Lecture 3

MMSE Estimation: & Information Measures

April 6, 2026

JOHN M. CIOFFI

Hitachi Professor Emeritus of Engineering

Instructor EE379B – Spring 2026

Announcements & Agenda

- Announcements
  - Problem Set 1 due Wednesday at 17:00
  - Most relevant reading – 1.5, 2.3, D.1, D.2, 4.1
  - Problem Set 2 due April 15 at 17:00

See Supplementary Lecture 3 on MMSE

- Problem Set 2 = PS2 due 4/19 at 17:00
  1. 4.29 biases and error probability
  2. 4.36 MMSE spatial equalizer
  3. 2.10 Entropy and Dimensionality
  4. 2.14 Bandwidth vs Power
  5. 2.20 MMSE and Entropy

- Agenda
  - Linear Matrix MMSE Examples
  - Information Measures – generalizing Gaussian to all distributions
  - Mutual Information and MMSE
  - Chain Rule

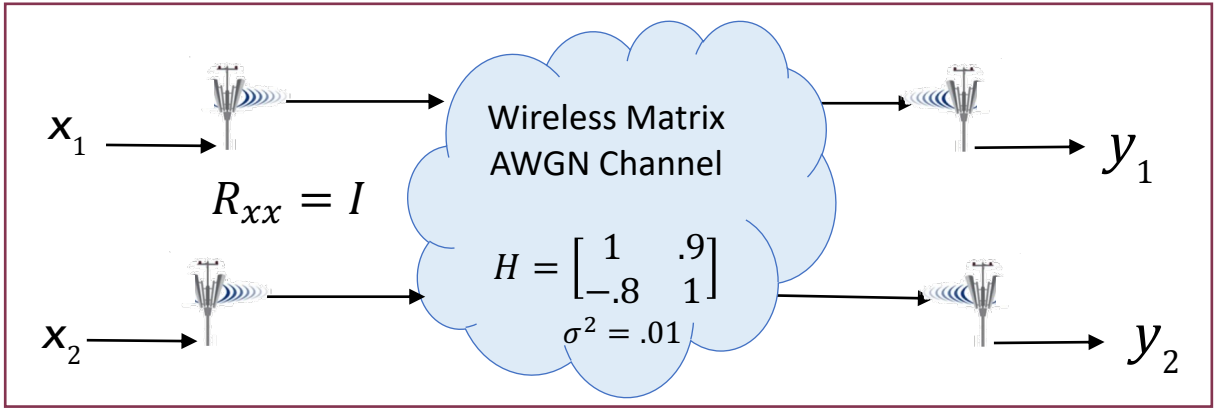
Linear Matrix MMSE Examples

See PS2.2 (Prob 4.36)

Linear MMSE: 2 x 2 Antenna System

See S3 on MMSE

$$\hat{x} = \underbrace{R_{xy} \cdot R_{yy}^{-1}}_{W} \cdot y$$




```
>> H=[1 .9
-.8 1];
>> Rxx=eye(2);
>> Rnn=.01*eye(2)
>> Ryy=H*Rxx*H'+Rnn;
>> Ryx=H;
>> W=(Ryx')*inv(Ryy) =
   0.5780   -0.5199
   0.4627    0.5780
>> W*H =
   0.9939    0.0003
   0.0003    0.9945
>> Ree=Rxx-W*Ryx =
   0.0061   -0.0003
  -0.0003    0.0055
>> snr=det(Rxx)/det(Ree);
>> log2(snr) =   14.8693
```

- This is Chapter 1's strong-crosstalk case.

  MMSE obtains the same L1:31 result, even without the "M" discrete modulator but why with no transmit M matrix?

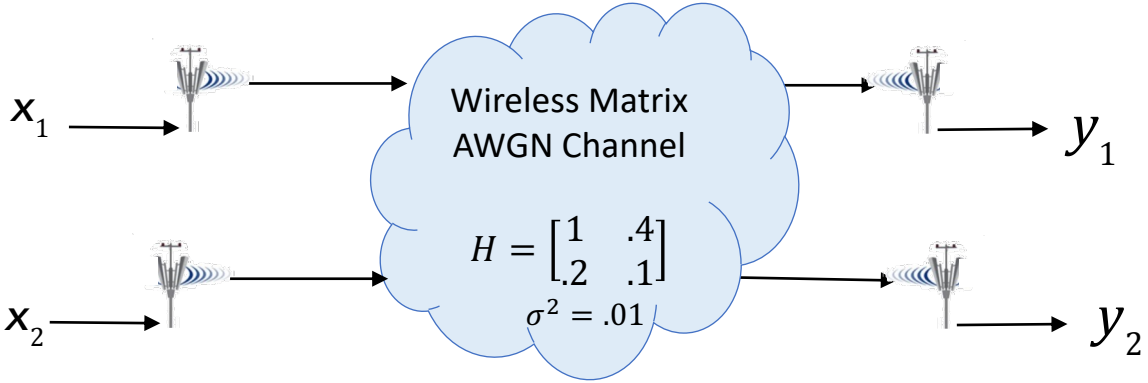
```
>> Mstar =

   0.4197    0.9076
   0.9076   -0.4197
```

$R_{xx} = I$ is close to water-fill (equal energy this channel);
$R_{xx} = M \cdot I \cdot M^*$; so "lucky" that its already close to best.

ML detector is only per-dimension independent if R_{xx} and R_{ee} are diagonal.

2 x 2 Antenna System



- This channel water-filled nonzero energy only on 1 dimension in L1.

```
>>H = 1.0000   0.4000
       0.2000   0.1000
>> Rxx=eye(2);
>> Ryy=H*Rxx*H'+Rnn;
>> Ryx=H;
>> W=(Ryx')*inv(Ryy) =
   0.9524  -0.4762
   0.0000   1.6667
>> W*H = 0.8571   0.3333
         0.3333   0.1667
>> Ree=Rxx-W*Ryx =
   0.1429  -0.3333
  -0.3333   0.8333
>> snr=det(Rxx)/det(Ree) = 126.0000
>> b=log2(snr) =  6.9773  (only for VC)
```

not diagonal; ML detect is NOT parallel

Previously in L1, $b = \log_2(1 + 2 \cdot g_2) = 6.93$ bits/subsymbol

But this time, two dimensions are used, and the ML detectors are interdependent

```
>> SNR=inv(diag(diag(Ree))) =  7.0000       0
                                    0   1.2000

>> log2(diag(SNR)) =
   2.8074
   0.2630
>> sum(log2(diag(SNR)))  =   3.0704
```

Thus, data rate loss can occur with independent detectors and MMSE.

(This loss can be recovered with Chapter 5's MMSE GDFE, or by using vector coding, so ∃ good solutions, not linear MMSE)

See PS2.2 (Prob 4.36)

Time-Frequency Block $1 + .9 \cdot D^{-1}$

```
>> H=toeplitz([1 zeros(1,7)]',[1 .9 zeros(1,7)]);
>> Rxx=eye(9);
>> Rnn=.181*eye(8);
>> Ryy=H*Rxx*H'+Rnn;
>> Ryx=H;
>> W=(Ryx')*inv(Ryy);
>> P=W*H;
>> size(P) % =   9   9

>> Ree=Rxx-W*Ryx;
>> snr=det(Rxx)/det(Ree) = 2.4089e+07
>> SNR=inv(diag(diag(Ree)));
>> bn = 0.5*log2(diag(SNR))' =
 0.8769   1.0096   1.0691   1.0907   1.0902   1.0673   1.0054   0.8681   0.6085
>> sum(bn/9) =   0.9651
>> 10*log10(2^(2*ans)-1) =  4.4885 dB
```

Repeat for 8→32:

>> sum(bn/33) = 1.0753

>> 10*log10(2^(2*ans)-1) = 5.3654 dB

Best infinite length is 5.7 dB.
(with dimension-by-dimension linear)

- See Chapter 3, 379A MMSE-LE example
- 379A-L14:17

Best with full ML is 8.8 dB, but requires input WF energy distribution, and vector Coding, or Chapter 5's GDFE.

Entropy and Estimation:
generalizing energy to all distributions

See PS2.3 (Prob 2.10)

Rest of L3 repeats EE379A early coding,
but here emphasizes the MMSE-canonical connection.

Information Measures and Energy/Mean-Square

Gaussian Distribution	Any Distribution
Mean-square energy $\mathcal{E}_x = \mathbb{E}[\|\boldsymbol{x}\|^2]$	Differential Entropy $\mathscr{H}_{\boldsymbol{x}}$ *Section 2.3.1*
Mean-square error $\sigma_e^2 = \mathbb{E}[\|\boldsymbol{e}\|^2]$	Conditional Entropy $\mathscr{H}_{\boldsymbol{x}/\boldsymbol{y}}$ *Section 2.3.2*
Signal-to-Noise $\mathcal{E}_x$ / σ_e^2	Mutual Information $\mathcal{I}(\boldsymbol{x};\boldsymbol{y})$ = $\mathscr{H}_{\boldsymbol{x}}$ - $\mathscr{H}_{\boldsymbol{x}/\boldsymbol{y}}$ *Section 2.3.2*

The information-carried by random variable/process generalizes the energy concepts from MMSE/Gaussian analysis to a general distribution.

These information measures correspond to bits/symbol quantities, and for the Gaussian case are basically the $\log_2$ of the corresponding energy measure.

It even works for discrete distributions (entropy, not differential entropy)

Entropy – measure information (source)

- **Entropy:** $\mathcal{H}_{\tilde{x}} = \mathbb{E}\left[\log_2\left(\frac{1}{p_{\tilde{x}}}\right)\right] = \sum_{i=0}^{|C|-1} p_{\tilde{x}}(i) \cdot \log_2\left(\frac{1}{p_{\tilde{x}}(i)}\right)$ Discrete $p_{\tilde{x}}(i)$

- Measures a discrete distribution's many values, its ***information***, by probability (think subsymbols).

- Generalizes bits/subsymbol, especially when the constellation size $|C| \geq M^{1/\bar{N}} = 2^{\tilde{b}}$.

example: $p_{\tilde{x}}(i) = \frac{1}{M}$ (uniform) → $|C| = 2^{\tilde{b}}$

Uniform → $\mathcal{H}_{\tilde{x}} = \log_2\left(M^{1/\bar{N}}\right) = \tilde{b}$ $(|C| = 2^{\tilde{b}+\tilde{\rho}})$ $\tilde{\rho} = 0$; uncoded)

- **Uniform distribution** has **maximum entropy**

$\mathcal{H}_{\tilde{x}} \leq \log_2|C|$ Binary example: $p_{\tilde{x}}(0) = \frac{1}{128}$ and $p_{\tilde{x}}(1) = \frac{127}{128}$

$$\mathcal{H}_{\tilde{x}} = \frac{\log_2(128)}{128} + \frac{127}{128} \cdot \log_2\left(\frac{128}{127}\right) = .06 < 1$$

See PS2.3 (Prob 2.10)

Information left after given another random vector

- Conditional entropy $\mathcal{H}_{\tilde{x}/\tilde{y}}=\mathbb{E}\left[\log_2\left(\frac{1}{p_{\tilde{x}/\tilde{y}}}\right)\right]=\sum_{j=0}^{|Y|-1}\sum_{i=0}^{|C|-1}p_{\tilde{x}\tilde{y}}(i,j)\cdot\log_2\left(\frac{1}{p_{\tilde{x}/\tilde{y}}(i,j)}\right)$

$$\mathcal{H}_{\tilde{x}/\tilde{y}}=\mathcal{H}_{\tilde{x}\tilde{y}}-\mathcal{H}_{\tilde{y}}$$

- Measures $\tilde{x}$'s residual randomness/info when $\tilde{y}$ is known/given

$\tilde{x}$; $\tilde{y}$	0	1	$p_{\tilde{x}}$
0	3/8	1/8	1/2
1	1/8	3/8	1/2
$p_{\tilde{y}}$	1/2	1/2	

$$\mathcal{H}_{\tilde{x}\tilde{y}}=\frac{6}{8}\cdot\log_2\frac{8}{3}+\frac{2}{8}\cdot\log_2 8=1.811$$

$$\mathcal{H}_{\tilde{x}}=1=\mathcal{H}_{\tilde{y}}$$

$$\mathcal{H}_{\tilde{x}/\tilde{y}}=1.811-1=.811\ bits/subsymbol$$

- If $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, then $\mathcal{H}_{\tilde{x}/\tilde{y}}=\mathcal{H}_{\tilde{x}}$

Continuous Distribution – DIFFERENTIAL Entropy

- Differential Entropy $\mathscr{H}_{\tilde{x}} = \mathbb{E}\left[\log_2\left(\frac{1}{p_{\tilde{x}}}\right)\right] = -\int_{-\infty}^{\infty} p_{\tilde{x}}(u) \cdot \log_2\left(\frac{1}{p_{\tilde{x}}(u)}\right) \cdot du$

- Differential Entropy $\mathscr{H}_{\tilde{x}}$ is not same as an integral-to-sum via a discrete approximation of $p_{\tilde{x}}(u)$.
  - They differ by a constant that depends on the approximation-interval size.
- Differential Entropy $\mathscr{H}_{\tilde{x}}$ does still however measure information content when subsymbols in codewords are chosen (usually at random) from $p_{\tilde{x}}(u)$.
- Maximum $\mathscr{H}_{\tilde{x}}$ occurs when $p_{\tilde{x}}(u)$ is **Gaussian** (any mean), with constant average energy.

$$\int_{-\infty}^{\infty} p_{\tilde{x}}(u) \cdot \|u\|^2 \cdot du = \mathcal{E}_{\tilde{x}}$$

Complex $\mathscr{H}_{\tilde{x}} = \log_2(\pi e \cdot \mathcal{E}_{\tilde{x}})$ bits/clpx-subsymbol

Real $\mathscr{H}_{x} = \frac{1}{2}\log_2(2\pi e \cdot \bar{\mathcal{E}}_{x})$ bits/dimension

- More generally, $\text{trace}\{R_{\tilde{x}\tilde{x}}\} = \mathcal{E}_{\tilde{x}}$.

$\mathscr{H}_{\tilde{x}} = \log_2|\pi e \cdot R_{\tilde{x}\tilde{x}}|$ bits/cplx-subsymbol

Gaussian = Uniform?

- First – qualification: this is in random-coding sense
  - Pick $2^{\bar{b}\cdot N}$ length-N codewords by sampling subsymbols from a stationary random variable with continuous dist.
- Yes, these codewords will have a uniform distribution in $N \to \infty$ dimensions (codewords ~ $2^{\overline{-b}\cdot N}$)
  - Upcoming AEP, asymptotic equipartition.
- The (often tacitly implied) constraint is that the code must have average energy $\mathcal{E}_{\boldsymbol{x}}$.
- These codewords thus are uniform over a hypersphere (all on surface actually too with prob 1!).
- The marginals in any finite number of dimensions of this hyper-spherical uniform are GAUSSIAN!

So YES, a Gaussian is in a way uniform, over a hypersphere's surface, differential dimensionalities, $\mathcal{H}_{\tilde{x}}$.

Not true for discrete, Where uniform is uniform, $\mathcal{H}_{\tilde{x}}$.

Gaussian MMSE & conditional entropy

- Complex scalar Gaussian x $\quad p_x(u) = \frac{1}{\pi\sigma_x^2} e^{-\frac{|x|^2}{\sigma_x^2}} \qquad \mathcal{H}_x = \log_2\{\pi \cdot e \cdot \sigma_x^2\}$

- Conditional x/y ? $\quad \mathcal{H}_{x/y} = \log_2\{\pi \cdot e \cdot \sigma_{x/y}^2\}$ $\qquad \sigma_{x/y}^2 = \sigma_x^2 - {r_{xy}^2}/{\sigma_y^2} = \text{MMSE}$

- Vector $\boldsymbol{x}$? $\quad \mathcal{H}_{\boldsymbol{x}} = \log_2\{(\pi e)^{\bar{N}} \cdot |R_{\boldsymbol{xx}}|\}$

$$\mathcal{H}_{\boldsymbol{x/y}} = \log_2\{(\pi e)^{\bar{N}} \cdot |R_{\boldsymbol{x/y}}^{\perp}|\}$$

$$R_{\boldsymbol{x/y}}^{\perp} = R_{\boldsymbol{xx}}^2 - R_{\boldsymbol{x/y}} \cdot R_{\boldsymbol{yy}}^{-1} \cdot R_{\boldsymbol{x/y}} = \text{MMSE}$$

(Appendix D on MMSE)

$\bar{N}$ is the number of complex dimensions = $N/2$

Relation to MMSE Estimation

- If $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ are jointly Gaussian, then $p_{\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}}$ is also Gaussian and has mean as MMSE estimate $\mathbb{E}[\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}]$ and autocorrelation $R_{\boldsymbol{ee}} = R_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}} - R_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}} \cdot R_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}}^{-1} \cdot R_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{x}}}$.

- $\mathscr{H}_{\tilde{x}/\tilde{y}} = \log_2|\pi e R_{\boldsymbol{ee}}|$ - that is, the entropy is essentially just the log of the MMSE (Gaussian).

  - Differential Entropy generalizes MMSE to any continuous probability distribution.
  - Measures the information content of the "miss" in estimating $\tilde{\boldsymbol{x}}$ from $\tilde{\boldsymbol{y}}$ for any $p_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}$.
  - Entropy does this for discrete distributions.

Mutual Information and SNR

Subsection 2.3.2

See PS2.5 (Prob 2.20)

For Gaussian, $\mathcal{I}$ and (geo) SNR are in 1-to-1 relationship: MMSE and best rate are essentially same thing.

Mutual Information ~ SNR

- Mutual Information is:

$$\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})=\mathbb{E}\left[\log_2\left(\frac{p_{\tilde{x}\tilde{y}}}{p_{\tilde{x}}\cdot p_{\tilde{y}}}\right)\right]=\mathcal{H}_{\tilde{x}}-\mathcal{H}_{\tilde{x}/\tilde{y}}=\mathcal{H}_{\tilde{y}}-\mathcal{H}_{\tilde{y}/\tilde{x}}.$$

$$=\mathcal{H}_{\tilde{x}}-\mathcal{H}_{\tilde{x}/\tilde{y}}=\mathcal{H}_{\tilde{y}}-\mathcal{H}_{\tilde{y}/\tilde{x}}\,.$$

  - For discrete example $\mathcal{I}$ = 1-.811 = .189 bits/subsymbol.

- $\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})$ is symmetric in $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ (MMSE forward and backward channel).

- $\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})$ measures common ("mutual") information between $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$, $\mathbb{E}\left[\log_2\left(\frac{p_{\tilde{x}/\tilde{y}}}{p_{\tilde{y}}}\right)\right]$.
  - On average, $\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})$ measures how much bigger is unconditional info versus conditional info, in bits.

- $\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})=\log_2\frac{|R_{\tilde{x}\tilde{x}}|}{|R_{\boldsymbol{ee}}|}=\log_2\frac{|R_{\tilde{y}\tilde{y}}|}{|R_{\boldsymbol{nn}}|}=\log_2\left(\left(1+SNR_{\boldsymbol{geo}}\right)^{\bar{N}}\right)$ for the matrix AWGN.

- OR as earlier for vector coding $\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})=\sum_{n=1}^{\bar{N}}\log_2 SNR_{mmse,n}$ for the matrix AWGN.

Law of Large Numbers, repeat 379A

Theorem 2.1.1 (Law of Large Numbers (LLN)) *The LLN observes that a stationary random variable z's sample average over its observations $\{z_n\}_{n=1,\ldots,N}$ converges to its mean with large N such that*

$$\lim_{N\to\infty} \quad Pr\left\{\left|\left(\frac{1}{N}\sum_{n=1}^{N} z_n\right) - \mathbb{E}[z]\right| > \epsilon\right\} \to 0 \quad \textit{weak form} \tag{2.13}$$

$$\lim_{N\to\infty} \quad Pr\left\{\frac{1}{N}\sum_{n=1}^{N} z_n = \mathbb{E}[z]\right\} = 1 \quad \textit{strong form} \ . \tag{2.14}$$

- Distribution of z must be the same (stationary) for all random selections.
- The random z can be function of random variable ($z = f(x)$)and the sample mean converges to $\mathbb{E}[f(x)]$.
  - E.g., $z_n = \|\boldsymbol{x}_n\|^2$ where the vector $\boldsymbol{x}_n$ might also have (a growing) N components (energy sample or length of the vector).
  - LLN then states that all the energy (really points in selection from any distribution with$\mathbb{E}[\|\boldsymbol{x}\|^2] \leq \mathcal{E}_x$) of a hypersphere are are at its surface with probability 1. Points on the interior have probability zero. It is also a sum of independent terms, and thus Gaussian (central limit theorem).
  - The marginal distributions for the vector $\boldsymbol{x}_n$'s element selections, and thus for $\boldsymbol{x}_n$ also, would be Gaussian if this $N\to\infty$-sequence has max entropy (uniform).
- The function of most interest in coding is $f(x) = -\log_2[p_x(x)]$- that is the function itself is probability distribution's log.
  - The **sample average** of this function **converges** to the **entropy**.
  - This suggests choosing codewords (this means each subsymbol in the codeword) at random from stationary distribution,
    - and then repeat at higher level for several codes chosen at random.
  - These are discrete codes, even when $\boldsymbol{x}$ is continuous, but their average follows the differential entropy (and mutual information).
  - Generalizes **sphere-packing** (which was for the AWGN only).

Random coding generalizes 379A sphere packing

- Pick subsymbols $\boldsymbol{x}_n$ randomly (independently) from (stationary) distribution $p_{\tilde{\boldsymbol{x}}}$ for each of $M = 2^b$ c'words.
  - This is one **random code**.
- Repeat the exercise for another code, and …. many more.
- Compute the average performance of all these random selected codes:
  - As $\overline{N} \to \infty$, this average performance is outstanding (as we'll see), as long as $\tilde{b} < \mathbb{I}(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{y}})$.
  - So at least one good one must exist.
- Entropy per subsymbol is

$$\begin{aligned} \widetilde{\mathcal{H}}_{\boldsymbol{x}} &= \frac{-1}{\overline{N}} \cdot E\left[\log_2(p_{\boldsymbol{x}})\right] \\ &= \frac{-1}{\overline{N}} \sum_{n=1}^{\overline{N}} E\left[\log_2(p_{\tilde{\boldsymbol{x}}_n})\right] \quad , \end{aligned}$$

- LLN with function $\log_2\left[p_{\tilde{z}}^{-1}\right]$ is the sample-average entropy estimate.

$$\hat{\widetilde{\mathcal{H}}}_{\tilde{\boldsymbol{x}}} = \frac{-1}{\overline{N}} \cdot \sum_{n=1}^{\overline{N}} \log_2\left[p(\tilde{\boldsymbol{x}}_n)\right] = \frac{-1}{\overline{N}} \cdot \log_2\left[p(\boldsymbol{x})\right] \quad \cdot \text{ , which converges to (constant) } \widetilde{\mathcal{H}}_{\boldsymbol{x}}$$

- The constant means the ave code has uniform distribution of codewords (asymptotically), $2^{\overline{N} \cdot \widetilde{\mathcal{H}}_{\boldsymbol{x}}}$ of them.

Asymptotic Equal Partition (AEP)

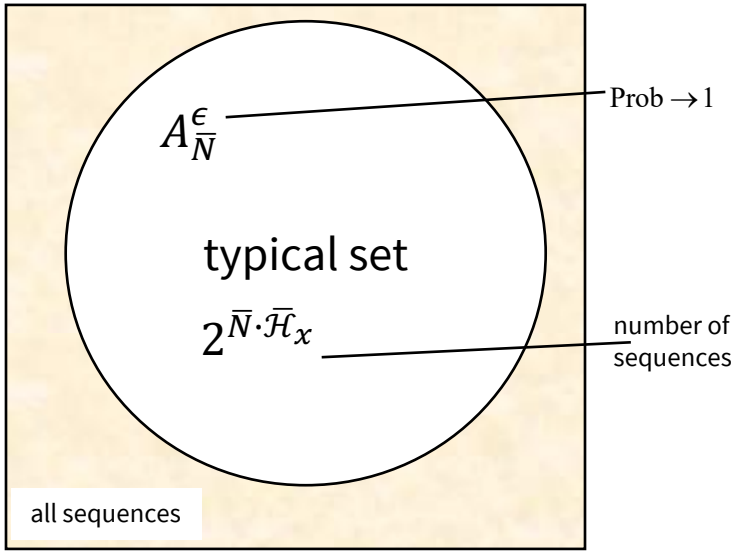
AEP Typical Sets ~ spheres/ball packing

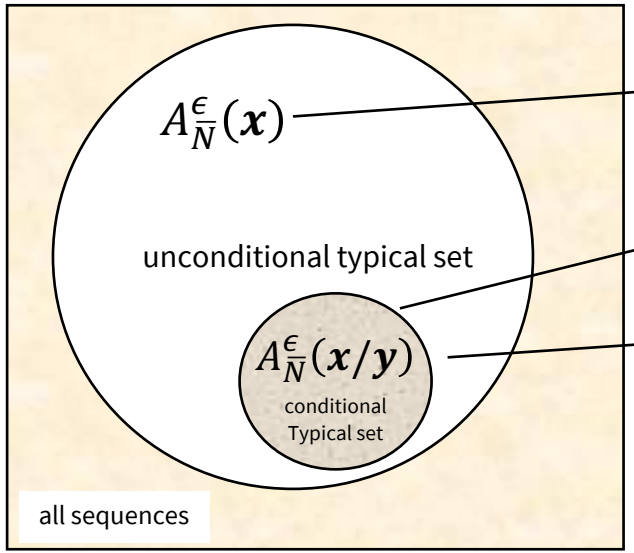
- The set is $A_{\overline{N}}^{\epsilon}(\boldsymbol{x}) \triangleq \left\{ \boldsymbol{x} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_{\overline{N}}] \;\middle|\; 2^{-\overline{N}\cdot\mathcal{H}_{\tilde{\boldsymbol{x}}}-\epsilon} \le p(\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_{\overline{N}}) \le 2^{-\overline{N}\cdot\mathcal{H}_{\tilde{\boldsymbol{x}}}+\epsilon} \right\}$

Lemma 2.3.6 **[AEP Lemma]** *For a typical set with $\overline{N} \to \infty$, the following are true:*

- $Pr\{A_{\overline{N}}^{\epsilon}(\boldsymbol{x})\} \to 1$
- *for any codeword* $\boldsymbol{x} \in A_{\overline{N}}^{\epsilon}$, $Pr\{\boldsymbol{x}\} \to 2^{-\overline{N}\cdot\mathcal{H}_{\tilde{\boldsymbol{x}}}}$

Decoder works well if only one codeword in conditional set for each $\boldsymbol{y}$ value, so good code spreads them uniformly.





$\mathcal{H}_{\tilde{\boldsymbol{x}}} = \mathbb{I}(\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{y}}) + \mathcal{H}_{\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}}$ bits/subsymbol

$\le \tilde{b} + \tilde{\rho}$

So pick $\tilde{\rho}$ large enough.

There are $2^{N\cdot H_{\tilde{\boldsymbol{x}}}} \cdot 2^{-N\cdot H_{\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}}} = 2^{N\cdot\mathbb{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})}$ little sets In the big set if "equally partitioned"

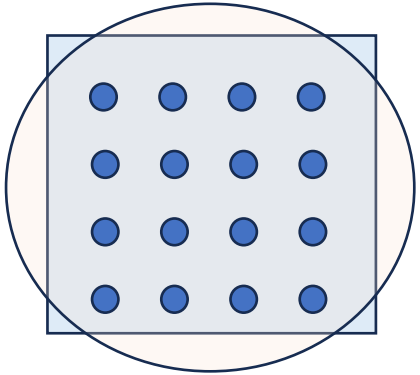
General Capacity Theorem from L3

$$\frac{|A_N^\epsilon(\boldsymbol{x})|}{|A_N^\epsilon(\boldsymbol{x}/\boldsymbol{y})|} \to 2^{\mathcal{I}(\boldsymbol{x};\boldsymbol{y})} \qquad \text{since } \mathcal{I}(\boldsymbol{x};\boldsymbol{y}) = \mathcal{H}_{\boldsymbol{x}} - \mathcal{H}_{\boldsymbol{x}/\boldsymbol{y}}$$

- Good codes will have only 1 codeword per conditional entropy subset.
- MAP detector decision region is then ~ $A_N^\epsilon(\boldsymbol{x}/\boldsymbol{y})$ - on average; but we can find it for one good code.
- If $A_N^\epsilon(\boldsymbol{x})$ were any larger, all codes (good or bad) will have at least one $A_N^\epsilon(\boldsymbol{x}/\boldsymbol{y})$ that contains 2+ codewords, which mean the MAP has to "flip a coin" – not good (high error prob).
- SHANNON's CAPACITY THEOREM
  - Number of codewords is limited by mutual info $b \leq \mathbb{I}(\boldsymbol{x}\,;\boldsymbol{y}\,)$.
  - Which is per-subsymbol equivalent with random code $\tilde{b} \leq \mathbb{I}(\tilde{\boldsymbol{x}}\,;\tilde{\boldsymbol{y}}\,)$.
  - If maximized over input distributions $\tilde{b} < \tilde{C} \leq \max_{p_{\tilde{\boldsymbol{x}}}} \mathbb{I}(\tilde{\boldsymbol{x}}\,;\tilde{\boldsymbol{y}}\,) \frac{\text{bits}}{\text{subsymbol}}$.

The uniform part is most important (from L3)

- The Gaussian distribution corresponds to marginal of uniform distribution over a hypersphere.
  - This uniform distributions marginals are asymptotically Gaussian.
  - This is a special case where uniform and Gaussian are basically the same.
  - Because all the Gaussian infinite-length vectors (codewords) have same energy (zero variance of the energy).
- All the points (really volume) are (is) at the surface.
- The Gaussian marginal dist'n is important only for shaping gain (< 1.53 dB).
- The (AEP) uniform spacing of points (no matter where the majority of them sit, surface or otherwise) remains for the fundamental gain.



The uniform spacing separates codewords in the union of the hypersquare (orthotope) and hypersphere.

Thus, good codes can be based on sequences from uniformly spaced PAM/QAM subsymbols.

And the rest is MMSE estimation,

With a chain-rule twist in some situations

Vector Coding is always MAP, ML, & MMSE, a special case.

Good Codes

- There are many of them for any given channel.
- They will approximate uniform distribution over some finite-dimensional space
  - Like a hypersphere if ave energy is the constraint.
- The uniform part is important, not just in probability, but in spacing through the N-dimensional codeword space.
  - Hyperspheres are often called “balls” in general spaces over finite fields
  - QPSK is a great code if $N \leq 2$, but there are better codes with larger N.
- **USE GOOD CODES** – they make everything you’ll see in this class later work better!

Chain Rule

Subsection 2.3.2

Chain Rule

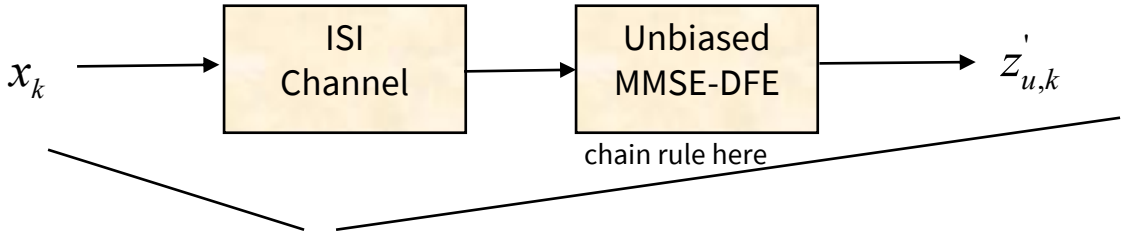
$$\mathsf{I}(\boldsymbol{x};\boldsymbol{y}) = \sum_{n=1}^{N} \mathsf{I}(\tilde{\boldsymbol{x}}_n;\boldsymbol{y}\,/[\tilde{\boldsymbol{x}}_{n-1} \quad \cdots \quad \tilde{\boldsymbol{x}}_1])$$

- If parallel independent channels (we know this by now!), the $[\tilde{\boldsymbol{x}}_{n-1} \quad \cdots \quad \tilde{\boldsymbol{x}}_1]$ provide no help;
  - just sum the individual channels $\mathsf{I}(\tilde{\boldsymbol{x}}_n;\tilde{\boldsymbol{y}}_n)$.
- But suppose not: each term represents a code (MMSE-related if Gaussian) problem with SNR, capacity, etc.

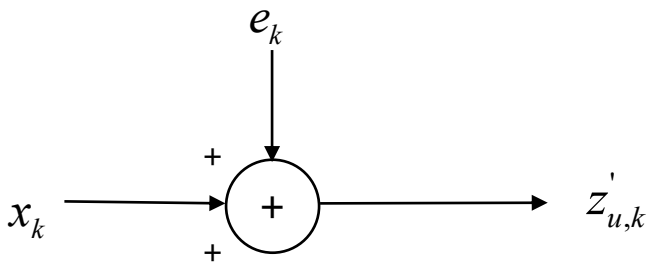
Matrix AWGN: GDFE (sometimes also called "successive decoding")

- Estimate (MMSE) and decode $[\tilde{\boldsymbol{x}}_{n-1} \quad \cdots \quad \tilde{\boldsymbol{x}}_1]$ first, then simpler single component problem.
  - So not just linear MMSE, linear MMSE + subtract "earlier" subsymbols' effect (nonlinear).
- It's parallel channels, but with a twist to make them independent step by step ("decision-feedback").

CDEF Example EE379A, L18



equivalent to



$$\mathcal{I}(\tilde{x};\tilde{\boldsymbol{y}}) = \mathcal{H}_{\tilde{x}_k} - \underbrace{\mathcal{H}_{\tilde{x}_k/[\tilde{\boldsymbol{y}} \quad \tilde{x}_{k-1.....-\infty}]}}_{MMSE-DFE}$$

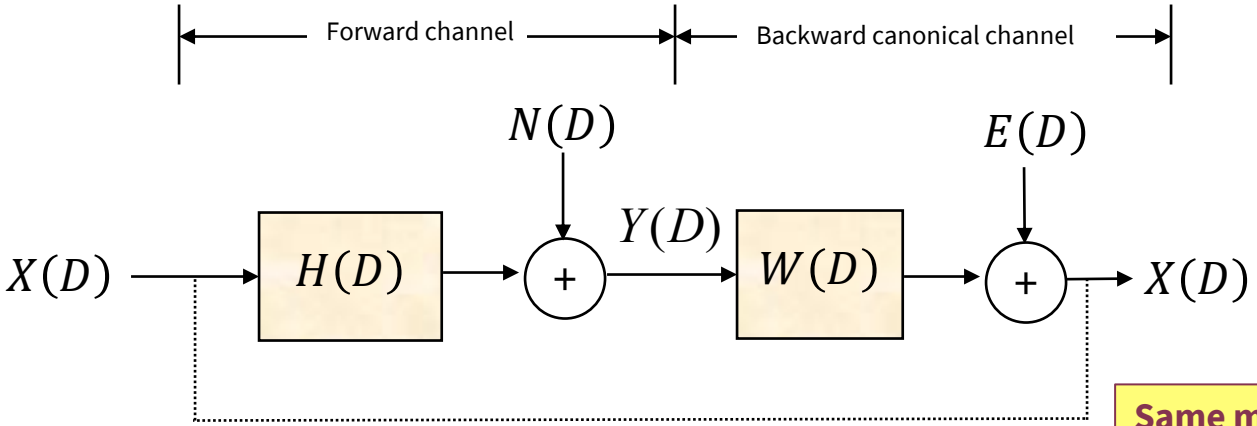
$$SNR = SNR_{mmse-dfe,u} = 2^{2\bar{\mathcal{I}}(\tilde{x};\tilde{y})} - 1$$

- The MMSE-DFE achieves the highest rate (with $\Gamma = 0$ dB) also:
  - $I = C$ if water-filling spectrum is at transmitter.
  - But, this spectra may be impossible with a single DFE, so can be several parallel DFES (see Section 3.12).
  - No error propagation (true if P_e is zero), and **canonical** (reliably achieves capacity).

There are many chain-rule orders.

- $\mathbb{I}(\boldsymbol{x};\boldsymbol{y}) = \sum_{n=1}^{N} \mathbb{I}\left(\tilde{\boldsymbol{x}}_{\pi(n)};\boldsymbol{y}\,/[\tilde{\boldsymbol{x}}_{\pi(n-1)} \quad \cdots \quad \tilde{\boldsymbol{x}}_{\pi(1)}]\right) = \sum_{n=1}^{N} \mathbb{I}\left(\tilde{\boldsymbol{y}}_{\pi(n)};\boldsymbol{x}\,/[\tilde{\boldsymbol{y}}_{\pi(n-1)} \quad \cdots \quad \tilde{\boldsymbol{y}}_{\pi(1)}]\right)$
- $N!$ orders exist for each of $\mathbb{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{I}(\boldsymbol{y};\boldsymbol{x})$.
- Every order corresponds to different set of parallel channels (some with feedback, a few without).
- But they all produce the same maximum data rate (achieved with good code that has zero gap).
- Thus, not only are there a lot of good codes – there are a lot of good MMSE-based modulation/demodulation designs also!

Forward and its Backward Canonical Models



Same mutual information

$$r(t) = h_c(t) * h_c^*(-t) = \|h\|^2 \cdot q(t)$$

$y(t)$ → $h_c^*(-t)$ → $\frac{1}{T}$ → $Y(D)$

$$Y(D) = R(D) \cdot X(D) + \underbrace{N(D)}_{\frac{\mathcal{N}_0}{2} \cdot R(D)}$$

Forward Canonical Model

$$X(D) = \underbrace{W(D)}_{MMSE-LE} \cdot Y(D) + \underbrace{E(D)}_{\frac{\mathcal{N}_0}{2} \cdot W(D)}$$

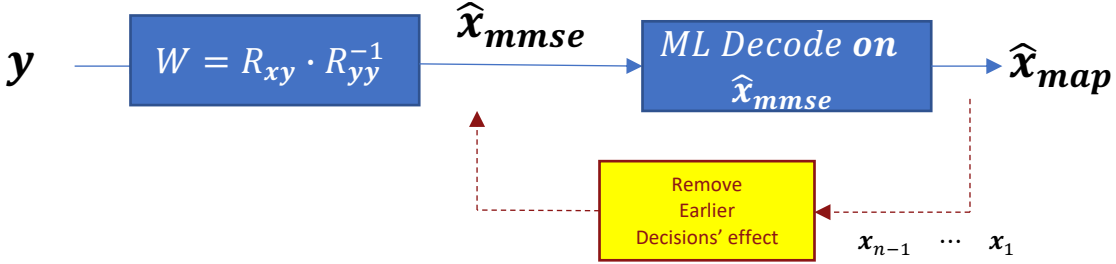
Backward Canonical Model
chain rule helps more here

For the filtered/matrix AWGN

- The MAP and MMSE determine the performance, and also the chain rule suggests a simpler decoder:

$$
\begin{aligned}
\mathcal{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}}) &= \mathscr{H}_{\tilde{\boldsymbol{y}}} - \mathscr{H}_{\tilde{\boldsymbol{y}}/\tilde{\boldsymbol{x}}} \\
&= \log_2\left(\frac{\left|R_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}}\right|}{\left|R_{\tilde{\boldsymbol{n}}\tilde{\boldsymbol{n}}}\right|}\right) \text{ bits/subsymbol} \\
&= \mathscr{H}_{\tilde{\boldsymbol{x}}} - \mathscr{H}_{\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}} \\
&= \log_2\left(\frac{\left|R_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}}\right|}{\left|R_{\boldsymbol{ee}}\right|}\right) \text{ bits/subsymbol} \\
&= \log_2 |I - W \cdot H| \quad \text{Forward} \\
&= \log_2 |I - H \cdot W| \quad \text{Backward}
\end{aligned}
$$

$$= \log_2(SNR_{mmse})$$



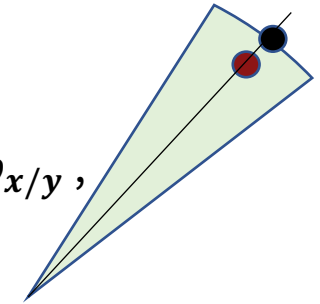
Still MAP if "previous" decisions are correct sequentially decodes

Optimal Detectors for Good Codes

- $\boldsymbol{x}$ codewords/subsymbols selected from Gaussian, $[\boldsymbol{x}\ \ \boldsymbol{y}]$ are jointly Gaussian (as is then $\boldsymbol{n}$ **).**
- **ML = MAP** since all good code's $\boldsymbol{x}$ codewords/symbols are equally likely (uniform, AEP):

$$\frac{MAP}{ML} \ni \min_{\{\tilde{\boldsymbol{x}}_k\}} \sum_{k=-\infty}^{\infty} \|\tilde{\boldsymbol{y}}_k - H \cdot \tilde{\boldsymbol{x}}_k\|^2 \neq \sum_{k=-\infty}^{\infty} \|\tilde{\boldsymbol{n}}_k\|^2.$$

Same as $\max_{\boldsymbol{x}} p_{\boldsymbol{x}/\boldsymbol{y}}$,
where $\boldsymbol{x}$ has ∞ length



- **MMSE = MAP** The smallest sum will reduce $\{\tilde{\boldsymbol{x}}_k\}$ magnitude slightly because it also shrinks noise (trade-off in sum):

$$MMSE \ni \min_{\{\tilde{\boldsymbol{x}}_k\}} \left\{ \lim_{K\to\infty} \frac{1}{2K+1} \sum_{K=-K}^{K} \|\tilde{\boldsymbol{x}}_k - W \cdot \tilde{\boldsymbol{y}}_k\|^2 \right\}$$

min over entire sum
W is MMSE filter

  - By LLN, this sum is MMSE & MAP and has optimum $\hat{\boldsymbol{x}} = E[\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}]$ …… on average over the random code set.
  - The bias removal is unnecessary because of the hyper-conical decision regions (like QPSK where decision regions don't change) for a zero-gap AWGN code, but we now know MMSE is a "DFE-like" structure (chain rule)

End Lecture 3