STANFORD

Lecture 3

MMSE Estimation: & Information Measures

April 12, 2023

JOHN M. CIOFFI

Hitachi Professor Emeritus of Engineering

Instructor EE392AA – Spring 2023

Announcements & Agenda

- Announcements
  - Problem Set 1 due today at 17:00
    - Grader volunteer? (Ex 10% / assignment plus pay)
  - Most Relevant Reading – 1.5, 2.3, D.1, D.2, 4.1
  - Problem Set 2 due April 19 at 17:00
  - **Class on April 14 , 3pm in STLC111**

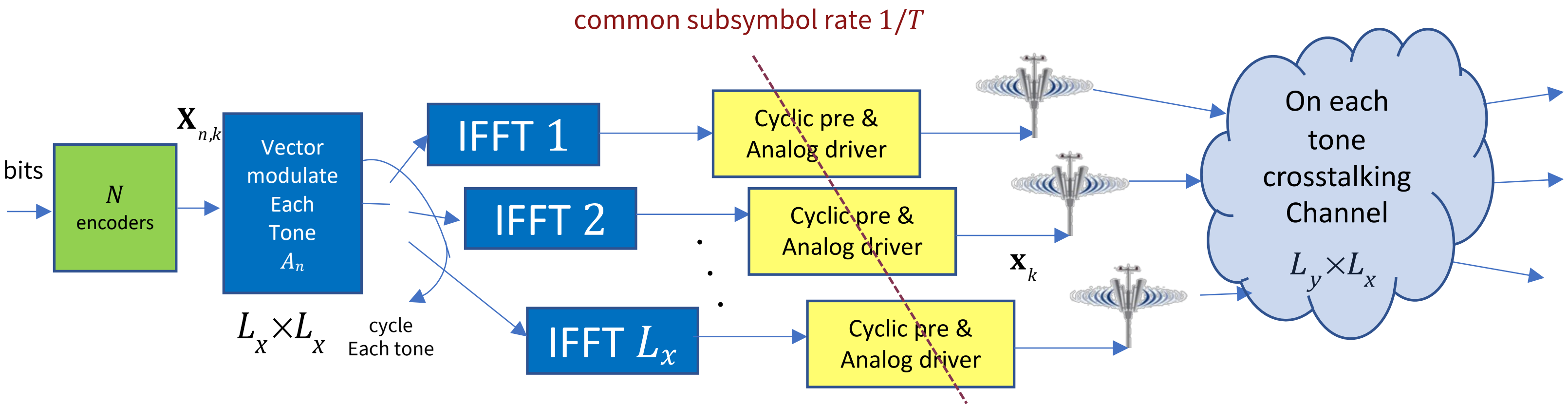
- Problem Set 2 = PS2 due 4/19 at 17:00
  1. 4.29 biases and error probability
  2. 4.36 MMSE spatial equalizer
  3. 2.10 Entropy and Dimensionality
  4. 2.14 Bandwidth vs Power
  5. 2.20 MMSE and Entropy

- Agenda
  - Vector DMT
  - General MMSE & Gaussian
    - Autocorrelation/Cross-Correlation
  - Linear MMSE & The Orthogonality Principle
    - Biases and SNRs
  - Examples
  - Information Measures – generalizing Gaussian to all distributions
  - Mutual Information and MMSE

Vector DMT

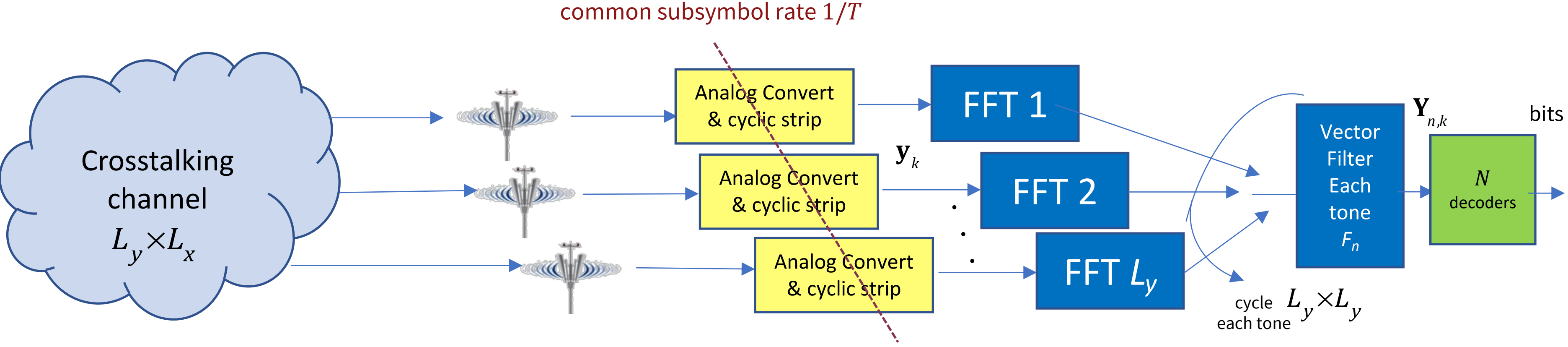
Section 4.7

Vector DMT/OFDM Transmitter



$$NL_x^2 + L_x N \log_2(N) << (L_x N)^2$$

Vector DMT/OFDM Receiver



$$\tilde{H}_n = F_n \cdot \Lambda_n \cdot M_n^*$$

$$NL_y^2 + L_y N \log_2(N) << (L_y N)^2$$

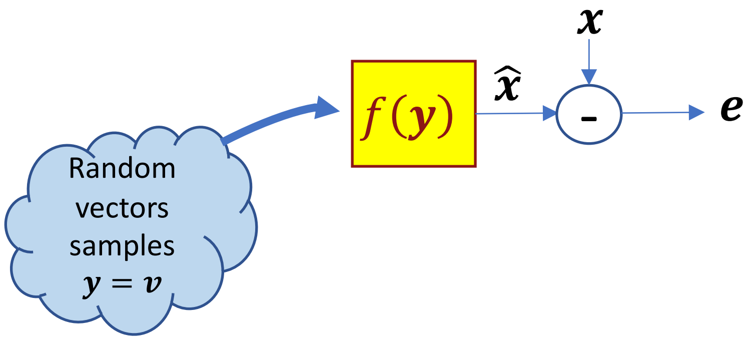
- Just much larger number of dimensions, each a scalar AWGN, $L = min(L_x, L_y)$
- $L \cdot N$ dimensions
- Can water-fill over them all (if total energy constraint, which is common)

General MMSE and Gaussian

Section D.1

The Estimation Problem

- Given random $\boldsymbol{x}$ and $\boldsymbol{y}$, want to estimate $\boldsymbol{x}$, $\widehat{\boldsymbol{x}} = f(\boldsymbol{y})$
  - Know both $p_{\boldsymbol{x},\boldsymbol{y}}$ and specific observed $\boldsymbol{y} = \boldsymbol{v}$
  - Continuous distributions $\boldsymbol{x}$ and $\boldsymbol{y}$



- The error:
  - $\boldsymbol{e} = \boldsymbol{x} - f(\boldsymbol{y})$

- Its mean-square
  - $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\|\boldsymbol{e}\|^2] = \mathbb{E}[\|\boldsymbol{x} - f(\boldsymbol{y})\|^2]$

- Its minimum

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

- Solution:
  - Conditional mean of $\boldsymbol{x}$ given $\boldsymbol{y}$; $\widehat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}/\boldsymbol{y}]$
  - $p_{\boldsymbol{x}/\boldsymbol{y}}$ the *à posteriori* distribution (used for MAP detector)
  - Proof: see Appendix D.1
  - Its linear: $(\widehat{\boldsymbol{x}_1 + \boldsymbol{x}_2}) = \widehat{\boldsymbol{x}}_1 + \widehat{\boldsymbol{x}}_2$

$$\widehat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}/\boldsymbol{y}]$$

Auto- & Cross- correlation

- **Autocorrelation** generalizes mean-square
  - Samples of $R_{\boldsymbol{xx}}(\tau) = \mathbb{E}[\boldsymbol{x}(t) \cdot \boldsymbol{x}^*(t-\tau)]$ when time is dimension
  - If correspond to samples of vector
    - Frequency-time and/or space-time
    - suppressed τ here if space time, but usually $\tau = 0$

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

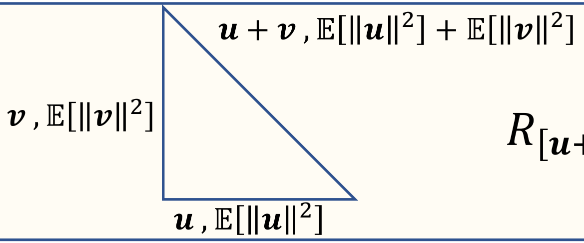
- **Energy** $\tau = 0$

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

- **Cross correlation** "generalizes" inner product
  - Samples of $R_{xx}(\tau) = \mathbb{E}[x(t) \cdot y^*(t-\tau)]$
  - Vectors can be different lengths L_x and L_y
  - "uncorrelated" (=0) → **orthogonal**

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

- **Pythagorus** IF uncorrelated $R_{\boldsymbol{uv}} = 0$
  - Generalizes "variances of uncorrelated random variables add"



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

The Joint Gaussian Distribution

- Completely specified by autocorrelation (and cross correlation)

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

- Its 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?
  - $|R_{yy}| > 0$
  - $|R_{xx}|$? $|R|$? – use pseudoinverse and determinant as product of **nonzero** eigenvalues

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

It's linear (for Gaussian)

$W = R_{xy} \cdot R_{yy}^{+}$ if singular

MMSE and AWGN Best Transmission are fundamentally connected

Linear MMSE & The Orthogonality Principle

Section D.2

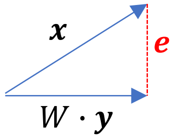
Linear MMSE: any joint distribution of x and y

- Given random $\boldsymbol{x}$ and $\boldsymbol{y}$, want to estimate $\boldsymbol{x}$, $\widehat{\boldsymbol{x}} = W \cdot \boldsymbol{y}$
  - Know both $p_{\boldsymbol{x},\boldsymbol{y}}$ and specific observed $\boldsymbol{y} = \boldsymbol{v}$
- The error:

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

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

Orthogonality Principle



- 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 estimator may not be linear if non-Gaussian

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

- Also again: $\widehat{(\boldsymbol{x}_1 + \boldsymbol{x}_2)} = \widehat{\boldsymbol{x}}_1 + \widehat{\boldsymbol{x}}_2$ or $\widehat{A \cdot \boldsymbol{x}} = A \cdot \widehat{\boldsymbol{x}}$

Vector and Matrix Norms

- The trace of an autocorrelation matrix 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

$$\boldsymbol{e}' = Q \cdot \boldsymbol{e} \qquad diagonal\!: \; R_{\boldsymbol{e}'\boldsymbol{e}'} = Q \cdot R_{\boldsymbol{ee}} \cdot Q^*$$

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

- The determinant of an autocorrelation matrix is the product of its eigenvalues

- 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→$Q \cdot W$) minimized, so then $|\, R_{\boldsymbol{ee}}|$= $|\, R_{\boldsymbol{e}'\boldsymbol{e}'}|$ is also minimized → Minimizing sum (trace) here is same as minimizing product (determinant).

Matrix SNR?

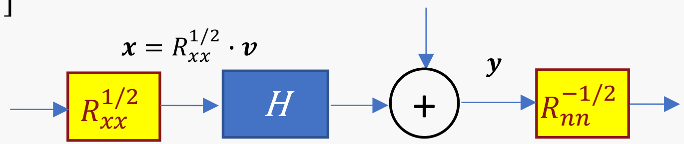
$\mathcal{E}_{\boldsymbol{x}} \rightarrow trace\,\{R_{\boldsymbol{xx}}\}$

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

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

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

$R_{\boldsymbol{vv}} = I$



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

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

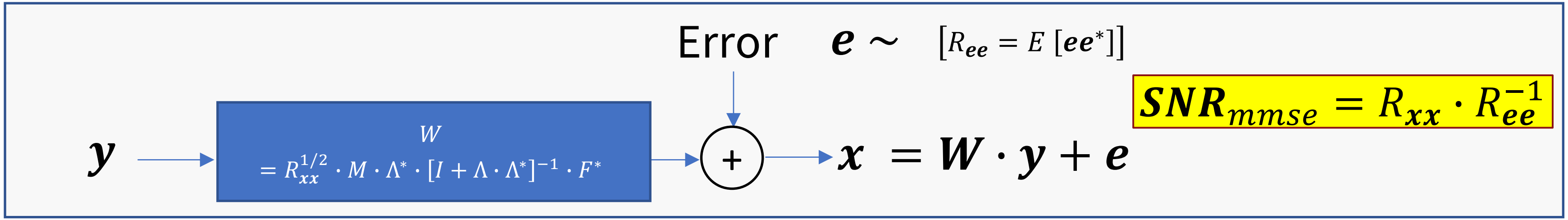
$$\widetilde{H} \triangleq R_{nn}^{-1/2} \cdot H \cdot R_{xx}^{1/2} = F \cdot \Lambda \cdot M^*$$

- It's like the parallel channels (take determinants) $SNR_{out} = |R_{\boldsymbol{yy}}| \cdot |R_{\boldsymbol{nn}}^{-1}| = \frac{|R_{\boldsymbol{yy}}|}{|R_{\boldsymbol{nn}}|}$
  - "Automatic" vector code
  - Bit rate: $b = \log_2(SNR_{out})$

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

- This set depends on $R_{\boldsymbol{xx}}$ choice, while earlier only $trace\,\{R_{\boldsymbol{xx}}\}$ was fixed
  - Water-fill $R_{\boldsymbol{xx}} = M \cdot diag\{\mathcal{E}_{water-fill}\} \cdot M^*$ maximizes

Backward Channel and Matrix SNR



- How about "backward channel" (MMSE) $\boldsymbol{SNR}$?

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

$$= SNR_{out}$$

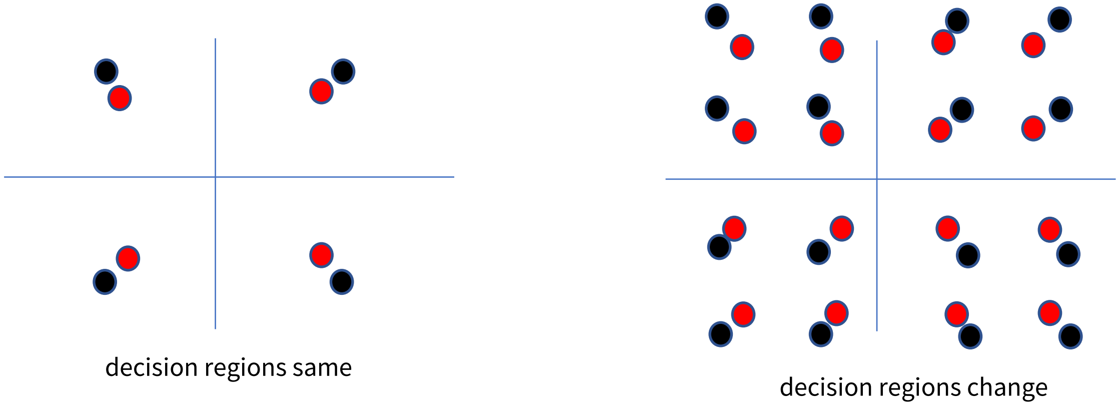
- Bit rate: $b = \log_2(SNR_{mmse})$
- $M^* \cdot W$ will estimate $\boldsymbol{x}'$ (linearity of MMSE estimates)
- Optimizing determinants is same as optimizing MSE/traces

Forward and backward have same SNR and "bit rate" (continuous $\boldsymbol{x}$ distribution)

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
  - $\mathbb{E}\left[{}^{\hat{x}}/_{x}\right] = \left(I - R_{ee} \cdot R_{xx}^{-1}\right) \cdot \boldsymbol{x} = \left(I - \boldsymbol{SNR}^{-1}\right) \cdot \boldsymbol{x}$



MMSE trades a little signal reduction for simultaneous noise reduction when minimizing the error

See PS2.1 (Prob 4.29)

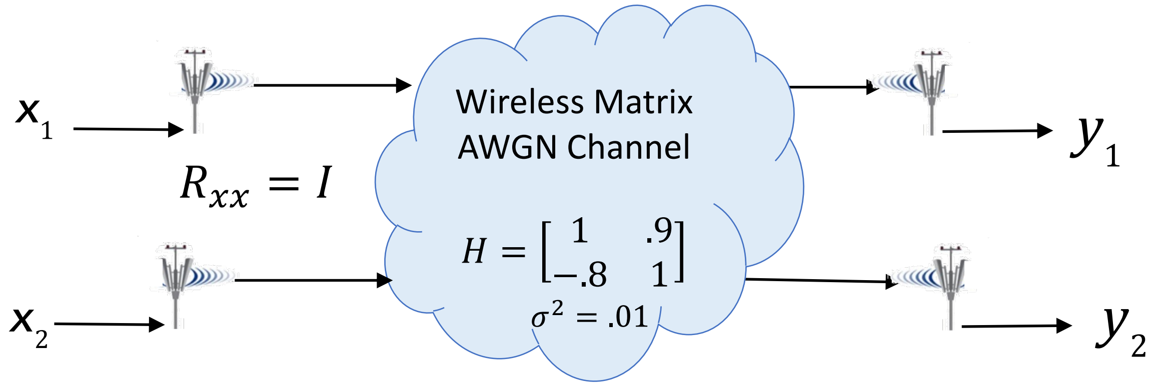
- For scalar case above, removal is scale up (by $\frac{SNR_{mmse}}{SNR_{mmse}-1}$)

- MIMO case, same per dimension, scale up (by $\frac{SNR_{mmse,n}}{SNR_{mmse,n}-1}$) **IF** MMSE R_{ee} is diagonal (vector coding)
  - IF not diagonal? (we'll learn what to do in later lectures)

Linear Matrix MMSE Examples

See PS2.2 (Prob 4.36)

2 x 2 Antenna System




```
>> 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
```

- Strong Crosstalk case from Chapter 1

  Basically the same Lecture 1 result, even without the "M" discrete modulator
  but why with no M on transmit?

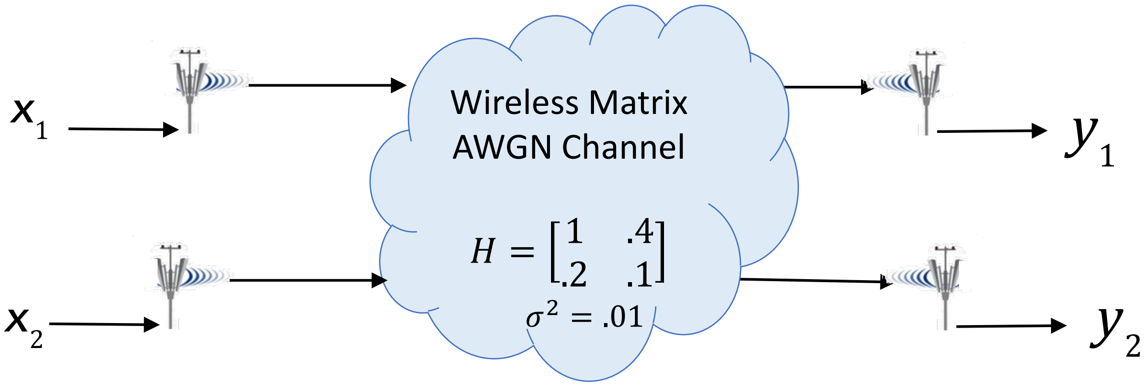
```
>> Mstar =

   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.

Previously in L1 was $\log_2(1 + 2 \cdot g_2) = 6.93$ bits/subsymbol
But this time, two dimensions are used, and the ML detectors are interdependent

```
>>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

```
>> 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

(All this loss can be recovered with Chapter 5's MMSE GDFE, in addition to using vector coding, so more than 1 solution)

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, MMSE-LE

Best with full ML is 8.8 dB, but requires Input WF energy distribution

Information Measures:
generalizing MMSE to all distributions

See PS2.3 (Prob 2.10)

Information Measures

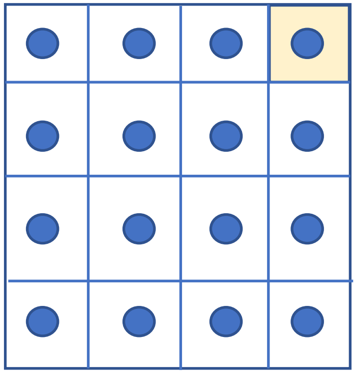
Gaussian Distribution	Any Distribution
Mean-square energy $\mathcal{E}_x = \mathbb{E}[\lvert \boldsymbol{x} \rvert^2]$	Entropy $\mathcal{H}_{\boldsymbol{x}}$ *Section 2.3.1*
Mean-square error $\sigma_e^2 = \mathbb{E}[\lvert \boldsymbol{e} \rvert^2]$	Conditional Entropy $\mathcal{H}_{\boldsymbol{x}/\boldsymbol{y}}$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

The information-carried by random variable/process generalizes the energy concepts from MMSE/Gaussian analysis to the spread/randomness of their distribution

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

Example leading to Entropy understanding

16 SQ QAM - UNCODED



$|C| = 16 \,; M = 16$

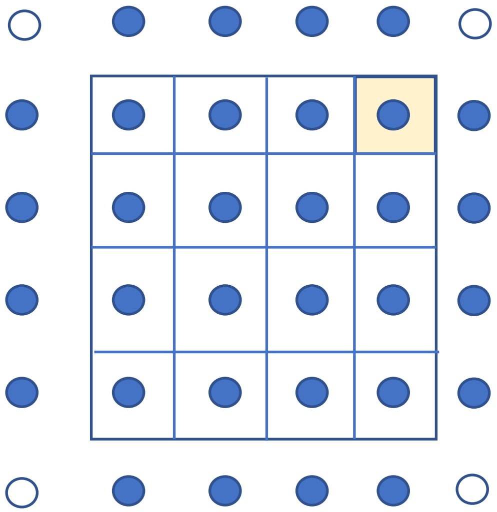
$b = 4 \,; \; \bar{b} = 2 \,; \; \tilde{b} = 4$

$N = \tilde{N} = 2 \,; \; \bar{N} = 1$ (or swap $\tilde{N}$ and $\bar{N}$)

32 CR QAM – UNCODED

$|C| = 32 \,; b = 5 \,; \; \bar{b} = 2.5$

$N = \tilde{N} = 2 \,; \; \bar{N} = 1$



6 PAM x 6 PAM CODED

$N = \bar{N} = 2 \,; \; \tilde{N} = 1 \,; \; |C| = 6 \; = 2^{2.59}$

If $\bar{b} = 2 \,, \bar{\rho} = 0.59$

(extra constellation points ~ redundancy)

- The subsymbol can have extra points, which means its coded
  - More redundant points, more dimensions → better codes

Information Theory Basics – Generalizes MMSE

- 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 distribution's, *information's,* many values, by probability (think subsymbols)

- generalizes bits/subsymbol where the constellation size $|C| \geq M^{1/\bar{N}} = 2^{\tilde{b}}$ the bits/subsymbol

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

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

- Uniform distribution has maximum entropy

$$\mathcal{H}_{\tilde{x}} \leq \log_2|C|$$

Binary example: $p_{\tilde{x}}(0) = \frac{1}{128}$ and $p_{\tilde{x}}(1) = \frac{127}{128}$

See PS2.3 (Prob 2.10)

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

Continuous Distribution – DIFFERENTIAL Entropy

- Differential Entropy

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

- Differential Entropy $\mathcal{H}_{\tilde{x}}$ is not same as approximating integral using discrete approx of $p_{\tilde{x}}(u)$
  - They differ by a constant that depends on the approximation-interval size

- Differential Entropy $\mathcal{H}_{\tilde{x}}$ does still however measure information content when subsymbols in codewords are chosen (usually at random) from $p_{\tilde{x}}(u)$.

- Maximum $\mathcal{H}_{\tilde{x}}$ 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 $$\mathcal{H}_{\tilde{x}} = \log_2(\pi e \mathcal{E}_{\tilde{x}}) \text{ bits/subsymbol}$$

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

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

$$\mathcal{H}_{\tilde{x}} = \log_2|\pi e R_{\tilde{x}\tilde{x}}| \text{ bits/complex-subsymbol}$$

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

Relation to MMSE Estimation

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

- $\mathcal{H}_{\tilde{\mathbf{x}}/\tilde{\mathbf{y}}} = \log_{\mathbf{2}}|\pi e R_{\boldsymbol{ee}}|$ - that is, the entropy is essentially just the log of the MMSE (Gaussian)
  - Entropy generalizes MMSE to any probability distribution
  - Measures the information content of the “miss” in estimating $\tilde{\boldsymbol{x}}$ from $\tilde{\boldsymbol{y}}$ for any $p_{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}}$

See PS2.5 (Prob 2.20)

Continuous distributions and entropy

- Complex Gaussian x

$$p_x(u) = \frac{1}{\pi\sigma_x^2} e^{-\frac{|x|^2}{\sigma_x^2}}$$

$$\mathcal{H}_x = \log_2\{\pi \cdot e \cdot \sigma_x^2\}$$

- Conditional x/y ?

$$\mathcal{H}_{x/y} = \log_2\left\{\pi \cdot e \cdot \sigma_{x/y}^2\right\}$$

$$\sigma_{x/y}^2 = \sigma_x^2 - {r_{xy}^2}/{\sigma_y^2} = \text{MMSE}$$

- Vector $\boldsymbol{x}$?

$$\mathcal{H}_{\boldsymbol{x}} = \log_2\left\{(\pi e)^{\bar{N}} \cdot |R_{\boldsymbol{xx}}|\right\}$$

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

(Appendix D on MMSE)

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

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

See PS2.5 (Prob 2.20)

Mutual Information and MMSE

Subsection 2.3.2

See PS2.5 (Prob 2.20)

Mutual Information ~ SNR

- Mutual Information $\mathbb{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}}) = \mathbb{E}\left[\log_2\left(\frac{p_{\tilde{x}\tilde{y}}}{p_{\tilde{x}}\cdot p_{\tilde{y}}}\right)\right] = \mathcal{H}_{\tilde{x}} - \mathcal{H}_{\tilde{x}/\tilde{y}} = \mathcal{H}_{\tilde{y}} - \mathcal{H}_{\tilde{y}/\tilde{x}}$
  - For discrete example = 1-.811 = .189 bits/subsymbol $= \mathcal{H}_{\tilde{\boldsymbol{x}}} - \mathcal{H}_{\tilde{\boldsymbol{x}}/\tilde{\boldsymbol{y}}} = \mathcal{H}_{\tilde{\boldsymbol{y}}} - \mathcal{H}_{\tilde{\boldsymbol{y}}/\tilde{\boldsymbol{x}}}$
- Symmetric in $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ (MMSE forward and backward channel)
- Amount of information common to $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$, $\mathbb{E}\left[\log_2\left(\frac{p_{\tilde{x}/\tilde{y}}}{p_{\tilde{y}}}\right)\right]$
  - On average, how much bigger is conditional versus uncond, in bits
- $\mathbb{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}})= \log_2\frac{|R_{\tilde{x}\tilde{x}}|}{|R_{ee}|} = \log_2\frac{|R_{\tilde{y}\tilde{y}}|}{|R_{nn}|} = \log_2\left(\left(1+SNR_{geo}\right)^{\bar{N}}\right)$ for the AWGN
- OR as earlier for vector coding $\mathbb{I}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{y}}) = \sum_{n=1}^{\bar{N}}\log_2 SNR_{mmse,n}$ for the AWGN

Law of Large Numbers

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*

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

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

- 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*-sequence has max entropy (uniform)
- The function of most interest in coding is $-\log_2[p_x(x)]$- that is the function itself is probability distribution's log
  - The sample average of this function converges to the entropy
  - Suggests choosing codewords (this means each subsymbol in the codeword) at random from stationary distribution
  - 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 entropy (and mutual information)

Random Coding

- 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

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

- LLN with function $p_{\tilde{\boldsymbol{x}}}$? (sample-average estimate of entropy)

$$\hat{\tilde{\mathcal{H}}}_{\tilde{\boldsymbol{x}}} = \frac{-1}{\overline{N}} \cdot \sum_{n=1}^{\overline{N}} \log_2\left[p(\tilde{\boldsymbol{x}}_n)\right] = \frac{-1}{\overline{N}} \cdot \log_2\left[p(\boldsymbol{x})\right] \quad .$$

LLN converges to (constant) $\tilde{\mathcal{H}}_{\boldsymbol{x}}$

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

Asymptotic Equal Partition (AEP)

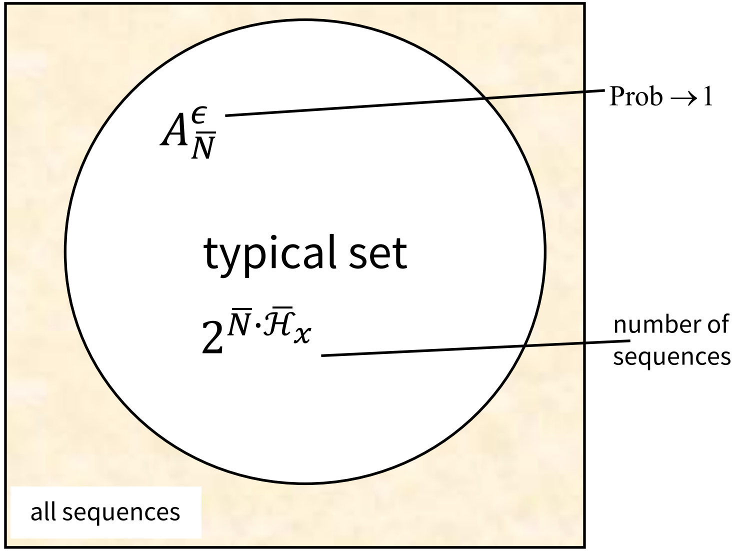
AEP Typical Sets

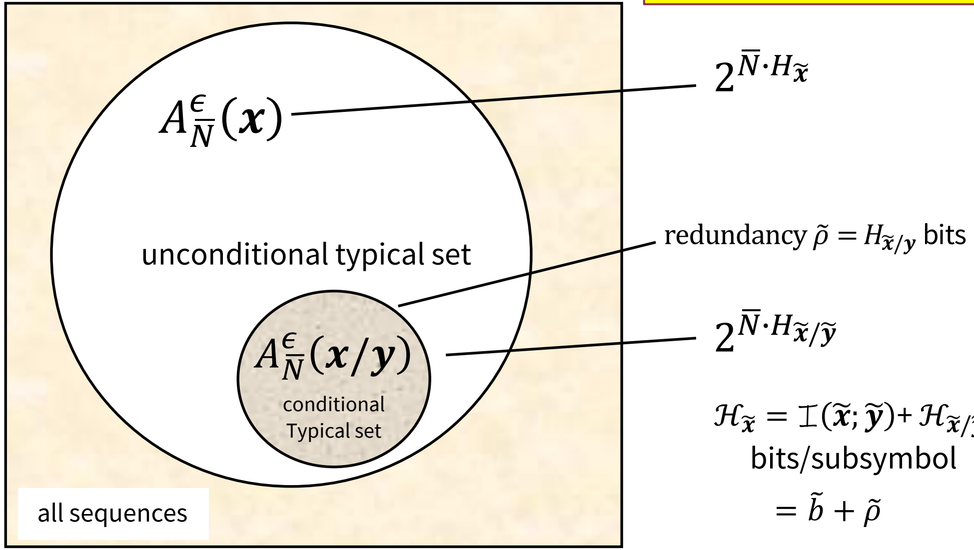
- The set is $$A_{\overline{N}}^{\epsilon}(\boldsymbol{x}) \triangleq \left\{ \boldsymbol{x} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_{\overline{N}}] \;\middle|\; 2^{-\overline{N}\cdot\mathcal{H}_{\tilde{\boldsymbol{x}}}-\epsilon} \leq p(\tilde{\boldsymbol{x}}_1,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

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





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

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

Formal Capacity Theorem

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

- Good codes will have only 1 codeword per conditional entropy subset
- MAP detector decision region is then ~ $A_{\bar{N}}^{\epsilon}(\boldsymbol{x}/\boldsymbol{y})$ - on average, but can find for one good code
- If $A_{\bar{N}}^{\epsilon}(\boldsymbol{x})$ were any larger, all codes (good or bad) will have at least one $A_{\bar{N}}^{\epsilon}(\boldsymbol{x}/\boldsymbol{y})$ 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 $b \le \mathbb{I}(\boldsymbol{x}\,;\boldsymbol{y}\,)$
  - Which is per-subsymbol equivalent with random code $\tilde{b} \le \mathbb{I}(\tilde{\boldsymbol{x}}\,;\tilde{\boldsymbol{y}}\,)$
  - If maximized over input distributions $\tilde{b} < \tilde{C} \le \max_{p_{\tilde{x}}} \mathbb{I}(\tilde{\boldsymbol{x}}\,;\tilde{\boldsymbol{y}}\,)$ bits/subsymbol

End Lecture 3