## Class Road Map - IT-I

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>9/25</td>
<td>Overview, Communications, Information, Entropy</td>
</tr>
<tr>
<td>L2</td>
<td>9/30</td>
<td>Entropy, Mutual Information, KL-Divergence</td>
</tr>
<tr>
<td>L3</td>
<td>10/2</td>
<td>More KL, Jensen, more Venn, Log Sum, Data Proc. Inequality</td>
</tr>
<tr>
<td>L4</td>
<td>10/7</td>
<td>Data Proc. Ineq., thermodynamics, Stats, Fano</td>
</tr>
<tr>
<td>L5</td>
<td>10/9</td>
<td>M. of Conv, AEP</td>
</tr>
<tr>
<td>L6</td>
<td>10/14</td>
<td>AEP, Source Coding, Types</td>
</tr>
<tr>
<td>LX</td>
<td>10/16</td>
<td>Makeup</td>
</tr>
<tr>
<td>L7</td>
<td>10/21</td>
<td>Types, Univ. Src Coding, Stoc. Procs, Entropy Rates</td>
</tr>
<tr>
<td>L8</td>
<td>10/23</td>
<td>Entropy rates, HMMs, Coding</td>
</tr>
<tr>
<td>L9</td>
<td>10/28</td>
<td>Kraft ineq., Shannon Codes, Kraft ineq. II, Huffman</td>
</tr>
<tr>
<td>L10</td>
<td>10/30</td>
<td>Huffman, Shannon/Fano/Elias</td>
</tr>
<tr>
<td>L11</td>
<td>11/4</td>
<td>Shannon/Fano/Elias, Games</td>
</tr>
<tr>
<td>LXX</td>
<td>11/6</td>
<td>In class midterm exam</td>
</tr>
<tr>
<td>L12</td>
<td>11/11</td>
<td>Vet’s day, makeup lecture: Arith. Coding, Background On Channel Capacity</td>
</tr>
<tr>
<td>L13</td>
<td>11/13</td>
<td>Channel Capacity, Ex. DMC</td>
</tr>
<tr>
<td>L14</td>
<td>11/18</td>
<td>Ex. DMC, Properties, Joint AEP, Shannon’s 2nd Theorem.</td>
</tr>
<tr>
<td>L15</td>
<td>11/20</td>
<td>Joint AEP, Shannon’s 2nd Theorem,</td>
</tr>
<tr>
<td>L17</td>
<td>11/27</td>
<td>Hamming Codes, Differential Entropy</td>
</tr>
<tr>
<td>L18</td>
<td>12/2</td>
<td>Diff. Entropy, Gaussian Diff. Entropy, Max Entropy, Gaussian Channel</td>
</tr>
<tr>
<td>L19</td>
<td>12/4</td>
<td>Gaussian Channel, Gaussian Channel Capacity, Preview IT-II, Review.</td>
</tr>
<tr>
<td>LXX</td>
<td>12/10</td>
<td>Final exam</td>
</tr>
</tbody>
</table>

**Finals Week: December 9th–13th.**
Cumulative Outstanding Reading TODOs

- Chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Chapters 3 and 4 in our book (Cover & Thomas, “Information Theory”).
- Sections 11.1 through 11.3 in our book (Cover & Thomas, “Information Theory”).
- Chapter 4 in our book (Cover & Thomas, “Information Theory”).
- Chapter 5 in our book (Cover & Thomas, “Information Theory”).
- Chapter 13 in our book (Cover & Thomas, “Information Theory”) (there is no chapter on arithmetic coding but the lecture slides will be complete, or see MacKay’s online text).
- Chapter 7 in our book (Cover & Thomas, “Information Theory”) on channel capacity.
- Chapter 8 and 9 in our book (Cover & Thomas, “Information Theory”) on differential entropy and the Gaussian channel.
Homework

- **Homework 1** on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), was due Tuesday, Oct 8th, 11:55pm.

- **Homework 2** on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Friday 10/18/2019, 11:45pm.

- **Homework 3** on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Tuesday 10/29/2019, 11:45pm.

- **Homework 4** on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Monday 11/4/2019, 11:45pm.

- **Homework 5** on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Monday 12/2/2019, 11:45pm.
Continuous/Differential Entropy

- Let $X$ now be a continuous r.v. with cumulative distribution

$$F(x) = \Pr(X \leq x)$$  \hspace{1cm} (18.5)

and $f(x) = \frac{d}{dx}F(x)$ is the density function.

- Let $S = \{x : f(x) > 0\}$ be the support set. Then

**Definition 18.2.1 (differential entropy $h(X)$)**

$$h(X) = - \int_S f(x) \log f(x) \, dx$$  \hspace{1cm} (18.6)

- Since we integrate over only the support set, no worries about $\log 0$.
- Let's jump right into some examples . . .
Continuous Entropy Of Uniform Distribution

Here, $X \sim U[0, a]$ with $a \in \mathbb{R}_{++}$.

Then

$$h(X) = -\int_{0}^{a} \frac{1}{a} \log \frac{1}{a} dx = -\log \frac{1}{a} \tag{18.5}$$

Note: continuous entropy can be both positive or negative.

How can entropy (which we know to mean “uncertainty”, or “information”) be negative?

In fact, entropy (as we’ve seen perhaps once or twice) can be interpreted as the exponent of the “count” or “volume” of a typical set.

Example: $2^{H(X)}$ is the number of things that happen, on average. We easily can have that $2^{H(X)} \ll |\mathcal{X}|$ which is what allows compression.

Consider a uniform r.v. $Y$ such that $2^{H(Y)} = |\mathcal{Y}|$.

Thus, negative exponent means the count, or “volume” is small.
Continuous Entropy Of Normal Distribution

- Normal (Gaussian) distributions are very important.
- Univariate (1D) Gaussian with $\mu$-mean and variance $\sigma^2$:

$$X \sim N(\mu, \sigma^2) \iff f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \quad (18.5)$$

- Let's compute the entropy of $f$ in nats.

$$h(X) = -\int f \ln f = -\int f(x) \left[ -\frac{(x - \mu)^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \quad (18.6)$$

$$\frac{E(X - \mu)^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \quad (18.7)$$

$$= \frac{1}{2} \ln e + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} \ln(2\pi e\sigma^2) \text{nats} \times \left( \frac{1}{\ln 2} \text{bits/nats} \right)$$

$$= \frac{1}{2} \log(2\pi e\sigma^2) \text{bits} \quad (18.8)$$

- Note: only a function of the variance $\sigma^2$, not the mean. Why?
- So entropy of a Gaussian is monotonically related to the variance.
Things are similar for the continuous case. Indeed

**Theorem 18.2.1**

Let $X_1, X_2, \ldots, X_n$ be a sequence of r.v.’s, i.i.d. $\sim f(x)$. Then

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \to E[-\log f(X)] = h(X)$$

(18.5)

this follows via the weak law of large numbers (WLLN) just like in the discrete case.

**Definition 18.2.2**

$$A_{\epsilon}^{(n)} = \{x_{1:n} \in S^n : | -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X)| \leq \epsilon\}$$

- Note: $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.
- Thus, we have upper/lower bounds on the probability

$$2^{-n(h+\epsilon)} \leq f(x_{1:n}) \leq 2^{-n(h-\epsilon)}$$

(18.6)
Differential vs. Discrete Entropy

- Let $X \sim f(x)$, and divide the range of $X$ up into bins of length $\Delta$. 
Differential vs. Discrete Entropy

Let $X \sim f(x)$, and divide the range of $X$ up into bins of length $\Delta$.

E.g., quantize range of $X$ using $n$ bits, $2^n$ values, so $\Delta = 2^{-n}$. 
Let $X \sim f(x)$, and divide the range of $X$ up into bins of length $\Delta$. 

- E.g., quantize range of $X$ using $n$ bits, $2^n$ values, so $\Delta = 2^{-n}$.

- We can then view this as follows:

\[
f(x) \quad \text{with} \quad \Delta = 2^{-n}
\]
Let $X \sim f(x)$, and divide the range of $X$ up into bins of length $\Delta$.

E.g., quantize range of $X$ using $n$ bits, $2^n$ values, so $\Delta = 2^{-n}$.

We can then view this as follows:

$$f(x) \Delta = 2^{-n}$$

Mean value theorem, i.e., that if continuous within bin $\exists x_i$ such that

$$f(x_i) = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x)dx \quad (18.1)$$
Create a quantized random variable $X^\Delta$ having those values so that

$$X^\Delta = x_i \quad \text{if} \quad i\Delta \leq X < (i + 1)\Delta$$  \hspace{1cm} (18.2)
Create a quantized random variable $X^\Delta$ having those values so that

$$X^\Delta = x_i \quad \text{if} \quad i\Delta \leq X < (i + 1)\Delta \quad (18.2)$$

This gives a discrete distribution

$$\Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx = \Delta \times f(x_i) \quad (18.3)$$
Differential vs. Discrete Entropy

- Create a quantized random variable $X^\Delta$ having those values so that
  
  \[ X^\Delta = x_i \quad \text{if} \quad i\Delta \leq X < (i + 1)\Delta \]  
  \hspace{1cm} (18.2)

- This gives a discrete distribution
  
  \[ \Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx = \Delta \times f(x_i) \]  
  \hspace{1cm} (18.3)

  and we can calculate the entropy

  \[ H(X^\Delta) \]  
  \hspace{1cm} (18.6)
Differential vs. Discrete Entropy

- Create a quantized random variable $X^\Delta$ having those values so that
  \[ X^\Delta = x_i \text{ if } i\Delta \leq X < (i + 1)\Delta \]  
  (18.2)

- This gives a discrete distribution
  \[ \Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx = \Delta \times f(x_i) \]  
  (18.3)

and we can calculate the entropy

\[
H(X^\Delta) = - \sum_{i=-\infty}^{\infty} p_i \log p_i = - \sum_i (\Delta f(x_i)) \log(f(x_i)\Delta) 
\]  
(18.4)

(18.6)
Differential vs. Discrete Entropy

- Create a quantized random variable $X^\Delta$ having those values so that

$$X^\Delta = x_i \quad \text{if} \quad i\Delta \leq X < (i + 1)\Delta \quad (18.2)$$

- This gives a discrete distribution

$$\Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \Delta \times f(x_i) \quad (18.3)$$

and we can calculate the entropy

$$H(X^\Delta) = - \sum_{i=-\infty}^{\infty} p_i \log p_i = - \sum_{i} (\Delta f(x_i)) \log(f(x_i)\Delta) \quad (18.4)$$

$$= - \sum_{i} \Delta f(x_i) \log f(x_i) - \sum_{i} \Delta f(x_i) \log \Delta \quad (18.5)$$

$$= (18.6)$$
Differential vs. Discrete Entropy

- Create a quantized random variable $X^\Delta$ having those values so that

$$X^\Delta = x_i \text{ if } i\Delta \leq X < (i + 1)\Delta \quad (18.2)$$

- This gives a discrete distribution

$$\Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx = \Delta \times f(x_i) \quad (18.3)$$

and we can calculate the entropy

$$H(X^\Delta) = - \sum_{i=-\infty}^{\infty} p_i \log p_i = - \sum_{i} (\Delta f(x_i)) \log(f(x_i)\Delta) \quad (18.4)$$

$$= - \sum_{i} \Delta f(x_i) \log f(x_i) - \sum_{i} \Delta f(x_i) \log \Delta \quad (18.5)$$

$$= - \sum_{i} \Delta f(x_i) \log f(x_i) - \log \Delta \quad (18.6)$$
This follows since (as expected)

\[
\sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx \right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1
\]

(18.7)
Differential vs. Discrete Entropy

- This follows since (as expected)

\[
\sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i \Delta}^{(i+1)\Delta} f(x) \, dx \right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1
\]

(18.7)

- Also, as \( \Delta \to 0 \), we have \(- \log \Delta \to \infty \) and (assuming all is integrable in a Riemannian sense)

\[
- \sum_i \Delta f(x_i) \log f(x_i) \to - \int f(x) \log f(x) \, dx
\]

(18.8)
This follows since (as expected)

\[
\sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx \right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1
\]  

(18.7)

Also, as \( \Delta \to 0 \), we have \(- \log \Delta \to \infty \) and (assuming all is integrable in a Riemannian sense)

\[
- \sum_i \Delta f(x_i) \log f(x_i) \to - \int f(x) \log f(x) \, dx
\]  

(18.8)

So, \( H(X^\Delta) + \log \Delta \to h(f) \) as \( \Delta \to 0 \).
This follows since (as expected)

\[ \sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx \right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1 \]  

(18.7)

Also, as \( \Delta \to 0 \), we have \(- \log \Delta \to \infty \) and (assuming all is integrable in a Riemannian sense)

\[ - \sum_i \Delta f(x_i) \log f(x_i) \to - \int f(x) \log f(x) \, dx \]  

(18.8)

So, \( H(X^\Delta) + \log \Delta \to h(f) \) as \( \Delta \to 0 \).

Loosely, \( h(f) \approx H(X^\Delta) + \log \Delta \) and for an \( n \)-bit quantization with \( \Delta = 2^{-n} \), we have

\[ H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n \]  

(18.9)
Differential vs. Discrete Entropy

This follows since (as expected)

$$\sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx \right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1$$

(18.7)

Also, as $\Delta \to 0$, we have $-\log \Delta \to \infty$ and (assuming all is integrable in a Riemannian sense)

$$-\sum_i \Delta f(x_i) \log f(x_i) \to -\int f(x) \log f(x) \, dx$$

(18.8)

So, $H(X^\Delta) + \log \Delta \to h(f)$ as $\Delta \to 0$.

Loosely, $h(f) \approx H(X^\Delta) + \log \Delta$ and for an $n$-bit quantization with $\Delta = 2^{-n}$, we have

$$H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n$$

(18.9)

This means that as $n \to \infty$, $H(X^\Delta)$ can get larger.
This follows since (as expected)

\[
\sum_i \Delta f(x_i) = \sum_i \Delta \left(\frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx\right) = \Delta \frac{1}{\Delta} \int f(x) \, dx = 1 
\]  

(18.7)

Also, as \( \Delta \to 0 \), we have \(- \log \Delta \to \infty\) and (assuming all is integrable in a Riemannian sense)

\[
- \sum_i \Delta f(x_i) \log f(x_i) \to - \int f(x) \log f(x) \, dx 
\]  

(18.8)

So, \( H(X^\Delta) + \log \Delta \to h(f) \) as \( \Delta \to 0 \).

Loosely, \( h(f) \approx H(X^\Delta) + \log \Delta \) and for an \( n\)-bit quantization with \( \Delta = 2^{-n} \), we have

\[
H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n 
\]  

(18.9)

This means that as \( n \to \infty \), \( H(X^\Delta) \) can get larger. Why?
I.e., we start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy, so $2^n$ values.
I.e., we start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy, so $2^n$ values.

For a discrete representation to represent $2^n$ values, we expect the entropy to go up with $n$, and as $n$ gets large so would the entropy, but then adjusted by $h(X)$ via

$$H(X^\Delta) \approx h(X) - \log \Delta = h(X) - \log 2^{-n} = h(X) + n$$
Differential vs. Discrete Entropy

- I.e., we start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy, so $2^n$ values.

- For a discrete representation to represent $2^n$ values, we expect the entropy to go up with $n$, and as $n$ gets large so would the entropy, but then adjusted by $h(X)$ via

$$H(X^\Delta) \approx h(X) - \log \Delta = h(X) - \log 2^{-n} = h(X) + n$$

- $H(X^\Delta)$ is the number of bits to describe this $n$-bit equally spaced quantization of the continuous random variable $X$. 

Differential vs. Discrete Entropy

- I.e., we start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy, so $2^n$ values.

- For a discrete representation to represent $2^n$ values, we expect the entropy to go up with $n$, and as $n$ gets large so would the entropy, but then adjusted by $h(X)$ via

\[ H(X^\Delta) \approx h(X) - \log \Delta = h(X) - \log 2^{-n} = h(X) + n \]

- $H(X^\Delta)$ is the number of bits to describe this $n$-bit equally spaced quantization of the continuous random variable $X$.

- $H(X^\Delta) \approx h(X) + n$ says that it might take either more than $n$ bits to describe $X$ at $n$-bit accuracy, or less than $n$ bits to describe $X$ at $n$-bit accuracy, depending on the concentration of $X$. 
Differential vs. Discrete Entropy

- I.e., we start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy, so $2^n$ values.

- For a discrete representation to represent $2^n$ values, we expect the entropy to go up with $n$, and as $n$ gets large so would the entropy, but then adjusted by $h(X)$ via
  $$H(X^\Delta) \approx h(X) - \log \Delta = h(X) - \log 2^{-n} = h(X) + n$$

- $H(X^\Delta)$ is the number of bits to describe this $n$-bit equally spaced quantization of the continuous random variable $X$.

- $H(X^\Delta) \approx h(X) + n$ says that it might take either more than $n$ bits to describe $X$ at $n$-bit accuracy, or less than $n$ bits to describe $X$ at $n$-bit accuracy, depending on the concentration of $X$.

- If $X$ is very concentrated $h(f) < 0$ then fewer bits. If $X$ is very spread out, then more than $n$ bits.
Joint Differential Entropy

- Like discrete case, we have entropy for vectors of r.v.s
Joint Differential Entropy

- Like discrete case, we have entropy for vectors of r.v.s
- The joint differential entropy is defined as:

\[ h(X_1, X_2, \ldots, X_n) = - \int f(x_1:n) \log f(x_1:n) dx_{1:n} \quad (18.10) \]
Joint Differential Entropy

- Like discrete case, we have entropy for vectors of r.v.s
- The joint differential entropy is defined as:

$$h(X_1, X_2, \ldots, X_n) = -\int f(x_1:n) \log f(x_1:n) dx_1:n$$  \hfill (18.10)

- Conditional differential entropy

$$h(X|Y) = -\int f(x, y) \log f(x|y) dxdy = h(X, Y) - h(Y)$$ \hfill (18.11)
Entropy of a Multivariate Gaussian

- When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$  \hspace{1cm} (18.12)
When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$  \hspace{1cm} (18.12)$$

then the entropy of $X$ has a nice form, in particular

$$h(X) = \frac{1}{2} \log \left[ (2\pi e)^n |\Sigma| \right] \text{ bits}$$  \hspace{1cm} (18.13)$$
Entropy of a Multivariate Gaussian

- When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$  \hspace{1cm} (18.12)$$

then the entropy of $X$ has a nice form, in particular

$$h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma|\right] \text{ bits}$$  \hspace{1cm} (18.13)$$

- Notice that the entropy is monotonically related to the determinant of the covariance matrix $\Sigma$ and is not at all dependent on the mean $\mu$. 
When \( X \) is distributed according to a multivariate Gaussian distribution, i.e.,

\[
X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu)}
\]  

(18.12)

then the entropy of \( X \) has a nice form, in particular

\[
h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma| \right] \text{ bits}
\]  

(18.13)

Notice that the entropy is monotonically related to the determinant of the covariance matrix \( \Sigma \) and is not at all dependent on the mean \( \mu \).

If we solve for \( |\Sigma| \) as a function of entropy, get \( \propto 2^{2h(X)} \).
Entropy of a Multivariate Gaussian

- When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

then the entropy of $X$ has a nice form, in particular

$$h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma|\right] \text{ bits}$$

- Notice that the entropy is monotonically related to the determinant of the covariance matrix $\Sigma$ and is not at all dependent on the mean $\mu$.
- If we solve for $|\Sigma|$ as a function of entropy, get $\propto 2^{2h(X)}$.
- The determinant of covariance is a form of spread, or dispersion of the distribution.
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

$h(X)$
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

$$h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right]$$

(18.14)
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

\[ h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right] \]

\[ = \frac{1}{2} E_f \left[ \text{tr} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \]

(18.14)

(18.15)

(18.20)
Entropy of a Multivariate Gaussian r.v. \( X \): Derivation

\[
h(X) = -\int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right]
\]

\[
= \frac{1}{2} E_f \left[ \text{tr} \left( x - \mu \right)^T \Sigma^{-1} \left( x - \mu \right) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)
\]

\[
= \frac{1}{2} E_f \left[ \text{tr} (x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)
\]
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

\[ h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right] \]

(18.14)

\[ = \frac{1}{2} E_f \left[ \text{tr} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \]

(18.15)

\[ = \frac{1}{2} E_f \left[ \text{tr}(x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \]

(18.16)

\[ = \frac{1}{2} \text{tr} E_f [(x - \mu)(x - \mu)^T] \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \]

(18.17)

\[ = \frac{1}{2} \ln \left( \frac{1}{2\pi e} \right) \]

(18.20)
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

\[
h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right] \]

(18.14)

\[
= \frac{1}{2} E_f \left[ \text{tr} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) 

(18.15)

\[
= \frac{1}{2} E_f \left[ \text{tr} (x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) 

(18.16)

\[
= \frac{1}{2} \text{tr} E_f \left[ (x - \mu)(x - \mu)^T \right] \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) 

(18.17)

\[
= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) 

(18.18)

\[
= \frac{1}{2} \ln \left( (2\pi e)^n \right) 

(18.20)
\]
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

$$h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right]$$

$$= \frac{1}{2} E_f \left[ \text{tr} \left( x - \mu \right)^T \Sigma^{-1} \left( x - \mu \right) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \tag{18.15}$$

$$= \frac{1}{2} E_f \left[ \text{tr}(x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \tag{18.16}$$

$$= \frac{1}{2} \text{tr} E_f \left[ (x - \mu)(x - \mu)^T \right] \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \tag{18.17}$$

$$= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \tag{18.18}$$

$$= \frac{1}{2} \text{tr} I + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \tag{18.19}$$

$$= \frac{1}{2} \ln \left( (2\pi e)^n |\Sigma| \right) \tag{18.20}$$
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

$$h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right]$$

(18.14)

$$= \frac{1}{2} E_f \left[ \text{tr} \left( x - \mu \right)^T \Sigma^{-1} (x - \mu) \right] + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.15)

$$= \frac{1}{2} E_f \left[ \text{tr} (x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.16)

$$= \frac{1}{2} \text{tr} E_f \left[ (x - \mu)(x - \mu)^T \right] \Sigma^{-1} + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.17)

$$= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.18)

$$= \frac{1}{2} \text{tr} I + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.19)

$$= \frac{n}{2} + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right]$$

(18.20)
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

$$h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right]$$

(18.14)

$$= \frac{1}{2} E_f \left[ \text{tr} \left( (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)$$

(18.15)

$$= \frac{1}{2} E_f \left[ \text{tr} (x - \mu)(x - \mu)^T \Sigma^{-1} \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)$$

(18.16)

$$= \frac{1}{2} \text{tr} E_f \left[ (x - \mu)(x - \mu)^T \right] \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)$$

(18.17)

$$= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)$$

(18.18)

$$= \frac{1}{2} \text{tr} I + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right)$$

(18.19)

$$= \frac{n}{2} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) = \frac{1}{2} \ln \left( (2\pi e)^n |\Sigma| \right)$$

(18.20)
Entropy of a Multivariate Gaussian r.v. $X$: Derivation

\[
h(X) = -\int f(x) \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right] \]

\[
= \frac{1}{2} E_f \left[ \text{tr} \left( (x - \mu)(x - \mu)^T \Sigma^{-1} \right) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \quad (18.15)
\]

\[
= \frac{1}{2} E_f \left[ \text{tr} \left( (x - \mu)(x - \mu)^T \Sigma^{-1} \right) \right] + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \quad (18.16)
\]

\[
= \frac{1}{2} \text{tr} \left[ E_f \left( (x - \mu)(x - \mu)^T \right) \right] \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \quad (18.17)
\]

\[
= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \quad (18.18)
\]

\[
= \frac{1}{2} \text{tr} I + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \quad (18.19)
\]

\[
= \frac{n}{2} + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) = \frac{1}{2} \ln \left( (2\pi e)^n |\Sigma| \right) \quad (18.20)
\]

This uses the "trace trick", that $\text{tr}(ABC) = \text{tr}(CAB)$.
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

\[ D(f \| g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \]  

(18.21)
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

\[ D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \geq 0 \] (18.21)

We can, like in the discrete case, use Jensen’s inequality to prove the non-negativity of \( D(f||g) \).
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form:

\[ D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \quad (18.21) \]

We can, like in the discrete case, use Jensen's inequality to prove the non-negativity of \( D(f||g) \).

**Mutual Information:**

\[ D(f(X, Y)||f(X)f(Y)) = I(X; Y) = h(X) - h(X|Y) \quad (18.22) \]
\[ = h(Y) - h(Y|X) \geq 0 \quad (18.23) \]
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

\[ D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \quad (18.21) \]

We can, like in the discrete case, use Jensen’s inequality to prove the non-negativity of \( D(f||g) \).

Mutual Information:

\[ D(f(X,Y)||f(X)f(Y)) = I(X;Y) = h(X) - h(X|Y) \quad (18.22) \]

\[ = h(Y) - h(Y|X) \geq 0 \quad (18.23) \]

Thus, since \( I(X;Y) \geq 0 \) we have again that conditioning reduces entropy, i.e., \( h(Y) \geq h(Y|X) \).
Chain rules and more

- We still have chain rules

\[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  (18.24)
Chain rules and more

- We still have chain rules

\[
h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \tag{18.24}
\]

- And bounds of the form

\[
\sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \tag{18.25}
\]
Chain rules and more

- We still have chain rules

\[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \] (18.24)

- And bounds of the form

\[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \] (18.25)

- For discrete entropy, we have monotonicity. i.e.,

\[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}) \]
Chain rules and more

- We still have chain rules

\[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  \hspace{1cm} (18.24)

- And bounds of the form

\[ \sum_i h(X_i | X_{1:n} \setminus \{i\}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \]  \hspace{1cm} (18.25)

- For discrete entropy, we have monotonicity. I.e.,

\[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}) \]

More generally

\[ f(A) = H(X_A) \]  \hspace{1cm} (18.26)

is monotonic non-decreasing in set \( A \) (i.e., \( f(A) \leq f(B), \forall A \subseteq B \)).
Chain rules and more

- We still have chain rules
  \[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  
  \hspace{1cm} (18.24)

- And bounds of the form
  \[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \]  
  \hspace{1cm} (18.25)

- For discrete entropy, we have monotonicity. I.e.,
  \[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}). \]
  More generally
  \[ f(A) = H(X_A) \]  
  \hspace{1cm} (18.26)
  is monotonic non-decreasing in set \( A \) (i.e., \( f(A) \leq f(B), \forall A \subseteq B \)).

- Is \( f(A) = h(X_A) \) monotonic?
Chain rules and more

- We still have chain rules
  \[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  
  (18.24)

- And bounds of the form
  \[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \]  
  (18.25)

- For discrete entropy, we have monotonicity. I.e.,
  \[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}) \]
  More generally
  \[ f(A) = H(X_A) \]  
  (18.26)
  is monotonic non-decreasing in set \( A \) (i.e., \( f(A) \leq f(B), \forall A \subseteq B \)).

- Is \( f(A) = h(X_A) \) monotonic? No, consider Gaussian entropy with diagonal \( \Sigma \) with small diagonal values. So \( h(X) = \frac{1}{2} \log [(2\pi e)^n |\Sigma|] \) can get smaller with more random variables.
Chain rules and more

- We still have chain rules

\[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  \hspace{1cm} (18.24)

- And bounds of the form

\[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \] \hspace{1cm} (18.25)

- For discrete entropy, we have monotonicity. I.e.,

\[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}) \]  

More generally

\[ f(A) = H(X_A) \]  \hspace{1cm} (18.26)

is monotonic non-decreasing in set \( A \) (i.e., \( f(A) \leq f(B), \forall A \subseteq B \)).

- Is \( f(A) = h(X_A) \) monotonic? No, consider Gaussian entropy with diagonal \( \Sigma \) with small diagonal values. So \( h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma|\right] \) can get smaller with more random variables.

- Similarly, when some variables independent, adding independent variables with negative entropy can decrease overall entropy.
Hadamard’s Inequality

Using differential entropy, we can sometimes get known results in linear algebra.
Using differential entropy, we can sometimes get known results in linear algebra.

From \( h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \), consider the case where \( X_{1:n} \) is jointly Gaussian \( \sim \mathcal{N}(\mu, K) \).
Using differential entropy, we can sometimes get known results in linear algebra.

From \( h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \), consider the case where \( X_{1:n} \) is jointly Gaussian \( \sim \mathcal{N}(\mu, K) \).

Then since \( \log \) is monotonic, we immediately get:

\[
|K| \leq \prod_{i=1}^{n} K_{ii} \quad (18.27)
\]

whenever \( K \) is positive semi-definite (a result known as Hadamard’s inequality).
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$. 

\[
X \sim \text{Gaussian} \quad \Rightarrow \quad X \quad \text{is \ symmetric \ positive \ definite}
\]
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^\top]$ which is known to be symmetric positive semidefinite.
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^\top]$ which is known to be symmetric positive semidefinite.
- Note that $C_{ij} = E[X_i X_j]$. 
Moments of Random Vectors

Let \( X_{1:n} \) be a continuous random vector.

- The first moment is \( \mu = E[X] \).
- The second moment is \( C = E[XX^\top] \) which is known to be symmetric positive semidefinite.
- Note that \( C_{ij} = E[X_iX_j] \).
- There are higher order moments as well, for example the third order moment has entries of the form \( C_{ijk} = E[X_iX_jX_k] \), the forth order moment has entries of the form \( C_{ijk\ell} = E[X_iX_jX_kX_\ell] \), and so on.
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^\top]$ which is known to be symmetric positive semidefinite.
- Note that $C_{ij} = E[X_iX_j]$.
- There are higher order moments as well, for example the third order moment has entries of the form $C_{ijk} = E[X_iX_jX_k]$, the forth order moment has entries of the form $C_{ijkl} = E[X_iX_jX_kX_\ell]$, and so on.
- Let $C^{(m)}$ be the $m^{th}$ order moment.
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^\top]$ which is known to be symmetric positive semidefinite.
- Note that $C_{ij} = E[X_iX_j]$.
- There are higher order moments as well, for example the third order moment has entries of the form $C_{ijk} = E[X_iX_jX_k]$, the fourth order moment has entries of the form $C_{ijkl} = E[X_iX_jX_kX_\ell]$, and so on.
- Let $C^{(m)}$ be the $m^{th}$ order moment.
- In general, an arbitrary (complex) random vector can have arbitrarily high non-zero $m^{th}$ order moments, and distributions can be characterized by their moments.
Moments of Random Vectors

- Let $X_{1:n}$ be a continuous random vector.
- The first moment is $\mu = E[X]$.
- The second moment is $C = E[XX^\top]$ which is known to be symmetric positive semidefinite.
- Note that $C_{ij} = E[X_iX_j]$.
- There are higher order moments as well, for example the third order moment has entries of the form $C_{ijk} = E[X_iX_jX_k]$, the forth order moment has entries of the form $C_{ijk\ell} = E[X_iX_jX_kX_\ell]$, and so on.
- Let $C^{(m)}$ be the $m^{th}$ order moment.
- In general, an arbitrary (complex) random vector can have arbitrarily high non-zero $m^{th}$ order moments, and distributions can be characterized by their moments.
- The multivariate Gaussian only requires the first and 2nd order moments, but all higher order moments are redundant given the first two (idea, the Gaussian can be parameterized exactly via its first and second order moments, statistics $X$ and $XX^\top$ are in 1-1 correspondence to standard mean/covariance).
Moments of Random Vectors

Consider first order moment \( C^{(1)} \) (a vector) and the second order moment \( C^{(2)} \) (a matrix). Consider all distributions that have these first and second order moments.
Consider first order moment $C^{(1)}$ (a vector) and the second order moment $C^{(2)}$ (a matrix). Consider all distributions that have these first and second order moments.

Out of all these distributions, which one would have the highest (differential) entropy?
Moments of Random Vectors

- Consider first order moment $C^{(1)}$ (a vector) and the second order moment $C^{(2)}$ (a matrix). Consider all distributions that have these first and second order moments.

- Out of all these distributions, which one would have the highest (differential) entropy?

- Consider what do moments do? They constrain the possible set of distributions to those having those particular moment values.
Moments of Random Vectors

- Consider first order moment $C^{(1)}$ (a vector) and the second order moment $C^{(2)}$ (a matrix). Consider all distributions that have these first and second order moments.

- Out of all these distributions, which one would have the highest (differential) entropy?

- Consider what do moments do? They constrain the possible set of distributions to those having those particular moment values.

- Intuitively: highest entropy should have fewest constraints, or highest entropy would be granted to the distributions with zero additional constraints on higher order moments. “constraint,” meaning being a particular value rather than being derivable from lower order moments.
Consider first order moment $C^{(1)}$ (a vector) and the second order moment $C^{(2)}$ (a matrix). Consider all distributions that have these first and second order moments.

Out of all these distributions, which one would have the highest (differential) entropy?

Consider what do moments do? They constrain the possible set of distributions to those having those particular moment values.

Intuitively: highest entropy should have fewest constraints, or highest entropy would be granted to the distributions with zero additional constraints on higher order moments. “constraint,” meaning being a particular value rather than being derivable from lower order moments.

Recall also that the first moment (the mean) doesn’t matter for entropy, i.e., $h(X + a) = h(X)$ for any constant $a$. 
Moments of Random Vectors

- Consider first order moment $C^{(1)}$ (a vector) and the second order moment $C^{(2)}$ (a matrix). Consider all distributions that have these first and second order moments.

- Out of all these distributions, which one would have the highest (differential) entropy?

- Consider what do moments do? They constrain the possible set of distributions to those having those particular moment values.

- Intuitively: highest entropy should have fewest constraints, or highest entropy would be granted to the distributions with zero additional constraints on higher order moments. “constraint,” meaning being a particular value rather than being derivable from lower order moments.

- Recall also that the first moment (the mean) doesn’t matter for entropy, i.e., $h(X + a) = h(X)$ for any constant $a$.

- In fact, we have:
A Gaussian has the maximum entropy over all distributions that have the same first and second moments. That is let $X \in \mathbb{R}^n$ be a vector random variable with $EX = 0$ and $EXX^\top = K$. Then

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

(18.28)

with equality when $X \sim \mathcal{N}(0, K)$. 

Theorem 18.7.1
proof of Theorem 18.7.1.

- Let \( g(X) \) be such that \( \int g(x)XX^\top dx = K \) (the covariance matrix).
proof of Theorem 18.7.1.

- Let \( g(X) \) be such that \( \int g(x)XX^\top dx = K \) (the covariance matrix).
- Let \( \eta(X) \sim \mathcal{N}(0, K) \) so that \( \int \eta(X)XX^\top dx = K \).
proof of Theorem 18.7.1.

- Let \( g(X) \) be such that \( \int g(x)XX^\top dx = K \) (the covariance matrix).
- Let \( \eta(X) \sim \mathcal{N}(0, K) \) so that \( \int \eta(X)XX^\top dx = K \).
- But \( \log \eta(X) \) has a quadratic form, i.e,

\[
\log \eta(x) = -\frac{1}{2} x^\top K^{-1} x - \frac{1}{2} \ln[(2\pi)^n |K|] \quad (18.29)
\]
proof of Theorem 18.7.1.

- Let $g(X)$ be such that $\int g(x)XX^\top dx = K$ (the covariance matrix).
- Let $\eta(X) \sim \mathcal{N}(0, K)$ so that $\int \eta(X)XX^\top dx = K$.
- But $\log \eta(X)$ has a quadratic form, i.e,

  $$\log \eta(x) = -\frac{1}{2} x^\top K^{-1} x - \frac{1}{2} \ln[(2\pi)^n |K|]$$  \hspace{1cm} (18.29)

- Thus, since $g$ and $\eta$ produce the same results for quadratic forms and by the trace trick, we have

  $$0 \leq D(g||\eta) = \int g(x) \log g(x)/\eta(x) dx = -h(g) - \int g \log \eta$$  \hspace{1cm} (18.30)

  $$= -h(g) - \int \eta \log \eta = -h(g) + h(\eta)$$  \hspace{1cm} (18.31)
proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$. 
proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy.
proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy.

- An instance of a much more general result about maximum entropy.
proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy.

- An instance of a much more general result about maximum entropy.
- Suppose we have a random variable $X$ and a vector of “feature functions” $f(x)$ and consider distributions that satisfy certain constraints $E_p f(X) = \mu$. 
proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy.

- An instance of a much more general result about maximum entropy.
- Suppose we have a random variable $X$ and a vector of “feature functions” $f(x)$ and consider distributions that satisfy certain constraints $E_p f(X) = \mu$.
- If, over all such distributions that satisfy the constraints, we maximize the entropy, we get a distribution of the form:

$$p(x) \propto \exp(\lambda f(x)) \quad (18.32)$$

where $\lambda$ a vector of parameters (Lagrange multipliers).
Differential vs. Discrete Entropy

Gaussians and Maximum Entropy

proof of Theorem 18.7.1.

- Thus, we get $h(g) \leq h(\eta)$.
- Finally, recall that the Gaussian achieves this entropy.

An instance of a much more general result about maximum entropy.

Suppose we have a random variable $X$ and a vector of “feature functions” $f(x)$ and consider distributions that satisfy certain constraints $E_p f(X) = \mu$.

If, over all such distributions that satisfy the constraints, we maximize the entropy, we get a distribution of the form: of the form

$$p(x) \propto \exp(\lambda f(x))$$

(18.32)

where $\lambda$ a vector of parameters (Lagrange multipliers).

- If $f(X) = (X_1, \ldots, X_n, \{X_iX_j\}_{ij})$, we get back the Gaussian.
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions \( p(y|x) \).
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in \mathcal{X}$, $p(y|x)$ models the form of distortion that $x$ undergoes when it is being sent from source to receiver.
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in X$, $p(y|x)$ models the form of distortion that $x$ undergoes when it is being sent from source to receiver.
- Real channels are continuous as are real signals. What really happens to a continuous random variable $X$ is that we have $Y = Z(X)$ where $Z$ is a random function that may or may not be dependent on $X$. 
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in \mathcal{X}$, $p(y|x)$ models the form of distortion that $x$ undergoes when it is being sent from source to receiver.
- Real channels are continuous as are real signals. What really happens to a continuous random variable $X$ is that we have $Y = Z(X)$ where $Z$ is a random function that may or may not be dependent on $X$.
- Quite hard to analyze in general, so we may consider only additive noise $Y = X + Z$ where $Z$ is a random variable.
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in \mathcal{X}$, $p(y|x)$ models the form of distortion that $x$ undergoes when it is being sent from source to receiver.
- Real channels are continuous as are real signals. What really happens to a continuous random variable $X$ is that we have $Y = Z(X)$ where $Z$ is a random function that may or may not be dependent on $X$.
- Quite hard to analyze in general, so we may consider only additive noise $Y = X + Z$ where $Z$ is a random variable.
- We further simplify by saying that $Z \perp \! \! \! \perp X$. 

---

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 18 - Dec 2nd, 2019
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in \mathcal{X}$, $p(y|x)$ models the form of distortion that $x$ undergoes when it is being sent from source to receiver.
- Real channels are continuous as are real signals. What really happens to a continuous random variable $X$ is that we have $Y = Z(X)$ where $Z$ is a random function that may or may not be dependent on $X$.
- Quite hard to analyze in general, so we may consider only additive noise $Y = X + Z$ where $Z$ is a random variable.
- We further simplify by saying that $Z \perp X$
- and moreover that $Z$ is Gaussian, leading to the ...
Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp \perp X_i \), for \( i = 1, \ldots, n \).
Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp X_i \), for \( i = 1, \ldots, n \).

- If \( \sigma^2 = 0 \), what is the capacity of this channel?
Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp X_i \), for \( i = 1, \ldots, n \).

- If \( \sigma^2 = 0 \), what is the capacity of this channel?

- If \( \sigma^2 = 0 \), capacity is infinite since one can perfectly send an arbitrarily precise real number (consider arithmetic coding, it sends a number all within \([0, 1)\)), number of bits per channel use is \( \infty \).
Above is our model, where $Y_i = X_i + Z_i$, with $Z_i \sim N(0, \sigma^2)$ and $Z_i \perp X_i$, for $i = 1, \ldots, n$.

- If $\sigma^2 = 0$, what is the capacity of this channel?
- If $\sigma^2 = 0$, capacity is infinite since one can perfectly send an arbitrarily precise real number (consider arithmetic coding, it sends a number all within $[0, 1)$), number of bits per channel use is $\infty$.
- If $\sigma^2 > 0$ and bounded, what is the capacity?
Gaussian Channel

\[ Z_i \sim N(0, \sigma^2) \]

\[ X_i \rightarrow \oplus \rightarrow Y_i \]

- Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp \perp X_i \), for \( i = 1, \ldots, n \).

- If \( \sigma^2 = 0 \), what is the capacity of this channel?

- If \( \sigma^2 = 0 \), capacity is infinite since one can perfectly send an arbitrarily precise real number (consider arithmetic coding, it sends a number all within \([0, 1)\)), number of bits per channel use is \( \infty \).

- If \( \sigma^2 > 0 \) and bounded, what is the capacity? Then, capacity is still also infinite (unbounded), since we can make input power as large as we want, effectively removing a finite strict subinterval within \([0, 1)\).
Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp \perp X_i \), for \( i = 1, \ldots, n \).

If \( \sigma^2 = 0 \), what is the capacity of this channel?

If \( \sigma^2 = 0 \), capacity is infinite since one can perfectly send an arbitrarily precise real number (consider arithmetic coding, it sends a number all within \([0, 1)\)), number of bits per channel use is \( \infty \).

If \( \sigma^2 > 0 \) and bounded, what is the capacity? Then, capacity is still also infinite (unbounded), since we can make input power as large as we want, effectively removing a finite strict subinterval within \([0, 1)\).

If input power is constrained as well (which is also more practical and realistic 😊), then the problem becomes interesting.
Power constraint

- Average power constraint: for any codeword of length $n$, we require that

\[
\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P
\]  

(18.33)

where $P$ is the average power $\approx EX^2$. 
Power constraint

• Average power constraint: for any codeword of length $n$, we require that

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P$$  \hspace{1cm} (18.33)

where $P$ is the average power $\approx EX^2$.

• This one allows a balance/tradeoff. I.e., we can use one large value if the others are small.
Power constraint

- Average power constraint: for any codeword of length $n$, we require that

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P$$  \hspace{1cm} (18.33)

where $P$ is the average power $\approx EX^2$.

- This one allows a balance/tradeoff. I.e., we can use one large value if the others are small.

- Other possible constraints might include a bound on the maximum absolute value, but this we do not analyze at this time.
Power constraint

- Average power constraint: for any codeword of length $n$, we require that

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P$$  \hspace{1cm} (18.33)

where $P$ is the average power $\approx EX^2$.

- This one allows a balance/tradeoff. I.e., we can use one large value if the others are small.

- Other possible constraints might include a bound on the maximum absolute value, but this we do not analyze at this time.

- Still others include “grouped” constraints (i.e., fix a time window size and bound the maximum of the averages within each window).
Power constraint

- Average power constraint: for any codeword of length $n$, we require that

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P$$  \hspace{1cm} (18.33)

where $P$ is the average power $\approx EX^2$.

- This one allows a balance/tradeoff. I.e., we can use one large value if the others are small.

- Other possible constraints might include a bound on the maximum absolute value, but this we do not analyze at this time.

- Still others include “grouped” constraints (i.e., fix a time window size and bound the maximum of the averages within each window).

- But lets stick with the one above in Equation (18.33).
Error of Simple Example of Gaussian Channel

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but let's analyze it nonetheless).
Error of Simple Example of Gaussian Channel

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).
- $X \in \{+\sqrt{P}, -\sqrt{P}\}$ means that $EX^2 = P$, so this satisfies the constraint. Why not use something with less power?
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but let's analyze it nonetheless).

\( X \in \{ +\sqrt{P}, -\sqrt{P} \} \) means that \( EX^2 = P \), so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as \( +\sqrt{P} \) if \( Y > 0 \) and \( -\sqrt{P} \) if \( Y < 0 \).
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).

$X \in \{+\sqrt{P}, -\sqrt{P}\}$ means that $EX^2 = P$, so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as $+\sqrt{P}$ if $Y > 0$ and $-\sqrt{P}$ if $Y < 0$.

Error:

$\text{(18.35)}$
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).

$X \in \{+\sqrt{P}, -\sqrt{P}\}$ means that $EX^2 = P$, so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as $+\sqrt{P}$ if $Y > 0$ and $-\sqrt{P}$ if $Y < 0$.

Error:

$$P_e$$

(18.35)
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).

* $X \in \{+\sqrt{P}, -\sqrt{P}\}$ means that $EX^2 = P$, so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as $+\sqrt{P}$ if $Y > 0$ and $-\sqrt{P}$ if $Y < 0$.

Error:

$$Pe = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P})$$  \hspace{1cm} (18.34)
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).

\[ X \in \{+\sqrt{P}, -\sqrt{P}\} \] means that \( EX^2 = P \), so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as \(+\sqrt{P}\) if \( Y > 0 \) and \(-\sqrt{P}\) if \( Y < 0 \).

Error:

\[
P_e = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P})
\]

\[
= \frac{1}{2} \Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P}|X = -\sqrt{P})
\]

\[ (18.34) \]

\[ (18.35) \]
Send 1 bit over channel at a time (obviously sub-optimal use of the channel, but lets analyze it nonetheless).

$X \in \{ +\sqrt{P}, -\sqrt{P} \}$ means that $EX^2 = P$, so this satisfies the constraint. Why not use something with less power?

For a uniform source distribution, decode as $+\sqrt{P}$ if $Y > 0$ and $-\sqrt{P}$ if $Y < 0$.

Error:

$$Pe = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P}) \quad (18.34)$$

$$= \frac{1}{2} \Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P}|X = -\sqrt{P})$$

$$= \Pr(Z > \sqrt{P}) \quad (18.35)$$
Simple Example of Gaussian Channel: Error Area

- The two separate error types ($\times 1/2$)

\[
\text{Leads to total error (} \times 1/2 \text{)}
\]

\[
\Pr (Z > \sqrt{P}) = 1 - \Phi(\sqrt{P} \sigma)
\]

where $\Phi$ is cumulative normal distribution, i.e.,

\[
\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt
\]
Simple Example of Gaussian Channel: Error Area

- The two separate error types ($\times 1/2$)

- Leads to total error ($\times 1/2$)

\[
\begin{align*}
\text{Pr}(Z > \sqrt{P}) &= 1 - \Phi(\sqrt{P} \sigma^2) \\
\Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt
\end{align*}
\]
Simple Example of Gaussian Channel: Error Area

- The two separate error types ($\times 1/2$)

\[ \text{0} \quad +\sqrt{P} \quad \text{or} \quad -\sqrt{P} \quad 0 \]

- Leads to total error ($\times 1/2$)

\[ -\sqrt{P} \quad 0 \quad +\sqrt{P} \]

- We have that

\[ \Pr(Z > \sqrt{P}) = 1 - \Phi\left(\frac{\sqrt{P}}{\sigma^2}\right) \tag{18.36} \]

where $\Phi$ is cumulative normal distribution, i.e.,

\[ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{18.37} \]
In fact, we have essentially just turned a Gaussian channel into a discrete BSC:

\[ \begin{array}{c|c|c}
\hline
X & 0 & 1 \\
\hline
0 & 1-p & p \\
1 & p & 1-p \\
\hline
\end{array} \]

\[ \begin{array}{c|c|c}
\hline
Y & 0 & 1 \\
\hline
0 & 1-p & p \\
1 & p & 1-p \\
\hline
\end{array} \]

where \( p = P_e \) for the Gaussian.
In fact, we have essentially just turned a Gaussian channel into a discrete BSC:

\[
\begin{array}{c c c c c}
0 & 1 - p & 0 \\
1 & p & 1 - p & 1 \\
\end{array}
\]

where \( p = P_e \) for the Gaussian.

This will be the common idea. We convert continuous channels into discrete ones with appropriate encodings.
In fact, we have essentially just turned a Gaussian channel into a discrete BSC:

\[
\begin{array}{cc}
0 & 1 - p \\
1 & p \\
\end{array}
\quad
\begin{array}{cc}
0 & 1 - p \\
1 & p \\
\end{array}
\]

where \( p = P_e \) for the Gaussian.

This will be the common idea. We convert continuous channels into discrete ones with appropriate encodings.

This is essentially a process of vector quantization (where under different quantization schemes, we study the tradeoffs that exist when coding). Tradeoffs take the form of rate vs. distortion under the power constraints.
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.
We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x):EX^2 \leq P} I(X;Y) \text{ bits} \quad (18.38)$$
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits}$$

(18.38)

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint \( P \)) is defined to be

\[
C = \max_{p(x): E X^2 \leq P} I(X; Y) \text{ bits}
\]  

(18.38)

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- \( I(X; Y) \) has a nice form in this case, as

\[
(18.40)
\]
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x) : EX^2 \leq P} I(X; Y) \text{ bits} \quad (18.38)$$

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- $I(X; Y)$ has a nice form in this case, as

$$I(X; Y)$$

(18.40)
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x) : \mathbb{E}X^2 \leq P} I(X; Y) \text{ bits} \quad (18.38)$$

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- $I(X; Y)$ has a nice form in this case, as

$$I(X; Y) = h(Y) - h(Y|X) \quad (18.40)$$
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x): E X^2 \leq P} I(X; Y) \text{ bits}$$  \hspace{1cm} (18.38)

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- $I(X; Y)$ has a nice form in this case, as

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X)$$  \hspace{1cm} (18.39)

$$= h(Y) - h(Z|X)$$  \hspace{1cm} (18.40)
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (18.38)$$

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.

- $I(X; Y)$ has a nice form in this case, as

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (18.39)$$

$$= h(Y) - h(Z|X) \quad (18.40)$$
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x): EX^2 \leq P} I(X;Y) \text{ bits} \quad (18.38)$$

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- $I(X;Y)$ has a nice form in this case, as

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X) \quad (18.39)$$

$$= h(Y) - h(Z|X) = h(Y) - h(Z) \quad (18.40)$$
Capacity of Gaussian Channel

- We need a capacity notion, but here under a power constraint.

Definition 18.9.1 (information capacity)

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits}$$  \hspace{1cm} (18.38)

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.

- $I(X; Y)$ has a nice form in this case, as

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X)$$  \hspace{1cm} (18.39)

$$= h(Y) - h(Z|X) = h(Y) - h(Z)$$  \hspace{1cm} (18.40)

- Strategy for computing $C$ same as before: 1) upper bound $I(X; Y)$, and 2) then find a (not nec. unique) $p(x)$ achieving the upper bound.
We need a capacity notion, but here under a power constraint.

**Definition 18.9.1 (information capacity)**

The (information) capacity (with power constraint $P$) is defined to be

$$C = \max_{p(x):EX^2 \leq P} I(X; Y) \text{ bits} \quad (18.38)$$

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
- $I(X; Y)$ has a nice form in this case, as
  $$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (18.39)$$
  $$= h(Y) - h(Z|X) = h(Y) - h(Z) \quad (18.40)$$
- Strategy for computing $C$ same as before: 1) upper bound $I(X; Y)$, and 2) then find a (not nec. unique) $p(x)$ achieving the upper bound.
- Like before, this does **not** mean that we necessary code with this particular $p(x)$, this particular $p(x)$ achieving $C$ may not be unique, nor will it necessarily be a good coding distribution or one we use.