STANFORD

*Supplementary Lecture 3*

# Minimum Mean-Square Estimation

*April 6, 2026*

JOHN M. CIOFFI

Hitachi Professor Emeritus of Engineering

Instructor EE379B – Spring 2026

# Agenda

- Agenda
  - General minimum-mean-square error (MMSE) & Gaussian
    - Autocorrelation/Cross-Correlation
  - Linear MMSE & The Orthogonality Principle
    - Biases and SNRs

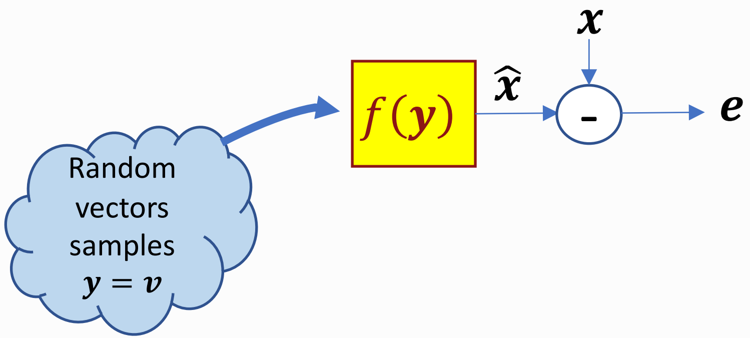
# General MMSE and Gaussian

*Section D.1*

**When all random processes are (stationary) Gaussian, Linear MMSE is the overall best MMSE.**

# The MMSE Estimation Problem

- Estimate $\boldsymbol{x}$ from $\boldsymbol{y}$ , $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$.
  - Designer knows $p_{\boldsymbol{x},\boldsymbol{y}}$ and specific observed $\boldsymbol{y} = \boldsymbol{v}$.
  - $\boldsymbol{x}$ and $\boldsymbol{y}$ 's distributions are continuous. (e.g., a random code on $\boldsymbol{x}$.)
- The **error** is:
  - $\boldsymbol{e} = \boldsymbol{x} - f(\boldsymbol{y})$.




- The **mean-square error (MSE)** is
  - $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\|\boldsymbol{e}\|^2] = \mathbb{E}[\|\boldsymbol{x} - f(\boldsymbol{y})\|^2]$.
- Its minimum over $f$ is the **MMSE**.

$$MMSE = \min_f \mathbb{E}\,[\|\boldsymbol{x} - f(\boldsymbol{y})\|^2].$$

- Solution [D.1] is
  - the conditional mean of $\boldsymbol{x}$ given $\boldsymbol{y}$; $\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}/\boldsymbol{y}]$,
  - from $p_{\boldsymbol{x}/\boldsymbol{y}}$ the $\grave{a}$ *posteriori* distribution (also used for MAP detector).
  - It's, it's an expectation, always linear: $(\widehat{\boldsymbol{x}_1 + \boldsymbol{x}_2}) = \hat{\boldsymbol{x}}_1 + \hat{\boldsymbol{x}}_2$.

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}/\boldsymbol{y}]$$
$$= W \cdot \boldsymbol{y} \text{ (if Gaussian)}$$

# Auto- & Cross- correlation

- **Autocorrelation** generalizes mean-square for random process. When stationary:
  - it samples $R_{\boldsymbol{xx}}(\tau) = \mathbb{E}[\boldsymbol{x}(t) \cdot \boldsymbol{x}^*(t-\tau)]$; where time is the dimension, and $\tau = kT'$ is **correlation interval**.
  - Vector process samples may correspond to:
    - frequency-time – **time samples** $x(kT)$ are the vector elements (Toeplitz).
    - space time usually sets $\tau = 0$ – **spatial samples** (think antennas),
      - often at each $n \cdot \left(\frac{1}{T}\right)$.

$$R_{\boldsymbol{xx}} = \mathbb{E}[\boldsymbol{x} \cdot \boldsymbol{x}^*] \quad R_{\boldsymbol{yy}} = \mathbb{E}[\boldsymbol{y} \cdot \boldsymbol{y}^*]$$

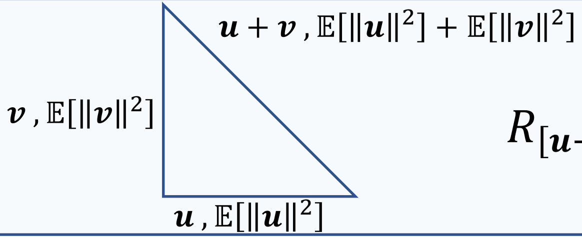
- **Energy** is over a symbol, including space dimensions:
  - Both cases $\tau = 0$ (non xtalk/energy) terms are on diagonal.

$$\mathcal{E}_x = trace\{R_{\boldsymbol{xx}}\} = \mathbb{E}[\boldsymbol{x}^* \cdot \boldsymbol{x}] = \mathbb{E}[\|\boldsymbol{x}\|^2]$$

- **Cross correlation** generalizes inner product:
  - It samples $R_{\boldsymbol{xy}}(\tau) = \mathbb{E}[\boldsymbol{x}(t) \cdot y^*(t-\tau)]$.
  - Vectors can be different lengths $L_x$ and $L_y$.
  - "uncorrelated" (=0) → **orthogonal.**
  - There are nondiagonal dimensions.

$$R_{\boldsymbol{xy}} = \mathbb{E}[\boldsymbol{x} \cdot \boldsymbol{y}^*] \qquad R_{\boldsymbol{yx}} = \mathbb{E}[\boldsymbol{y} \cdot \boldsymbol{x}^*]$$

- **Pythagorus** IF uncorrelated $R_{\boldsymbol{uv}} = 0$.
  - Generalizes "variances of uncorrelated random variables add."
  - $\mathbb{E}[\acute{\boldsymbol{x}} \cdot g(\boldsymbol{y})] = 0$ always, so $\boldsymbol{u} = \hat{\boldsymbol{x}}, \boldsymbol{v} = \boldsymbol{e}$ works.




$$R_{[\boldsymbol{u}+\boldsymbol{v}][\boldsymbol{u}+\boldsymbol{v}]} = R_{\boldsymbol{uu}} + R_{\boldsymbol{vv}}$$

# The Joint Gaussian Distribution

- is completely specified by its autocorrelation (and cross correlation) with zero mean:

$$R \triangleq R_{\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}[\boldsymbol{x}^* \quad \boldsymbol{y}^*]} = \begin{bmatrix} R_{\boldsymbol{xx}} & R_{\boldsymbol{xy}} \\ R_{\boldsymbol{yx}} & R_{\boldsymbol{yy}} \end{bmatrix}.$$

- A Gaussian's marginal distributions for $\boldsymbol{x}$ and $\boldsymbol{y}$
  - are also Gaussian.
- Its conditional distributions are Gaussian.
  - In particular with non-zero mean $\mathbb{E}[\boldsymbol{x}/\boldsymbol{y}]$.

$$\text{real:} \quad p(\boldsymbol{x},\boldsymbol{y}) = (2\pi)^{-\frac{N_x+N_y}{2}} \cdot |R|^{-1/2} \cdot e^{-\frac{1}{2}\left\{[\boldsymbol{x}^*\boldsymbol{y}^*]\cdot R^{-1}\cdot\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right\}}$$

$$\text{complex:} \quad p(\boldsymbol{x},\boldsymbol{y}) = (\pi)^{-[N_x+N_y]} \cdot |R|^{-1} \cdot e^{-\left\{[\boldsymbol{x}^*\boldsymbol{y}^*]\cdot R^{-1}\cdot\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\right\}}$$

- Singularity?
  - $|R_{\boldsymbol{yy}}| > 0$ with nonsingular noise on $\boldsymbol{y} = H \cdot \boldsymbol{x} + \boldsymbol{n}$.
  - $|R_{\boldsymbol{xx}}|$ ? $|R|$? – use pseudoinverse, and determinant redefines as product of **nonzero** eigenvalues.

$$\mathbb{E}[\boldsymbol{x}/\boldsymbol{y}] = \underbrace{R_{\boldsymbol{xy}} \cdot R_{\boldsymbol{yy}}^{-1}}_{W} \cdot \boldsymbol{y}$$

It's linear (for Gaussian).

$W = R_{\boldsymbol{xy}} \cdot R_{\boldsymbol{yy}}^{+}$ if $\boldsymbol{y}$ is singular.

**MMSE and AWGN's best transmission are fundamentally connected.**

# Linear MMSE & The Orthogonality Principle

*Section D.2*

**Decomposition into pass spaces and null spaces becomes critical in canonical design (think Vector Coding).**

# Linear MMSE: any joint distribution of $x$ and $y$

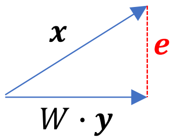
- Given random $\boldsymbol{x}$ and $\boldsymbol{y}$, receiver estimates $\boldsymbol{x}$, $\widehat{\boldsymbol{x}} = W \cdot \boldsymbol{y}$ .
  - It knows both $p_{\boldsymbol{x},\boldsymbol{y}}$ and specific observed $\boldsymbol{y} = \boldsymbol{v}$.
- The error simplifies to:

$$\boldsymbol{e} = \boldsymbol{x} - \sum_{n=1}^{N} \boldsymbol{w}_n \cdot \boldsymbol{y}_n = \boldsymbol{x} - W \cdot \boldsymbol{y}.$$

- The linear MMSE is
  - $\mathbb{E}_{\,\boldsymbol{x},\boldsymbol{y}}[\|\boldsymbol{e}\|^2] = \mathbb{E}\,[\|\boldsymbol{x} - W \cdot \boldsymbol{y}\|^2]$ .

**Orthogonality Principle**

- Its minimum occurs when $\mathbb{E}[\boldsymbol{e} \cdot \boldsymbol{y}_n^*] = \boldsymbol{0}$ for all $n$:
  - Proof is in Appendix D.2.
  - That is, the error and the estimator's input are uncorrelated.




- Minimum $\widehat{\boldsymbol{x}} = \underbrace{R_{\boldsymbol{xy}} \cdot R_{\boldsymbol{yy}}^{-1}}_{W} \cdot \boldsymbol{y}$ , linear in $\boldsymbol{y}$ , **so true MMSE if Gaussian.**
  - The true MMSE $f(\boldsymbol{y})$ may not be linear if non-Gaussian.
- Always: $(\widehat{\boldsymbol{x}_1 + \boldsymbol{x}_2}) = \widehat{\boldsymbol{x}}_1 + \widehat{\boldsymbol{x}}_2$ or $\widehat{A \cdot \boldsymbol{x}} = A \cdot \widehat{\boldsymbol{x}}$ .

MMSE Matrix: $R_{\boldsymbol{ee}} = R_{\boldsymbol{xx}} - R_{\boldsymbol{xy}} \cdot R_{\boldsymbol{yy}}^{-1} \cdot R_{\boldsymbol{yx}} = R_{\boldsymbol{x}/\boldsymbol{y}}^{\perp}$

$= R_{\boldsymbol{xx}} - W \cdot R_{\boldsymbol{yx}}$

# Vector and Matrix Norms

- An autocorrelation matrix' trace is its norm (and also equal to mean-squared length of random vector).
  - MMSE $=\mathbb{E}[||\boldsymbol{e}||^2] = trace\{R_{\boldsymbol{ee}}\}$ .
- The **trace** of a square autocorrelation matrix is also equal to the **sum** of its eigenvalues $\mathcal{E}_{e',n}$ :
  - $R_{\boldsymbol{ee}} = Q^* \cdot R_{\boldsymbol{e'e'}} \cdot Q$ is a symmetric matrix' eigen decomposition with diag $\{R_{\boldsymbol{e'e'}}\}$ holding eigenvalues, columns of $Q$ = eigenvectors.

$$\boldsymbol{e}' = Q \cdot \boldsymbol{e} \qquad diagonal\colon R_{\boldsymbol{e'e'}} = Q \cdot R_{\boldsymbol{ee}} \cdot Q^* \, .$$

$$||\boldsymbol{e}||^2 = ||\boldsymbol{e}'||^2 \text{ because } QQ^* = Q^*Q = I \, .$$

- An autocorrelation matrix' **determinant** is the **product** of its eigenvalues $\mathcal{E}_{e',n}$:
- MMSE $=\mathbb{E}[||\boldsymbol{e}||^2]$ and $\ln|R_{\boldsymbol{ee}}| = \sum_n \ln \mathcal{E}_{e',n}$ .
- The minimization of each component of $\boldsymbol{e}$ is *variables separable* (has its own row of $W$),
  - so then the sum is minimized,
  - but this means each of the $\boldsymbol{e}'$ also ($W \rightarrow Q \cdot W$) is minimized,
  - so then $|R_{\boldsymbol{ee}}| = |R_{\boldsymbol{e'e'}}|$ is also minimized
  - → Minimizing sum (trace) here **is same as minimizing product (determinant). MMSE as trace or determinant.**

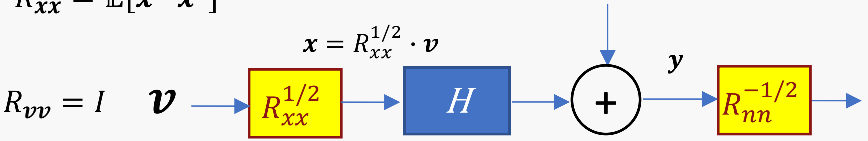
# Matrix SNR?

$\mathcal{E}_{\boldsymbol{x}} \rightarrow trace\,\{R_{\boldsymbol{xx}}\}$

$R_{\boldsymbol{xx}} = \mathbb{E}[\boldsymbol{x} \cdot \boldsymbol{x}^*]$

AWGN $\boldsymbol{n} \sim [R_{\boldsymbol{nn}} = \mathbb{E}\,[\boldsymbol{nn}^*]]$

$SNR_{out} = R_{\boldsymbol{yy}} \cdot R_{\boldsymbol{nn}}^{-1}$




$\boldsymbol{y}' = \widetilde{H} \cdot \boldsymbol{v} + \boldsymbol{n}'$

$R_{\boldsymbol{yy}} = H \cdot R_{\boldsymbol{xx}} \cdot H^* + R_{\boldsymbol{nn}}$

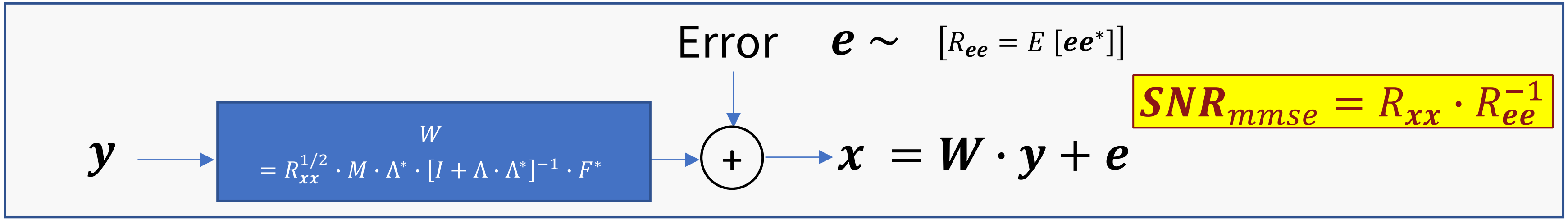
$$\widetilde{H} \triangleq R_{nn}^{-1/2} \cdot H \cdot R_{xx}^{1/2} = F \cdot \Lambda \cdot M^*$$

- It's like L2's parallel channels (take determinants) $SNR_{out} = |R_{\boldsymbol{yy}}| \cdot |R_{\boldsymbol{nn}}^{-1}| = \frac{|R_{\boldsymbol{yy}}|}{|R_{\boldsymbol{nn}}|}$ .
  - Vector code from $\boldsymbol{v}$ to $\boldsymbol{y}'$.
  - Bit rate is then $b = \log_2(SNR_{out})$.

$$SNR_{out} = \frac{|R_{\boldsymbol{yy}}|}{|R_{\boldsymbol{nn}}|} = \frac{|H \cdot R_{\boldsymbol{xx}} \cdot H^* + R_{\boldsymbol{nn}}|}{|R_{\boldsymbol{nn}}|} = |\widetilde{H} \cdot \widetilde{H}^* + I| = |\Lambda^2 + I| = \prod_{n=1}^{N}(SNR_n + 1)$$

- This $SNR_{out}$ set depends on $R_{\boldsymbol{xx}}$ choice, likely with fixed $trace\,\{R_{\boldsymbol{xx}}\}$.
  - Water-fill $R_{\boldsymbol{xx}} = M \cdot diag\{\mathcal{E}_{water-fill}\} \cdot M^*$ maximizes the matrix SNR, or effectively, its determinant.

# Backward Channel and Matrix SNR




- How about "backward channel's," $\boldsymbol{x} = \boldsymbol{W} \cdot \boldsymbol{y} + \boldsymbol{e}$ , (MMSE) $\boldsymbol{SNR}$?

$$SNR_{mmse} = \frac{|R_{xx}|}{|R_{ee}|} = |W \cdot R_{yy} \cdot W^* + R_{ee}| / |R_{ee}| = |\Lambda^2 + I| = \prod_{n=1}^{N} (1 + SNR_n)$$

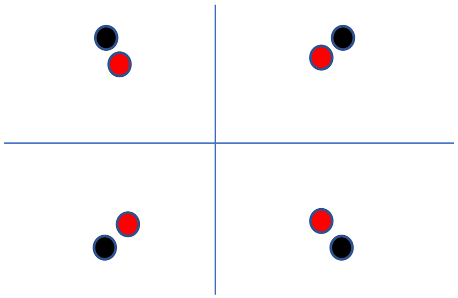
$$= SNR_{out}$$

- Bit rate is again $b = \log_2(SNR_{mmse})$, same as L3's $\mathbb{I}(\boldsymbol{x};\boldsymbol{y})$, is symmetric.
- $M^* \cdot W$ is MMSE estimate of $\boldsymbol{v}$ (linearity of MMSE estimates).
- Optimizing determinants is same as optimizing MSE, or traces.

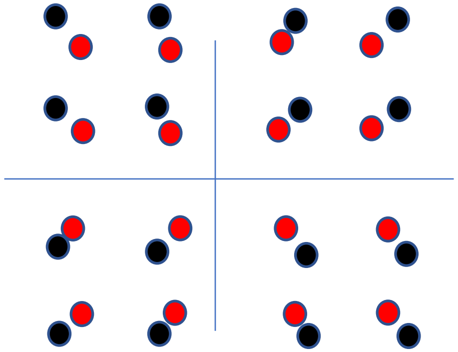
**Forward and backward have same SNR and "bit rate" (continuous $x$ distribution, or Random AEP good-code sense.)**

# MMSE is always a Biased Estimate

- Biased-Estimate Definition: $\mathbb{E}\left[{}^{\hat{x}}/_{x}\right] \neq \boldsymbol{x}$.
- MMSE estimates always have bias (if noise is nonzero), See Appendix D.2:
  - $\mathbb{E}\left[{}^{\hat{x}}/_{x}\right] = \left(I - R_{e'e'} \cdot R_{xx}^{-1}\right) \cdot \boldsymbol{x} = \left(I - \boldsymbol{SNR}^{-1}\right) \cdot \boldsymbol{x}$



decision regions same



decision regions change

MMSE trades a little signal reduction for simultaneous noise reduction when minimizing the error, now on every dimension.

See PS2.1 (Prob 4.29)

- For scalar case above, removal is scale up (by $\frac{SNR_{mmse}}{SNR_{mmse}-1}$ ).
- MIMO case, same per dimension, scale up (by $\frac{SNR_{mmse,n}}{SNR_{mmse,n}-1}$ ) **WHEN** MMSE $R_{e'e'}$ is diagonal (vector coding).
  - IF not diagonal? (See L11 later in 379B.)

# End Supplementary Lecture 3