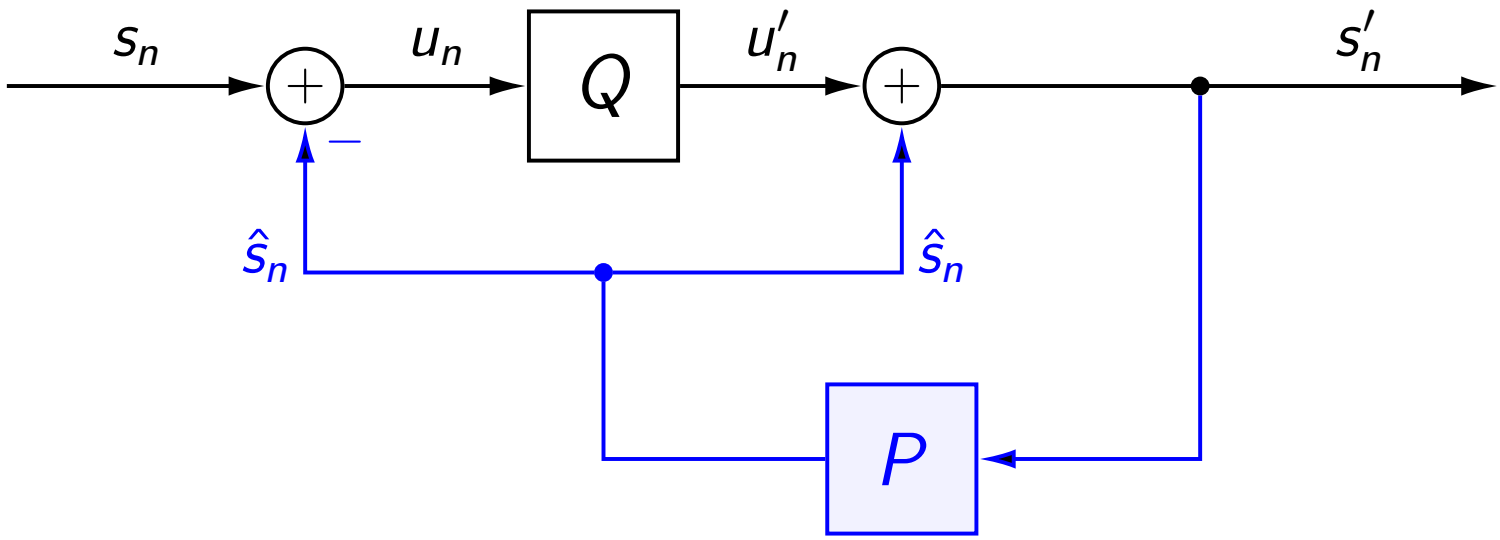
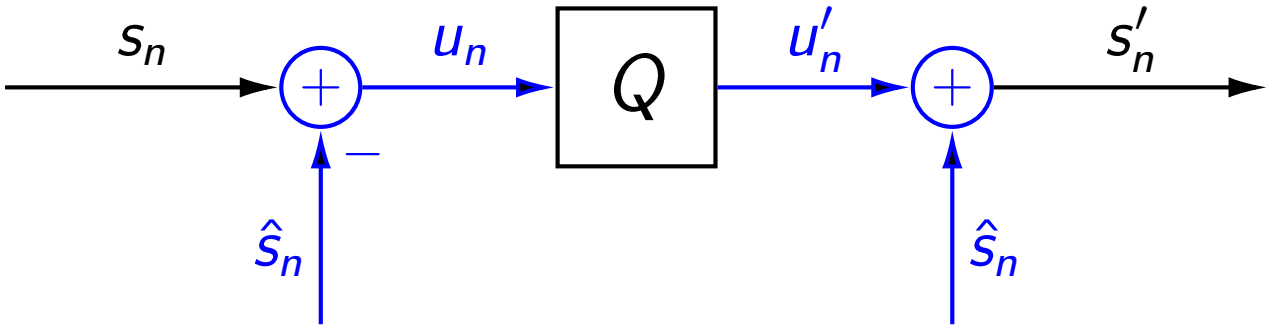
Predictive Coding



Idea of Predictive Coding



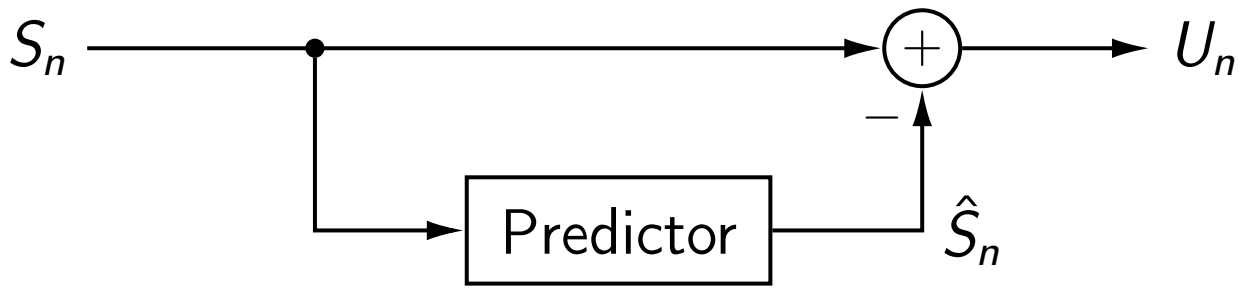
Idea: Reduce dependencies before quantization

1. Predict a sample using an estimate (function of past samples)
2. Quantize residual between sample and its prediction
3. Add quantized residual and prediction to obtain reconstructed sample

Questions

- How to predict a sample using past samples?
- How to combine prediction and quantization?

Prediction: Statistical Estimation Procedure



- Select **set of observed random variables** $\mathcal{B}_n$
  - ➡ Typically: N preceding random variables

$$\mathcal{B}_n = \{S_{n-1}, S_{n-2}, \cdots, S_{n-N}\} \tag{1}$$

- **Predictor for** S_n: Deterministic function of observation set $\mathcal{B}_n$

$$\hat{S}_n = A(\mathcal{B}_n) \tag{2}$$

- Prediction error

$$U_n = S_n - \hat{S}_n = S_n - A(\mathcal{B}_n) \tag{3}$$

Distortion in Predictive Coding

- Signal modification due to prediction

$$\begin{aligned} u_k &= s_k - \hat{s}_k \quad \Longrightarrow \quad s_k = u_k + \hat{s}_k \\ s'_k &= u'_k + \hat{s}_k \end{aligned}$$

- Additive distortion measures (MSE distortion for $p = 2$)

$$\begin{aligned} d_N(\boldsymbol{s}, \boldsymbol{s'}) &= \frac{1}{N} \sum_{k=0}^{N-1} |s_k - s'_k|^p = \frac{1}{N} \sum_{k=0}^{N-1} |u_k + \hat{s}_k - u'_k - \hat{s}_k|^p \\ &= \frac{1}{N} \sum_{k=0}^{N-1} |u_k - u'_k|^p = d_N(\boldsymbol{u}, \boldsymbol{u'}) \end{aligned} \tag{4}$$

➡ **Distortion between original and reconstructed samples is equal to distortion between original and reconstructed prediction residuals**

Performance of Predictive Coding

Operational distortion-rate function $D(R)$ for predictive coding

- Same distortion in original and prediction error domain

➡ $D(R)$ is equal to operational distortion-rate function $D_U(R)$ for (scalar) quantization of prediction residuals

- Operational distortion-rate function for scalar quantization

$$D(R) = D_U(R) = \sigma_U^2 \cdot g(R) \tag{5}$$

with σ_U^2: variance of the prediction residual
$g(R)$: depends only on the type of the distribution of the residuals

Design criterion for predictor

- Neglect dependency on the distribution type of prediction residual

➡ Define: **Optimal predictor $A(\mathcal{B}_n)$ for given observation set $\mathcal{B}_n$ minimizes prediction error variance σ_U^2**

Optimal Prediction

Typical optimization criterion

- Minimize prediction error energy

$$\varepsilon_U^2 = \mathrm{E}\{\, U_n^2 \,\} = \mathrm{E}\Big\{ \big(S_n - \hat{S}_n \big)^2 \Big\} = \mathrm{E}\Big\{ \big(S_n - A(\mathcal{B}_n) \big)^2 \Big\} \tag{6}$$

- Note: Minimization of prediction error energy

$$\begin{aligned} \varepsilon_U^2 &= \mathrm{E}\{\, U_n^2 \,\} = \mathrm{E}\{\, (U_n - \mu_U + \mu_U)^2 \,\} \\ &= \mathrm{E}\{\, (U_n - \mu_U)^2 \,\} + \mathrm{E}\{\, \mu_U^2 \,\} + 2\,\mu_U\,\mathrm{E}\{\, U_n - \mu_U \,\} \\ &= \sigma_U^2 + \mu_U^2 \end{aligned} \tag{7}$$

  implies minimization of variance σ_U^2 and mean μ_U

Solution of minimization problem

- Conditional mean (see proof in [Wiegand and Schwarz])

$$\hat{S}_n = A(\mathcal{B}_n) = \mathrm{E}\{\, S_n \,|\, \mathcal{B}_n \,\} \tag{8}$$

➡ General case requires storage of large tables (impractical and complex)

Optimal Prediction for Autoregressive Sources

Autoregressive sources

- Model for random processes with dependencies between samples
- Autoregressive process of order m (AR(m) process) is given by

$$S_n = \mu_S + \sum_{k=1}^{m} a_k \cdot (S_{n-k} - \mu_S) + Z_n \tag{9}$$

$$= \mu_S \cdot (1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m) + \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(m)} + Z_n \tag{10}$$

where

- μ_S is the mean of the AR(m) process
- $\{Z_n\}$ is a zero-mean iid process
- $\boldsymbol{S}_{n-1}^{(m)} = (S_{n-1}, \cdots, S_{n-m})^{\mathrm{T}}$ is the vector of the past m samples
- $\boldsymbol{a}_m = (a_1, \cdots, a_m)^{\mathrm{T}}$ is a constant parameter vector
- $\boldsymbol{e}_m = (1, \cdots, 1)^{\mathrm{T}}$ is an m-dimensional unit vector

Optimal Prediction for Autoregressive Sources

- AR(m) process

$$S_n = \mu_S \left(1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m\right) + \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(m)} + Z_n \tag{11}$$

- Consider observation set of $N \geq m$ past samples

$$\mathcal{B}_n = \boldsymbol{S}_{n-1}^{(N)} = (S_{n-1}, S_{n-2}, \cdots, S_{n-N})^{\mathrm{T}} \qquad \text{with} \qquad N \geq m \tag{12}$$

➡ **Optimal predictor** for observation set $\mathcal{B}_n$

$$\begin{aligned}
\mathrm{E}\{\, S_n \,|\, \mathcal{B}_n \,\} &= \mathrm{E}\left\{ \mu_S \left(1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m\right) + \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(m)} + Z_n \;\middle|\; \boldsymbol{S}_{n-1}^{(N)} \right\} \\
&= \mu_S \left(1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m\right) + \boldsymbol{a}_m^{\mathrm{T}} \mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(m)} \;\middle|\; \boldsymbol{S}_{n-1}^{(N)} \right\} + \mathrm{E}\left\{ Z_n \;\middle|\; \boldsymbol{S}_{n-1}^{(N)} \right\} \\
&= \mu_S \left(1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m\right) + \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(m)}
\end{aligned} \tag{13}$$

➡ **Optimal predictor is an affine function of the past m samples**

Affine Prediction

- Introduce structural constraints yielding simple predictor
- Given: Any observation vector of K past samples

$$\boldsymbol{S}_n^K = (S_{n-a}, S_{n-b}, \cdots)^{\mathrm{T}} \qquad \text{with} \qquad a > 0,\ b > 0, \cdots \tag{14}$$

➡ **Affine predictor**

$$\hat{S}_n = A(\,\boldsymbol{S}_n^K\,) = h_0 + \boldsymbol{h}_K^{\mathrm{T}} \cdot \boldsymbol{S}_n^K \tag{15}$$

where $\boldsymbol{h}_K = (h_1, \cdots, h_K)^{\mathrm{T}}$ is a constant vector and h_0 a constant offset

- Variance σ_U^2 of prediction residual only depends on $\boldsymbol{h}_K$

$$\begin{aligned}
\sigma_U^2(h_0, \boldsymbol{h}_K) &= \mathrm{E}\Big\{ \big(U_n - \mathrm{E}\{\,U_n\,\}\big)^2 \Big\} \\
&= \mathrm{E}\Big\{ \Big(S_n - h_0 - \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{S}_n^K - \mathrm{E}\Big\{\, S_n - h_0 - \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{S}_n^K \,\Big\}\Big)^2 \Big\} \\
&= \mathrm{E}\Big\{ \Big(S_n - \mathrm{E}\{\,S_n\,\} - \boldsymbol{h}_K^{\mathrm{T}}\,\Big(\boldsymbol{S}_n^K - \mathrm{E}\Big\{\,\boldsymbol{S}_n^K\,\Big\}\Big)\Big)^2 \Big\}
\end{aligned} \tag{16}$$

Affine Prediction

- Mean squared prediction error ε_U^2

$$\begin{aligned}\varepsilon_U^2(h_0, \boldsymbol{h}_K) &= \sigma_U^2(\boldsymbol{h}_K) + \mu_U^2(h_0, \boldsymbol{h}_K) \\ &= \sigma_U^2(\boldsymbol{h}_K) + \mathrm{E}\Big\{ S_n - h_0 - \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{S}_n^K \Big\}^2 \\ &= \sigma_U^2(\boldsymbol{h}_K) + \big(\mu_S\,(1 - \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{e}_K) - h_0\big)^2 \end{aligned} \tag{17}$$

➡ Mean squared prediction error ε_U^2 is minimized if we

- Choose $\boldsymbol{h}_k$ for minimizing the variance σ_U^2
- Set the constant offset h_0 equal to

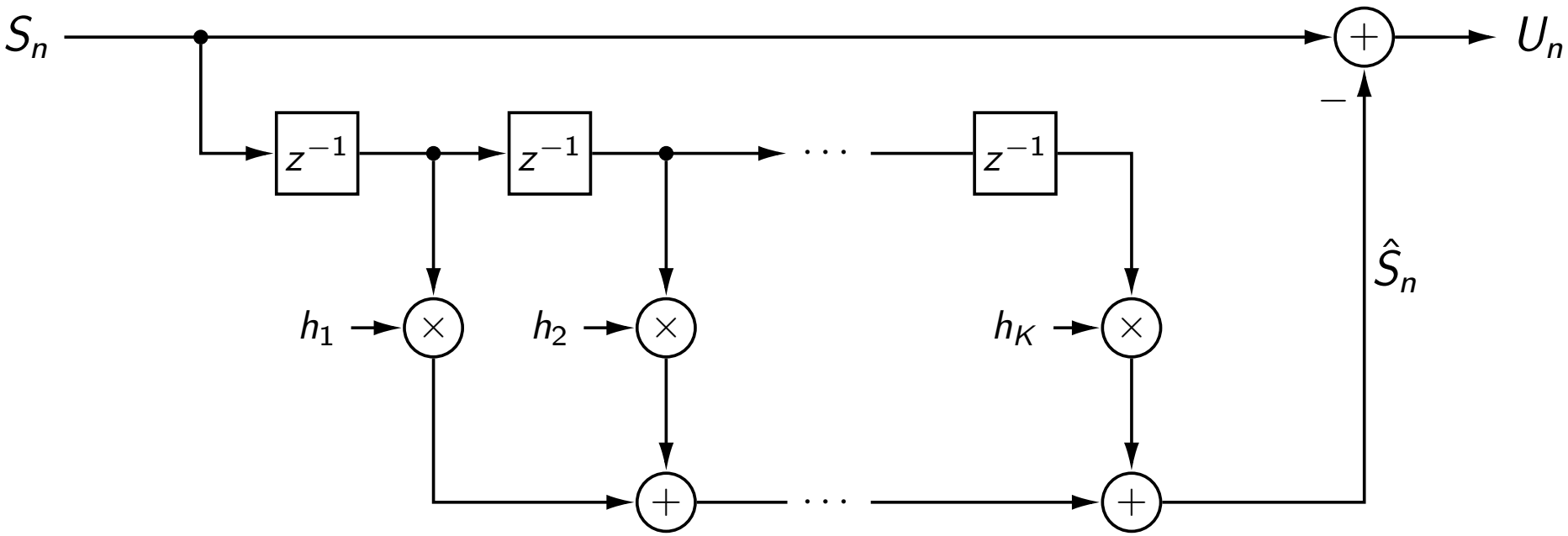
$$h_0 = \mu_S\,(1 - \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{e}_K) = \mu_S\,\left(1 - \sum_{k=1}^{K} h_k\right) \tag{18}$$

➡ Can restrict our considerations to **linear predictors**

$$\hat{S}_n = A(\,\boldsymbol{S}_n^K\,) = \boldsymbol{h}_K^{\mathrm{T}} \cdot \boldsymbol{S}_n^K \tag{19}$$

and **minimization of prediction error variance** σ_U^2

Linear Prediction



- Linear predictor

$$\hat{S}_n = \boldsymbol{h}_K^{\mathrm{T}} \cdot \boldsymbol{S}_n^K = \sum_{i=1}^{K} h_i \cdot S_{n-i}$$

$$= h_1 \cdot S_{n-1} + h_2 \cdot S_{n-2} + \cdots + h_K \cdot S_{n-K} \tag{20}$$

Linear Prediction

- Simplify notation for observation vector

$$\boldsymbol{S}_n^K = \boldsymbol{S}_K = (S_{n-a}, S_{n-b}, \cdots)^{\mathrm{T}} \tag{21}$$

- Linear predictor

$$\hat{S}_n = \boldsymbol{h}_K^{\mathrm{T}} \cdot \boldsymbol{S}_K \tag{22}$$

- Prediction error

$$U_n = S_n - \hat{S}_n = S_n - \boldsymbol{h}_K^{\mathrm{T}} \cdot \boldsymbol{S}_K \tag{23}$$

➡ Variance of prediction error

$$\begin{aligned}
\sigma_U^2(\boldsymbol{h}_K) &= \mathrm{E}\Big\{ \left(U_n - \mathrm{E}\{ U_n \}\right)^2 \Big\} \\
&= \mathrm{E}\Big\{ \Big(S_n - \boldsymbol{h}_K^{\mathrm{T}} \boldsymbol{S}_K - \mathrm{E}\Big\{ S_n - \boldsymbol{h}_K^{\mathrm{T}} \boldsymbol{S}_K \Big\}\Big)^2 \Big\} \\
&= \mathrm{E}\Big\{ \Big(\left(S_n - \mathrm{E}\{ S_n \}\right) - \boldsymbol{h}_K^{\mathrm{T}} \left(\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \}\right)\Big)^2 \Big\}
\end{aligned}$$

Linear Prediction

- Variance of prediction error

$$\begin{aligned}
\sigma_U^2(\boldsymbol{h}_K) &= \mathrm{E}\Big\{ \big((S_n - \mathrm{E}\{ S_n \}) - \boldsymbol{h}_K^{\mathrm{T}} (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \})\big)^2 \Big\} \\
&= \mathrm{E}\Big\{ (S_n - \mathrm{E}\{ S_n \})^2 \Big\} \\
&\quad - 2\, \boldsymbol{h}_K^{\mathrm{T}} \cdot \mathrm{E}\big\{ (S_n - \mathrm{E}\{ S_n \}) (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \}) \big\} \\
&\quad + \boldsymbol{h}_K^{\mathrm{T}} \cdot \mathrm{E}\Big\{ (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \}) (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \})^{\mathrm{T}} \Big\} \cdot \boldsymbol{h}_K^{\mathrm{T}} \\
&= \sigma_S^2 - 2\, \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{c}_n + \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{C}_K\, \boldsymbol{h}_K \qquad (24)
\end{aligned}$$

with $\boldsymbol{C}_K$ and $\boldsymbol{c}_n$ being given by

$$\boldsymbol{C}_K = \mathrm{E}\Big\{ (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \}) (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \})^{\mathrm{T}} \Big\} \qquad (25)$$

$$\boldsymbol{c}_n = \mathrm{E}\big\{ (S_n - \mathrm{E}\{ S_n \}) (\boldsymbol{S}_K - \mathrm{E}\{ \boldsymbol{S}_K \}) \big\} \qquad (26)$$

- ➡ $\boldsymbol{C}_K$: Auto-covariance matrix of observation set $\boldsymbol{S}_K$
- ➡ $\boldsymbol{c}_n$: Cross-covariance vector of observation set $\boldsymbol{S}_K$ and sample S_n

Linear Prediction

- Linear prediction given an observation set $\boldsymbol{S}_K$

$$\hat{S}_n = \boldsymbol{h}_K^{\mathrm{T}} \, \boldsymbol{S}_K$$

➡ **Variance of prediction error signal** $U_n = S_n - \hat{S}_n$

$$\sigma_U^2 \;=\; \sigma_S^2 - 2\,\boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{c}_n + \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{C}_K\,\boldsymbol{h}_K$$

$$\text{with} \quad \boldsymbol{C}_K \;=\; \mathrm{E}\Big\{ \big(\boldsymbol{S}_K - \mathrm{E}\{\,\boldsymbol{S}_K\,\}\big)\,\big(\boldsymbol{S}_K - \mathrm{E}\{\,\boldsymbol{S}_K\,\}\big)^{\mathrm{T}} \Big\}$$

$$\boldsymbol{c}_n \;=\; \mathrm{E}\big\{\, S_n\,\big(\boldsymbol{S}_K - \mathrm{E}\{\,\boldsymbol{S}_K\,\}\big)\,\big\}$$

➡ Similarly: **Mean squared prediction error**

$$\varepsilon_U^2 \;=\; \varepsilon_S^2 - 2\,\boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{r}_n + \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{R}_K\,\boldsymbol{h}_K$$

$$\text{with} \quad \boldsymbol{R}_K \;=\; \mathrm{E}\Big\{\, \boldsymbol{S}_K\,\boldsymbol{S}_K^{\mathrm{T}} \,\Big\} = \boldsymbol{C}_K + \mu_S^2\,\boldsymbol{e}_K\boldsymbol{e}_K^{\mathrm{T}}$$

$$\boldsymbol{r}_n \;=\; \mathrm{E}\{\, S_n\,\boldsymbol{S}_K\,\} = \boldsymbol{c}_n + \mu_S^2\,\boldsymbol{e}_K$$

Optimal Linear Prediction

- Minimize prediction error variance

$$\sigma_U^2(\boldsymbol{h}_k) = \sigma_S^2 - 2\boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{c}_n + \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{C}_K\,\boldsymbol{h}_K \tag{27}$$

➡ Set derivative with respect to $\boldsymbol{h}_K$ equal to zero

$$\frac{\partial}{\partial \boldsymbol{h}_K}\,\sigma_U^2 = -2\,\boldsymbol{c}_n + 2\,\boldsymbol{C}_K\,\boldsymbol{h}_K = \boldsymbol{0} \tag{28}$$

➡ **Yule-Walker equations** / **normal equations** (linear equation system)

$$\boxed{\boldsymbol{C}_K\,\boldsymbol{h}_K = \boldsymbol{c}_n} \tag{29}$$

➡ Optimal filter $\boldsymbol{h}_K$ can be derived by

- Determining auto-covariance matrix $\boldsymbol{C}_K$ for observation set $\boldsymbol{S}_K$
- Determining cross-covariance vector $\boldsymbol{c}_N$ between S_n and $\boldsymbol{S}_K$
- Solving linear equation system $\boldsymbol{C}_K \cdot \boldsymbol{h}_K = \boldsymbol{c}_N$

Prediction Error Variance for Optimal Linear Prediction

- Prediction error variance

$$\sigma_U^2(\boldsymbol{h}_k) = \sigma_S^2 - 2\boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{c}_n + \boldsymbol{h}_K^{\mathrm{T}}\,\boldsymbol{C}_K\,\boldsymbol{h}_K$$

- Optimal prediction coefficients $\boldsymbol{h}^*$ are given by solution of

$$\boldsymbol{C}_K \cdot \boldsymbol{h}_K^* = \boldsymbol{c}_n$$

➡ Prediction error variance for optimal prediction

$$\begin{aligned}
\sigma_U^2(\boldsymbol{h}_K^*) &= \sigma_S^2 - 2\,(\boldsymbol{h}_K^*)^{\mathrm{T}}\boldsymbol{c}_n + (\boldsymbol{h}_K^*)^{\mathrm{T}}\boldsymbol{C}_K\,\boldsymbol{h}_K^* \\
&= \sigma_S^2 - 2\,(\boldsymbol{h}_K^*)^{\mathrm{T}}\boldsymbol{c}_n + (\boldsymbol{h}_K^*)^{\mathrm{T}}\boldsymbol{c}_n \\
&= \sigma_S^2 - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{h}_K^* \qquad (30) \\
&= \sigma_S^2 - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_K^{-1}\boldsymbol{c}_n \qquad (31)
\end{aligned}$$

➡ Optimal prediction: Signal variance σ_S^2 is reduced by $\boldsymbol{c}_n^T\boldsymbol{C}_K^{-1}\boldsymbol{c}_n = \boldsymbol{c}_n^T\boldsymbol{h}_N^*$

Example: AR(1) Process with $K = 1$

- Consider AR(1) process with correlation coefficient ϱ
- Prediction using only directly preceding sample

$$\hat{S}_n = h_1 \cdot S_{n-1}$$

➡ Auto-covariance matrix and cross-covariance vector

$$\boldsymbol{C}_1 = \mathrm{E}\{ (S_{n-1} - \mathrm{E}\{ S_{n-1} \}) (S_{n-1} - \mathrm{E}\{ S_{n-1} \}) \} = \sigma_S^2$$

$$\boldsymbol{c}_n = \mathrm{E}\{ (S_n - \mathrm{E}\{ S_n \}) (S_{n-1} - \mathrm{E}\{ S_{n-1} \}) \} = \varrho \cdot \sigma_S^2$$

➡ Yule-Walker equations

$$\sigma_S^2 \cdot h_1 = \varrho \cdot \sigma_S^2$$

➡ Optimal prediction

$$h_1 = \varrho \qquad \Longrightarrow \qquad \hat{S}_n = \varrho \cdot S_{n-1}$$

➡ Resulting prediction error variance

$$\sigma_U^2 = \sigma_S^2 - h_1 \, c_n = \sigma_S^2 \, (1 - \varrho^2)$$

Example: AR(1) Process with $K = 2$

- Consider AR(1) process with correlation coefficient ϱ
- Prediction using the two directly preceding sample

$$\hat{S}_n = h_1 \cdot S_{n-1} + h_2 \cdot S_{n-2}$$

➡ Yule-Walker equations

$$\sigma_S^2 \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \sigma_S^2 \begin{bmatrix} \varrho \\ \varrho^2 \end{bmatrix}$$

➡ Optimal prediction

$$h_1 = \varrho \qquad \text{and} \qquad h_2 = 0$$

➡ No improvement relative to prediction with only previous samples

The Orthogonality Principle

- Property for optimal affine predictors

$$\begin{aligned}
\mathrm{E}\{\, U_n \boldsymbol{S}_K \,\} &= \mathrm{E}\Big\{ \big(S_n - h_0 - \boldsymbol{h}_K^{\mathrm{T}} \boldsymbol{S}_K\big)\, \boldsymbol{S}_K \Big\} \\
&= \mathrm{E}\{\, S_n \boldsymbol{S}_K \,\} - h_0\, \mathrm{E}\{\, \boldsymbol{S}_K \,\} - \mathrm{E}\Big\{ \boldsymbol{S}_K \boldsymbol{S}_K^{\mathrm{T}} \Big\}\, \boldsymbol{h}_K \\
&= \boldsymbol{c}_n + \mu_S^2\, \boldsymbol{e}_K - h_0\, \mu_S\, \boldsymbol{e}_K - \big(\boldsymbol{C}_K + \mu_S^2\, \boldsymbol{e}_K\, \boldsymbol{e}_K^{\mathrm{T}}\big)\, \boldsymbol{h}_K \\
&= \boldsymbol{c}_n - \boldsymbol{C}_K \boldsymbol{h}_K + \mu_S\, \boldsymbol{e}_K \big(\, \mu_S\, (1 - \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{e}_K) - h_0 \big) \qquad (32)
\end{aligned}$$

- Inserting the optimal solutions

$$\boldsymbol{h}_K^* = \boldsymbol{C}_K^{-1} \cdot \boldsymbol{c}_n \qquad \text{and} \qquad h_0^* = \mu_S\, (1 - \boldsymbol{h}_K^{\mathrm{T}}\, \boldsymbol{e}_K) \qquad (33)$$

yields

$$\mathrm{E}\{\, U_n \boldsymbol{S}_K \,\} = \boldsymbol{0} \qquad (34)$$

➡ For optimal affine prediction, the correlation between the observation vector and the prediction residual is zero

One-Step Linear Prediction

- Observation set: K directly preceding samples

$$\mathcal{B}_n = \boldsymbol{S}_K = (S_{n-1}, S_{n-2}, \cdots, S_{n-K})^{\mathrm{T}} \tag{35}$$

- Covariances

$$\phi_k = \mathrm{E}\{ \left(S_n - \mathrm{E}\{ S_n \}\right)\left(S_{n+k} - \mathrm{E}\{ S_{n+k} \}\right) \} \tag{36}$$

➡ Yule-Walker / normal equations

$$\boldsymbol{C}_K \cdot \boldsymbol{h}_K = \boldsymbol{c}_n \tag{37}$$

➡ Normal equations in matrix form

$$\begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_{K-1} \\ \phi_1 & \phi_0 & \cdots & \phi_{K-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K-1} & \phi_{K-2} & \cdots & \phi_0 \end{bmatrix} \begin{bmatrix} h_1^K \\ h_2^K \\ \vdots \\ h_K^K \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_K \end{bmatrix} \tag{38}$$

One-Step Linear Prediction

- Normal equations in matrix form ($\boldsymbol{C}_K \cdot \boldsymbol{h}_K = \boldsymbol{c}_n$)

$$\begin{bmatrix} \phi_0 & \phi_1 & \cdots & \phi_{K-1} \\ \phi_1 & \phi_0 & \cdots & \phi_{K-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{K-1} & \phi_{K-2} & \cdots & \phi_0 \end{bmatrix} \begin{bmatrix} h_1^K \\ h_2^K \\ \vdots \\ h_K^K \end{bmatrix} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_K \end{bmatrix} \tag{39}$$

- Changing the equation to: $\boldsymbol{c}_n - \boldsymbol{C}_k \cdot \boldsymbol{h}_k = \boldsymbol{0}$

$$\begin{bmatrix} \phi_1 & \phi_0 & \phi_1 & \cdots & \phi_{K-1} \\ \phi_2 & \phi_1 & \phi_0 & \cdots & \phi_{K-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_K & \phi_{K-1} & \phi_{K-2} & \cdots & \phi_0 \end{bmatrix} \begin{bmatrix} 1 \\ -h_1^K \\ -h_2^K \\ \vdots \\ -h_K^K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{40}$$

One-Step Linear Prediction: Augmented Normal Equations

- Remember: Prediction error variance for optimal prediction using K preceding samples

$$\sigma_K^2 = \sigma_S^2 - \boldsymbol{c}_n^{\mathrm{T}}\, \boldsymbol{C}_K^{-1}\, \boldsymbol{c}_n = \sigma_S^2 - \boldsymbol{c}_n^{\mathrm{T}}\, \boldsymbol{h}_K \tag{41}$$

$$= \phi_0 - h_1\phi_1 - h_2\phi_2 - \cdots - h_K\phi_K \tag{42}$$

➡ Add as another equation (first row) to our equation system

$$\underbrace{\begin{bmatrix} \phi_0 & \phi_1 & \phi_2 & \cdots & \phi_K \\ \phi_1 & \phi_0 & \phi_1 & \cdots & \phi_{K-1} \\ \phi_2 & \phi_1 & \phi_0 & \cdots & \phi_{K-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_K & \phi_{K-1} & \phi_{K-2} & \cdots & \phi_0 \end{bmatrix}}_{\boldsymbol{C}_{K+1}} \underbrace{\begin{bmatrix} 1 \\ -h_1^K \\ -h_2^K \\ \vdots \\ -h_K^K \end{bmatrix}}_{\boldsymbol{a}_K} = \begin{bmatrix} \sigma_K^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{43}$$

- The resulting equation system is called **augmented normal equations**

Optimal One-Step Prediction: Prediction Error Variance

- One set of augmented normal equations for each size K of the observation set

➡ Combine augmented normal equation for $K = 0, 1, 2, \cdots, N$ into one linear equation system

$$\boldsymbol{C}_{N+1} \begin{bmatrix} \boldsymbol{a}_N & \boldsymbol{a}_{N-1} & \cdots & \boldsymbol{a}_1 & \boldsymbol{a}_0 \end{bmatrix} = \begin{bmatrix} \cdots \end{bmatrix} \tag{44}$$

➡ Resulting equation system (X: arbitrary values)

$$\boldsymbol{C}_{N+1} \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -h_1^N & 1 & \ddots & 0 & 0 \\ -h_2^N & -h_1^{N-1} & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ -h_N^N & -h_{N-1}^{N-1} & \cdots & -h_1^1 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_N^2 & X & \cdots & X & X \\ 0 & \sigma_{N-1}^2 & \ddots & X & X \\ 0 & 0 & \ddots & X & X \\ \vdots & \vdots & \ddots & \sigma_1^2 & X \\ 0 & 0 & \cdots & 0 & \sigma_0^2 \end{bmatrix}$$

Optimal One-Step Prediction: Prediction Error Variance

$$\boldsymbol{C}_{N+1} \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -h_1^N & 1 & \ddots & 0 & 0 \\ -h_2^N & -h_1^{N-1} & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ -h_N^N & -h_{N-1}^{N-1} & \cdots & -h_1^1 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_N^2 & X & \cdots & X & X \\ 0 & \sigma_{N-1}^2 & \ddots & X & X \\ 0 & 0 & \ddots & X & X \\ \vdots & \vdots & \ddots & \sigma_1^2 & X \\ 0 & 0 & \cdots & 0 & \sigma_0^2 \end{bmatrix}$$

➡ Taking the determinate of both side yields

$$|\boldsymbol{C}_{N+1}| = \sigma_N^2 \cdot \sigma_{N-1}^2 \cdot \ldots \cdot \sigma_2^2 \cdot \sigma_1^2 \cdot \sigma_0^2 \tag{45}$$

➡ **Prediction error variance σ_N^2 for optimal prediction using the N directly preceding samples**

$$\sigma_N^2 = \frac{|\boldsymbol{C}_{N+1}|}{|\boldsymbol{C}_N|} \tag{46}$$

One-Step Prediction for Autoregressive Processes

- Recall: AR(m) process with mean μ_S and $\boldsymbol{a}_m = (a_1, \cdots, a_m)^{\mathrm{T}}$

$$S_n = Z_n + \mu_S(1 - \boldsymbol{a}_m^T \boldsymbol{e}_m) + \boldsymbol{a}_m^T \boldsymbol{S}_{n-1}^{(m)} \tag{47}$$

- Consider: Prediction using $N \geq m$ preceding samples, define

$$\boldsymbol{a}_N = (a_1, \cdots, a_m, 0, \cdots, 0)^{\mathrm{T}} \tag{48}$$

➡ Prediction error for affine prediction

$$\begin{aligned} U_n &= S_n - h_0 - \boldsymbol{h}_N^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(N)} \\ &= Z_n + \mu_S(1 - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{e}_m) + \boldsymbol{a}_m^T \boldsymbol{S}_{n-1}^{(m)} - h_0 - \boldsymbol{h}_K^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(N)} \\ &= Z_n + \mu_S(1 - \boldsymbol{a}_N^{\mathrm{T}} \boldsymbol{e}_N) - h_0 + \left(\boldsymbol{a}_N - \boldsymbol{h}_N\right)^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(N)} \end{aligned} \tag{49}$$

➡ Prediction error for affine prediction with optimal offset $h_0 = \mu_S\,(1 - \boldsymbol{h}_N^{\mathrm{T}}\,\boldsymbol{e}_N)$

$$\begin{aligned} U_n &= Z_n - \mu_S(\boldsymbol{a}_N - \boldsymbol{h}_N)^{\mathrm{T}}\,\boldsymbol{e}_N + \left(\boldsymbol{a}_N - \boldsymbol{h}_N\right)^{\mathrm{T}} \boldsymbol{S}_{n-1}^{(N)} \\ &= Z_n + \left(\boldsymbol{S}_{n-1}^{(N)} - \mu_S\,\boldsymbol{e}_N\right)^{\mathrm{T}}\,\left(\boldsymbol{a}_N - \boldsymbol{h}_N\right) \end{aligned} \tag{50}$$

One-Step Prediction for Autoregressive Processes

- Prediction error for affine prediction with optimal offset

$$U_n = Z_n + \left(\boldsymbol{S}_{n-1}^{(N)} - \mu_S \, \boldsymbol{e}_N \right)^{\mathrm{T}} \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \tag{51}$$

➡ Correlation between prediction error and observation vector

$$\begin{aligned}
\mathrm{E}\left\{ U_n \, \boldsymbol{S}_{n-1}^{(N)} \right\} &= \mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(N)} \, Z_n \right\} + \mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(N)} \left(\boldsymbol{S}_{n-1}^{(N)} - \mu_S \, \boldsymbol{e}_N \right)^{\mathrm{T}} \right\} \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \\
&= \left(\mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(N)} \left(\boldsymbol{S}_{n-1}^{(N)} \right)^{\mathrm{T}} \right\} - \mu_S \, \mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(N)} \right\} \boldsymbol{e}_N^{\mathrm{T}} \right) \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \\
&= \left(\mathrm{E}\left\{ \boldsymbol{S}_{n-1}^{(N)} \left(\boldsymbol{S}_{n-1}^{(N)} \right)^{\mathrm{T}} \right\} - \mu_S^2 \, \boldsymbol{e}_N \boldsymbol{e}_N^{\mathrm{T}} \right) \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \\
&= \boldsymbol{C}_N \cdot \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right)
\end{aligned} \tag{52}$$

One-Step Prediction for Autoregressive Processes

- Correlation between prediction error and observation vector

$$\mathrm{E}\Big\{ U_n \, \boldsymbol{S}_{n-1}^{(N)} \Big\} = \boldsymbol{C}_N \cdot \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \tag{53}$$

➡ Orthogonality principle

$$\mathrm{E}\Big\{ U_n \, \boldsymbol{S}_{n-1}^{(N)} \Big\} = 0 \tag{54}$$

➡ **Optimal prediction coefficients for $N \geq m$**

$$\boldsymbol{h}_N = \boldsymbol{a}_N = (a_1, \cdots, a_m, 0, \cdots, 0)^{\mathrm{T}} \tag{55}$$

➡ Resulting prediction error

$$\begin{aligned} U_n &= Z_n + \left(\boldsymbol{S}_{n-1}^{(N)} - \mu_S \, \boldsymbol{e}_N \right)^{\mathrm{T}} \left(\boldsymbol{a}_N - \boldsymbol{h}_N \right) \\ &= Z_n \qquad \text{(no remaining dependencies between samples)} \end{aligned} \tag{56}$$

One-Step Prediction for AR(1) Processes

- AR(1) process

$$S_n = Z_n + \mu_S(1-\varrho) + \varrho \cdot S_{n-1} \tag{57}$$

➡ Yule-Walker equations (for $K = 2$)

$$\sigma_S^2 \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \sigma_S^2 \begin{bmatrix} \varrho \\ \varrho^2 \end{bmatrix}$$

➡ Optimal prediction filter

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} \varrho \\ 0 \end{bmatrix}$$

➡ Resulting prediction error variance

$$\begin{aligned} \sigma_U^2 &= \sigma_S^2 - [h_1 \; h_2] \cdot \begin{bmatrix} \sigma_S^2 \varrho \\ \sigma_S^2 \varrho^2 \end{bmatrix} \\ &= \sigma_S^2 - \varrho \cdot \sigma_S^2 \varrho - 0 \cdot \sigma_S^2 \varrho^2 \\ &= \sigma_S^2 \cdot (1 - \varrho^2) \end{aligned}$$

One-Step Prediction for AR(1) Processes

- Prediction error variance for optimal prediction

$$\sigma_U^2 = \frac{|\boldsymbol{C}_2|}{|\boldsymbol{C}_1|} = \frac{\sigma_S^4 - \sigma_S^4\,\varrho^2}{\sigma_S^2} = \sigma_S^2\,(1-\varrho^2) \tag{58}$$

- Prediction residual for filter h_1

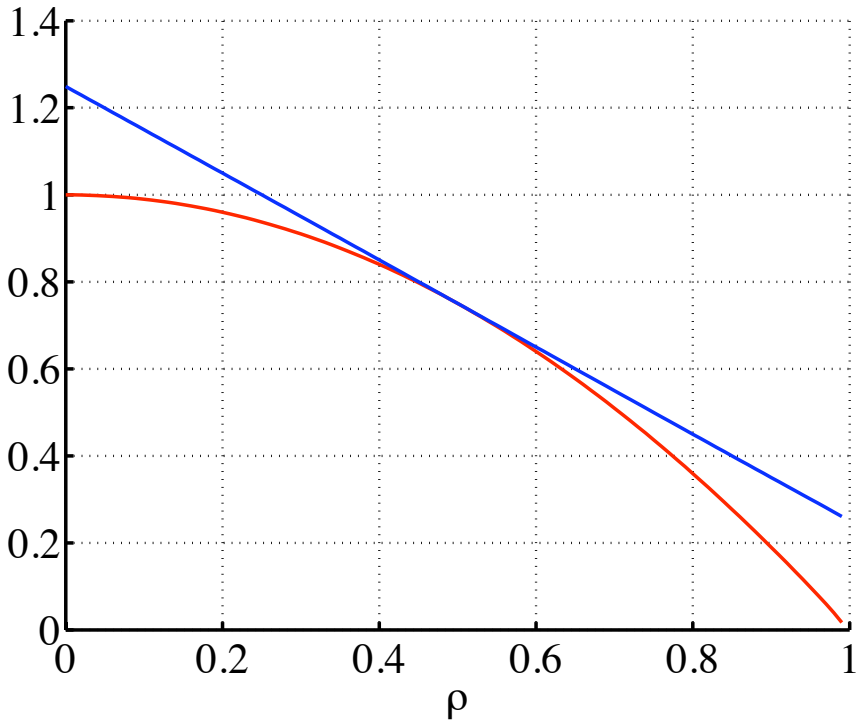
$$U_n = S_n - h_1\,S_{n-1}$$

- Prediction error variance

$$\begin{aligned}\sigma_U^2(h_1) &= \mathrm{E}\{\,(U_n - \mathrm{E}\{\,U_n\,\})^2\,\} \\ &= \sigma_S^2(1 + h_1^2 - 2\varrho h_1)\end{aligned}$$

- Note: Setting derivative to zero also yields the result $h_1^* = \varrho$

σ_U^2/σ_S^2 for $h = \varrho$ and $h = 0.5$



Prediction Gain

- Remember: High-rate approximation for scalar quantization

$$D(R) = \varepsilon^2 \cdot \sigma^2 \cdot 2^{-2R} \tag{59}$$

where ε^2 depends on quantization method and distribution type

Prediction Gain

- Improvement of predictive coding relative to direct quantization (high rates)
- Neglecting modification of distribution type (and usage of rec. samples)

➡ Prediction gain is defined according to

$$G_P = \frac{\sigma_S^2}{\sigma_U^2} \tag{60}$$

➡ Prediction gain for optimal linear prediction

$$G_P = \frac{\sigma_S^2}{\sigma_U^2} = \frac{\sigma_S^2}{\sigma_S^2 - \boldsymbol{c}_n^{\mathrm{T}}\, \boldsymbol{C}_N^{-1}\, \boldsymbol{c}_n} \tag{61}$$

Prediction Gain for AR(1) Processes

- Prediction gain for optimal prediction for AR(1) process

$$G_P(h^*) = \frac{\sigma_S^2}{\sigma_U^2} = \frac{\sigma_S^2}{\sigma_S^2\,(1-\varrho^2)} = \frac{1}{1-\varrho^2} \tag{62}$$

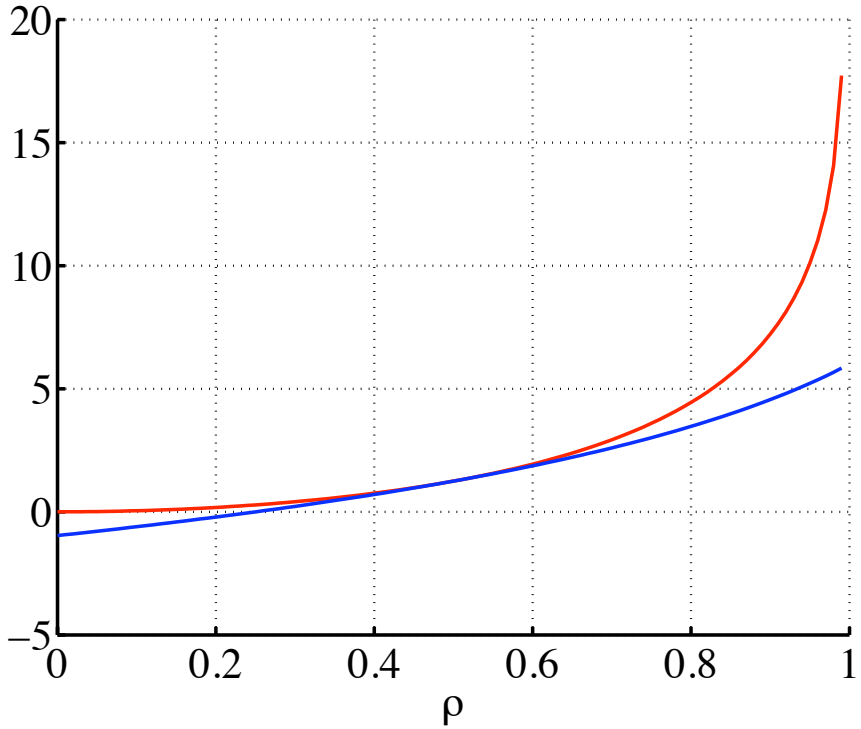
- Prediction gain for filter h_1

$$\begin{aligned} G_P(h_1) &= \frac{\sigma_S^2}{\sigma_S^2(1+h_1^2-2\varrho h_1)} \\ &= \frac{1}{1+h_1^2-2\varrho h_1} \end{aligned}$$

- At high bit rates:
  SNR improvement achieved by predictive coding

$$\Delta\text{SNR} = 10\log_{10} G_P$$

$10\log_{10} G_P(\varrho)$ for $h = \varrho$ and $h = 0.5$



Asymptotic Prediction Gain

Upper Bound for Prediction Gain

- Consider optimal one-step prediction using $N \to \infty$ preceding samples
- One-step prediction of a random variable S_n given the countably infinite set of preceding random variables $\{S_{n-1}, S_{n-2}, \cdots\}$ and $\{h_0, h_1, \cdots\}$

$$U_n = S_n - h_0 - \sum_{k=1}^{\infty} h_k \, S_{n-k} \tag{63}$$

- Orthogonality criterion: U_n is uncorrelated with all S_{n-k} for $k > 0$
- Furthermore, U_{n-k} for $k > 0$ is fully determined by a linear combination of past input values S_{n-k-i} for $i \geq 0$
- Hence, U_n is uncorrelated with U_{n-k} for $k > 0$

$$\phi_{UU}(k) = \sigma^2_{U,\infty} \cdot \delta(k) \qquad \Longleftrightarrow \qquad \Phi_{UU}(\omega) = \sigma^2_{U,\infty} \tag{64}$$

where $\sigma^2_{U,\infty}$ is the asymptotic one-step prediction error variance for $N \to \infty$

Asymptotic Prediction Error Variance

- For one-step prediction we showed

$$|\boldsymbol{C}_N| = \sigma_{N-1}^2 \cdot \sigma_{N-2}^2 \cdot \sigma_{N-3}^2 \cdot \ldots \cdot \sigma_0^2 \tag{65}$$

  which yields

$$\frac{1}{N} \ln |\boldsymbol{C}_N| = \ln |\boldsymbol{C}_N|^{\frac{1}{N}} = \frac{1}{N} \sum_{k=0}^{N-1} \ln \sigma_k^2 \tag{66}$$

- If a sequence of numbers $\alpha_0, \alpha_1, \alpha_2, \cdots$ approaches a limit α_∞, the average value approaches the same limit

$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \alpha_k = \alpha_\infty \tag{67}$$

- Hence, we can write

$$\lim_{N \to \infty} \ln |\boldsymbol{C}_N|^{\frac{1}{N}} = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} \ln \sigma_k^2 = \ln \sigma_{U,\infty}^2 \tag{68}$$

Asymptotic Prediction Error Variance

- Reformulating expression

$$\lim_{N\to\infty} \ln |\boldsymbol{C}_N|^{\frac{1}{N}} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \ln \sigma_k^2 = \ln \sigma_{U,\infty}^2 \tag{69}$$

yields

$$\sigma_{U,\infty}^2 = \exp\left(\lim_{N\to\infty} \ln |\boldsymbol{C}_N|^{\frac{1}{N}} \right) = \lim_{N\to\infty} |\boldsymbol{C}_N|^{\frac{1}{N}} \tag{70}$$

- Determinant of $N \times N$ matrix: Product of its eigenvalues $\xi_k^{(N)}$

$$\begin{aligned} \lim_{N\to\infty} |\boldsymbol{C}_N|^{\frac{1}{N}} &= \lim_{N\to\infty} \left(\prod_{k=0}^{N-1} \xi_k^{(N)} \right)^{\frac{1}{N}} = 2^{\lim_{N\to\infty} \log_2 \left(\prod_{k=0}^{N-1} \xi_k^{(N)} \right)^{\frac{1}{N}}} \\ &= 2^{\lim_{N\to\infty} \log_2 \left(\prod_{k=0}^{N-1} \xi_k^{(N)} \right)^{\frac{1}{N}}} = 2^{\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \log_2 \xi_k^{(N)}} \end{aligned} \tag{71}$$

Asymptotic Prediction Error Variance

- Asymptotic prediction error variance

$$\sigma^2_{U,\infty} = 2^{\left(\lim\limits_{N\to\infty} \frac{1}{N} \sum\limits_{k=0}^{N-1} \log_2 \xi_k^{(N)} \right)} \tag{72}$$

- Remember: Theorem of Grenander and Szegö

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} G\left(\xi_i^{(N)}\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G\left(\Phi_{SS}(\omega)\right) \, \mathrm{d}\omega \tag{73}$$

with $\Phi_{SS}(\omega)$ being the Fourier series of ϕ_k

$$\Phi_{SS}(\omega) = \sum_{k=-\infty}^{\infty} \phi_k \cdot e^{-\mathrm{i}\omega k} \quad \text{with} \quad \phi_k = \mathrm{E}\{ (S_n - \mu)(S_{n+k} - \mu) \} \tag{74}$$

Asymptotic Prediction Gain

➡ **Asymptotic prediction error variance**

$$\sigma_{U,\infty}^2 = \lim_{N\to\infty} |\boldsymbol{C}_N|^{\frac{1}{N}} = 2^{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_2 \Phi_{ss}(\omega)\,\mathrm{d}\omega\right)} \tag{75}$$

➡ **Asymptotic prediction error gain**

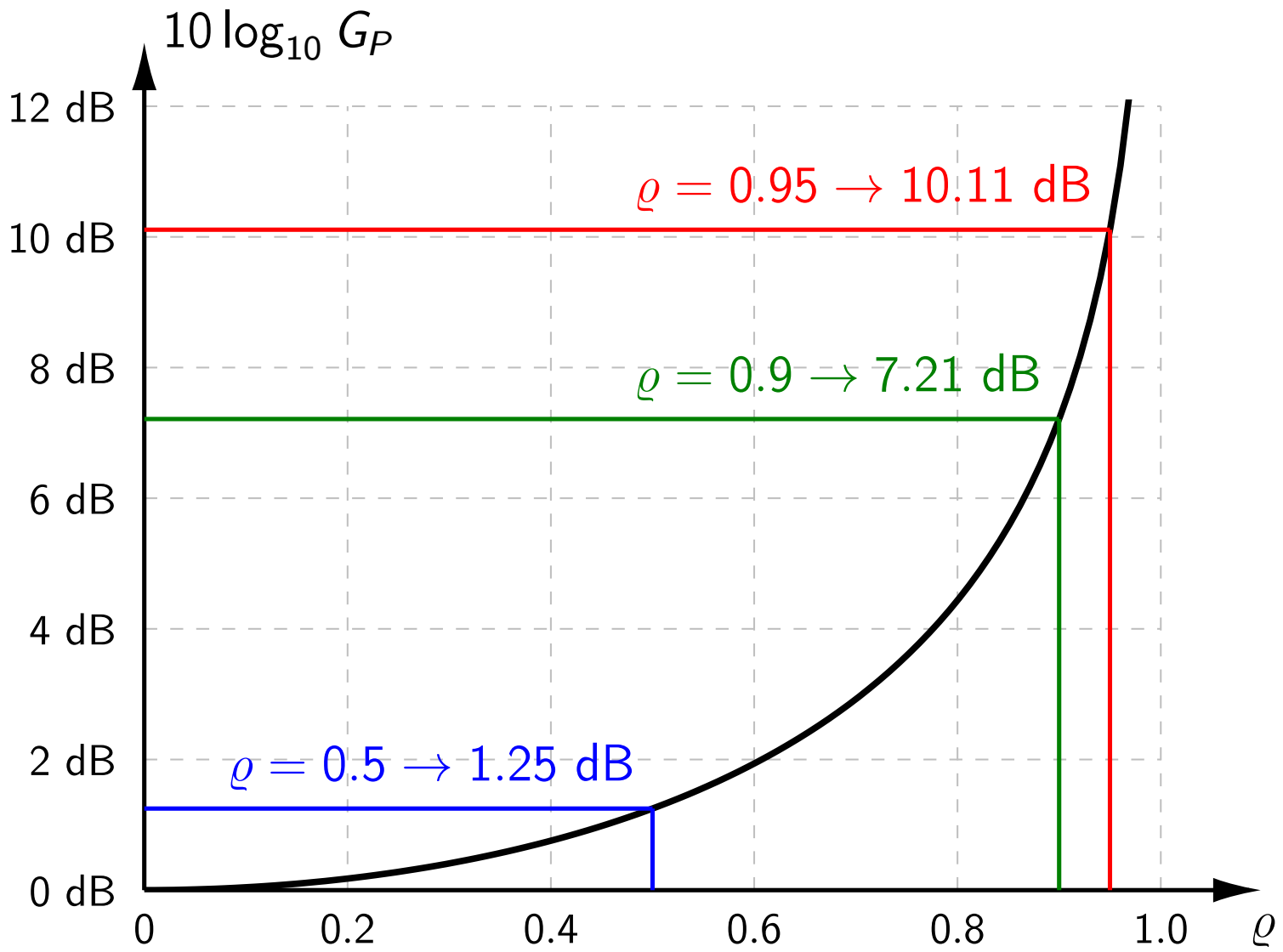
$$G_P^{\infty} = \frac{\sigma_S^2}{\sigma_{U,\infty}^2} = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{ss}(\omega)\,\mathrm{d}\omega}{2^{\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_2 \Phi_{ss}(\omega)\,\mathrm{d}\omega\right)}} \tag{76}$$

➡ Same expression as for asymptotic transform coding gain

- Asymptotic prediction gain for Markov processes (only previous sample)

$$G_P^{\infty} = \frac{1}{1-\varrho^2} \tag{77}$$

Asymptotic Prediction Gain for Markov Processes



Prediction in Images: Intra-Picture Prediction

- Picture samples are typically coded in raster-scan order
- Samples can be predicted using already coded samples

Example modes for picture prediction

- 1-d horizontal prediction:

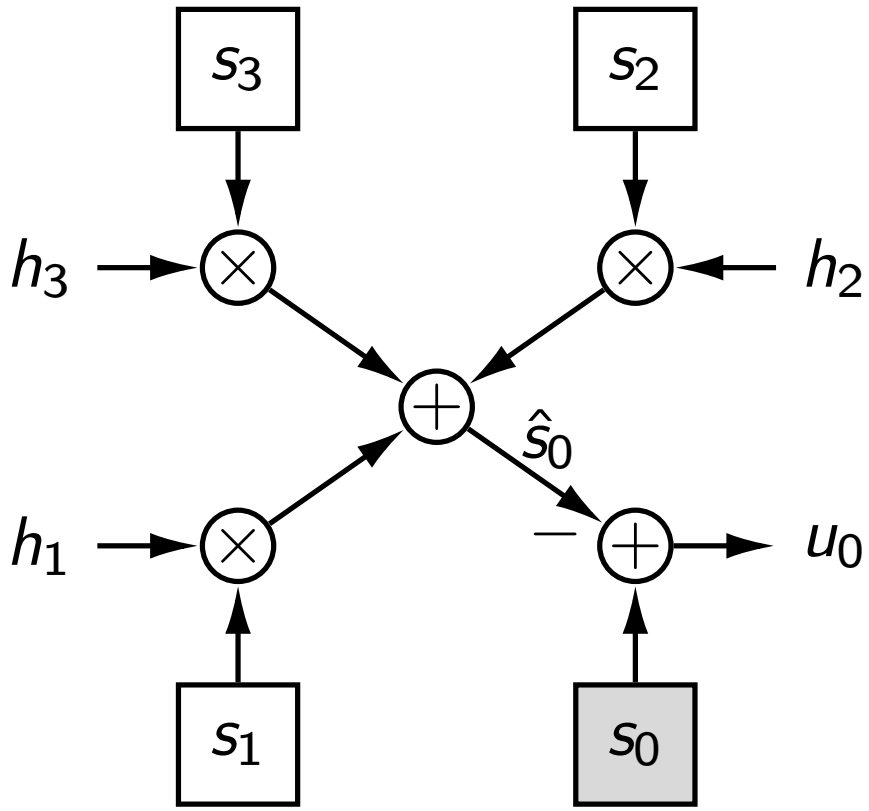
$$\hat{s}_0 = h_1 \cdot s_1$$

- 1-d vertical prediction:

$$\hat{s}_0 = h_2 \cdot s_2$$

- 2-d prediction:

$$\hat{s}_0 = \sum_{k=1}^{3} h_k s_k$$

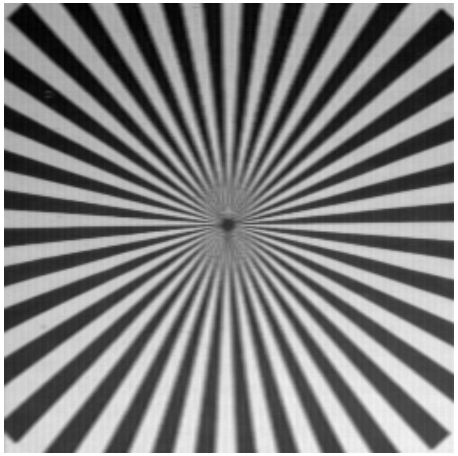


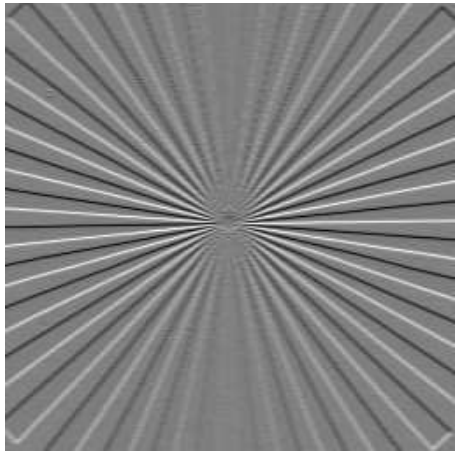
Prediction Example: Test Pattern

original

$\sigma_S^2 = 4925.81$

(using $\mu_S = 127$)





vert. prediction

$h_1 = 0$

$h_2 = 0.932$

$h_3 = 0$

$\sigma_U^2 = 646.67$

$G_P = 8.82$ dB

hor. prediction

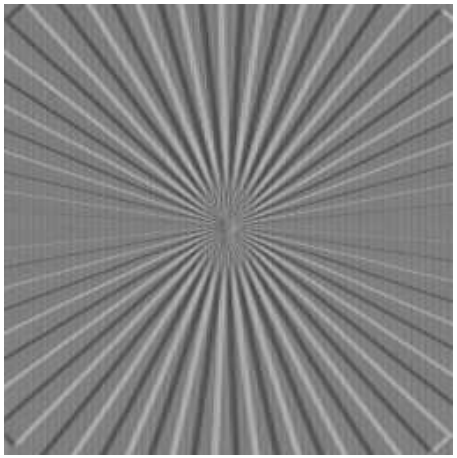
$h_1 = 0.953$

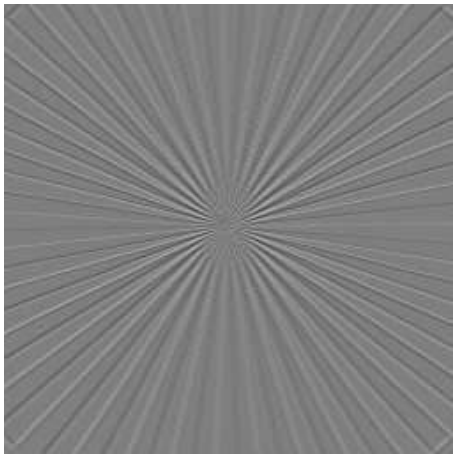
$h_2 = 0$

$h_3 = 0$

$\sigma_U^2 = 456.17$

$G_P = 10.33$ dB





2-d prediction

$h_1 = 0.911$

$h_2 = 0.871$

$h_3 = -0.788$

$\sigma_U^2 = 109.90$

$G_P = 16.51$ dB

Prediction Example: Picture "Lena"

256×256 **center**

$\sigma_S^2 = 2746.43$

(using $\mu_S = 127$)

vert. prediction

$h_1 = 0$

$h_2 = 0.977$

$h_3 = 0$

$\sigma_U^2 = 123.61$

$G_P = 13.47$ dB

hor. prediction

$h_1 = 0.962$

$h_2 = 0$

$h_3 = 0$

$\sigma_U^2 = 212.36$

$G_P = 11.12$ dB

2-d prediction

$h_1 = 0.623$

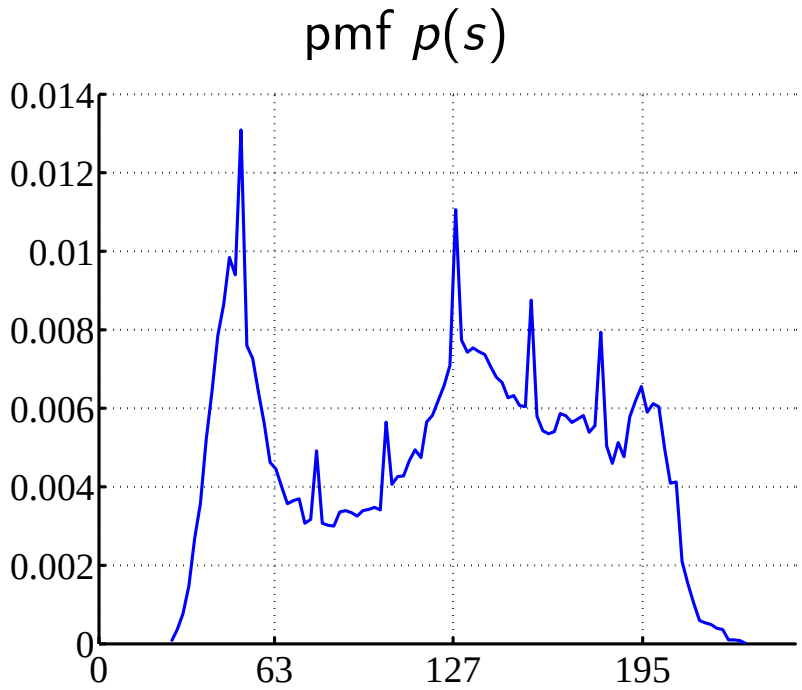
$h_2 = 0.835$

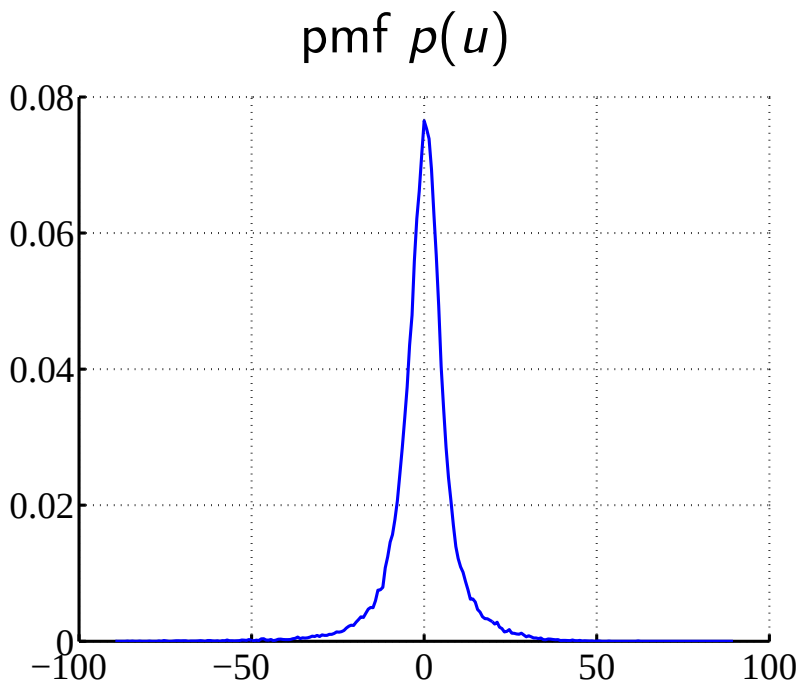
$h_3 = -0.480$

$\sigma_U^2 = 80.35$

$G_P = 15.34$ dB

Prediction Example: Histogram for Picture "Lena"





- Entropy changes significantly (rounding prediction to integer)

$$H(S) = 7.44\,\text{bits/sample} \qquad H(U) = 4.97\,\text{bits/sample}$$

➡ Prediction in Lossless Coding:

- Low-complexity alternative to conditional coding
- Examples: PNG, JPEG-LS

Summary

Prediction

- Estimate random variable from already observed random variables
- Optimal predictor: Conditional mean

Linear and affine prediction

- Simple and efficient structure
- Optimal predictor given by Yule-Walker equations
- AR(m) processes: Optimal predictor has m coefficients
- Optimal prediction error is orthogonal to input signal
- Non-matched predictor can increase signal variance