+1}|}{|R_m|} = \alpha_m$$

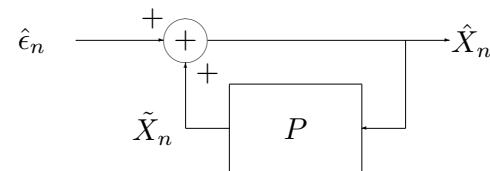
achieved by m th order autoregressive process with psd

$$S(f) = \alpha_m / |A(f)|^2$$

LP problem again

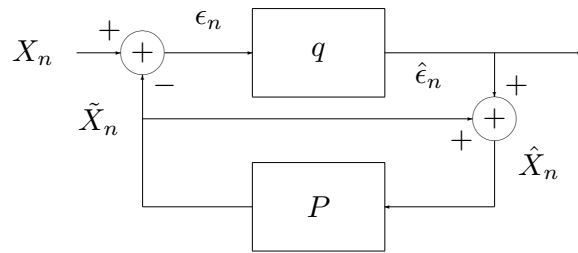
Design the prediction filter assuming perfect residuals, i.e., prediction residuals based on original signal

Substitute the quantized error $\hat{\epsilon} = q(\epsilon)$ for ϵ_n in the model, so that the decoder becomes



The decoder produces the possible reproduction waveforms, which should look like original signals (at least in second order properties)

The corresponding encoder is then



Basic DPCM or predictive quantizer

Intuition is that if quantization rate is high, prediction residuals based on quantized past should behave much like prediction residuals based on true past, so that predictor remains good for actual error sequence

coefficients from past reproductions. No need for side information. Smaller delay.

Common to add adaptive postfilter to filter signal to “hide” quantization noise. E.g., for speech coding, noise masked at formant peaks. Postfilter enhances areas where noise masked, diminishes area where noise strong

Note: ADPCM is a *waveform coder*, tries to reproduce the original waveform.

ADPCM: adaptive DPCM, adaptive predictive quantization

Can adapt either predictor P to local signal behavior (e.g., using LP techniques), or quantizer q (typically step size of uniform quantizer)

Predictor can be adapted in either of two ways:

forward adaptation compute the predictor coefficients based on data to be coded for the next frame or block of data. Send both prediction parameters (side information) and quantizer output sequence

backward adaptation compute the predictor coefficients based on the previous frame of data, which the decoder already knows and has reproduction. Compute based on *reproduction*, not original signal.

With backward adaptation, decoder can compute new predictor

Linear Predictive Coding (LPC)

Example of a *vocoder*, not a waveform coder. Extracts a description of the signal which is quantized, sent, and locally synthesized. Waveforms do not resemble each other, but second order statistics do.

Originally developed by Itakura and Saito (1966) using ML approach, but popularized as LPC by Atal and others using LP approach (1969).

Itakura and Atal considered primary pioneers.

Fascinating history with an impact on the development of the IP network protocol.

Most early work at NTT, Bell, Speech Communications Research Lab, UCSB

Used heavily in secure digital telephony.

First commercially successful product: Speak and Spell, TI. Also arguably the first DSP chip.

Model a “frame” of speech based on samples $x = (x_0, x_1, \dots, x_{N-1})$ ($\approx 25 - 50\text{ms}$)

1. estimate the autocorrelation $R(k)$ for $k = 0, 1, \dots, m$ (typically $m = 10$), then find the optimal predictor P or prediction error filter $A = 1 - P$

A byproduct of the optimization is the “gain” or prediction error variance $\alpha_m = E[(X_n - \tilde{X}_n)^2]$

2. quantize (or vector quantize) the parameters describing the model, e.g., a_1, \dots, a_m and α_m

often other equivalent parameters are quantized, e.g.,

- reflection coefficients
- autocorrelation coefficients
- inverse filter coefficients
- cepstrum
- line spectral pairs (LSP)

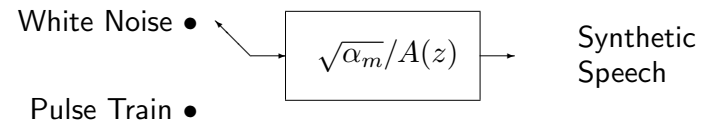
Mathematically equivalent when no quantization, but quantization errors effect each differently.

3. decide if speech is voiced (e.g., vowel) or unvoiced (e.g., plosives or “shhhh”) If voiced, estimate pitch (many methods)
4. transmit bits to describe parameters to receiver: prediction coefficients, gain, voicing, and pitch

Provides an autoregressive process (all-pole) $\sqrt{\alpha_m}/A(z)$ with

$$A(z) = 1 + \sum_{k=1}^M a_k z^{-k}.$$

5. reconstruction: if voiced, connect pulse sequence of estimated period to A , if unvoiced drive by white noise of variance = gain. The output of the filter is the reconstructed speech.



Simplistic: no voicing or pitch estimation details

Find model (α_m, A) as before. Coding occurs when the final model is selected from a discrete set, e.g., quantize separate parameters or parameter vector. Local synthesis at decoder.

Not a waveform coder, no attempt made to actually reconstruct waveform. A new signal is synthesized which attempts to approximate original local spectrum. A *vocoder*

Atal and Shroeder used LP methods to design ADPCM (APC), a residual excited LPC (later variations called "RELPC") New predictor was designed for each window and used in predictive quantizer. (1967, 1968). Open-loop residual quantization.

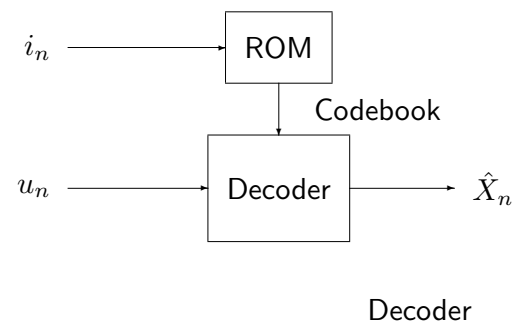
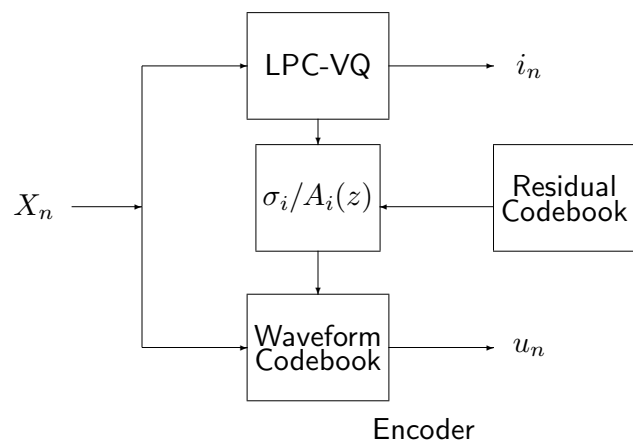
Code Excited Linear Prediction (CELP)

Modern version of analysis-by-synthesis. Closed-loop residual quantization.

Codebook for residuals. Given input vector, find residual codeword that produces best closed loop match out of LPC (or LPC-VQ) to original input.

Effectively produces a waveform codebook for searching by passing residual codewords through model linear filter.

Stewart et al. (1981, 1982), Shroeder and Atal (1984, 1985)



Atal, Gersho & others made practicable by incorporating perceptual weighting filters, i.e., perceptually meaningful distortion measures in search and adaptive postprocessing to improve quality.

Dual rate systems often combine LPC for low rate with CELP for high rate. Many variations.