# Class Road Map - IT-I

<table>
<thead>
<tr>
<th>Class</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>9/26</td>
<td>Overview, Communications, Information, Entropy</td>
</tr>
<tr>
<td>L2</td>
<td>10/1</td>
<td>Props. Entropy, Mutual Information,</td>
</tr>
<tr>
<td>L3</td>
<td>10/3</td>
<td>KL-Divergence, Convex, Jensen, and properties.</td>
</tr>
<tr>
<td>L5</td>
<td>10/10</td>
<td>AEP, Compression</td>
</tr>
<tr>
<td>L6</td>
<td>10/15</td>
<td>Compression, Method of Types,</td>
</tr>
<tr>
<td>L7</td>
<td>10/17</td>
<td>Types, U. Coding., Stoc. Processes, Entropy rates,</td>
</tr>
<tr>
<td>L8</td>
<td>10/22</td>
<td>Entropy rates, HMMs, Coding, Kraft,</td>
</tr>
<tr>
<td>L9</td>
<td>10/24</td>
<td>Kraft, Shannon Codes, Huffman, Shannon/Fano/Elias</td>
</tr>
<tr>
<td>L10</td>
<td>10/28</td>
<td>Huffman, Shannon/Fano/Elias</td>
</tr>
<tr>
<td>L11</td>
<td>10/29</td>
<td>Shannon Games,</td>
</tr>
<tr>
<td>LXX</td>
<td>10/31</td>
<td>Midterm, in class.</td>
</tr>
<tr>
<td>L12</td>
<td>11/7</td>
<td>Arith. Coding, Channel Capacity</td>
</tr>
<tr>
<td>L13</td>
<td>11/12</td>
<td>Channel Capacity</td>
</tr>
<tr>
<td>L14</td>
<td>11/14</td>
<td>Channel Capacity, Shannon’s 2nd thm</td>
</tr>
<tr>
<td>L15</td>
<td>11/19</td>
<td>Shannon’s 2nd thm, zero error codes, feedback</td>
</tr>
<tr>
<td>L16</td>
<td>11/21</td>
<td>Joint thm, coding, hamming,</td>
</tr>
<tr>
<td>L17</td>
<td>11/26</td>
<td>hamming, diff. entropy</td>
</tr>
<tr>
<td>L18</td>
<td>12/3</td>
<td>diff. entropy</td>
</tr>
<tr>
<td>LXX</td>
<td>12/10</td>
<td>Final exam, 4:30pm</td>
</tr>
</tbody>
</table>

Finals Week: December 12th–16th.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Chapter 3 in our book (Cover & Thomas, “Information Theory”).
- Section 11.1 (method of types).
- Chapter 4 and 5 in our book (Cover & Thomas, “Information Theory”)
- Read stream code chapter 6 in “Information Theory, Inference, and Learning Algorithms” by David J.C. MacKay (available online http://www.inference.phy.cam.ac.uk/mackay/itila/)
- Read Chapter 7 in our book (Cover & Thomas, “Information Theory”).
- Read Chapter 8 in our book (Cover & Thomas, “Information Theory”).
Homework

Homework 7 on our web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/), due Sunday night, at 11:45pm.
Email me if you want to skype/google hangout over the next week and we can arrange a time.

Final assignment now online, need to upload pdf scan we send you by Tuesday Dec 11th, at 4:00pm.

Final exam time/location: Tuesday, December 10, 2013, 4:30-6:20pm, PCAR 297
Continuous/Differential Entropy

- Let $X$ now be a continuous r.v. with cumulative distribution
  \[ F(x) = \Pr(X \leq x) \]  
  (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 $h(X)$)**

\[ h(X) = - \int_S f(x) \log f(x) \, dx \]  
(18.6)

- Since we integrate over only the support set, no worries about $\log 0$.
- Perhaps it is best to do some examples.
Continuous Entropy Of Uniform Distribution

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

$$h(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = - \log \frac{1}{a}$$  \hspace{1cm} (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 “volume” of a typical set.
- Example: $2^{H(X)}$ is the number of things that happen, on average, and can have $2^{H(X)} \ll |\mathcal{X}|$.
- Consider a uniform r.v. $Y$ such that $2^{H(X)} = |\mathcal{Y}|$.
- Thus having a negative exponent just means the volume is small.
Continuous Entropy Of Normal Distribution

- Normal (Gaussian) distributions are very important.
- We have:
  \[ X \sim N(0, \sigma^2) \Leftrightarrow f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}x^2/\sigma^2} \]  
  (18.5)
- Let's compute the entropy of \( f \) in nats.
  \[ h(X) = -\int f \ln f = -\int f(x) \left[ -\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \]  
  (18.6)
  \[ \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \]  
  (18.7)
  \[ = \frac{1}{2} \ln e + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} \ln(2\pi e\sigma^2) \text{nats} \times \left( \frac{1}{\ln 2} \text{bits/nats} \right) \]
  \[ = \frac{1}{2} \log(2\pi e\sigma^2) \text{bits} \]  
  (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**

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

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

This follows since (as expected)

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

(18.18)

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

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

(18.19)

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

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

(18.20)

This means that as $n \to \infty$, $H(X^\Delta)$ can get larger. Why?
Like discrete case, we have entropy for vectors of r.v.s
Like discrete case, we have entropy for vectors of r.v.s

The joint differential entropy is defined as:

$$h(X_1, X_2, \ldots, X_n) = -\int f(x_{1:n}) \log f(x_{1:n}) dx_{1:n} \quad (18.1)$$
Joint Differential Entropy

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

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

(18.1)

- Conditional differential entropy

\[ h(X|Y) = - \int f(x, y) \log f(x|y) \, dx \, dy = h(X, Y) - h(Y) \]  

(18.2)
When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$  \hfill (18.3)

Notice that the entropy is monotonically related to the determinant of the covariance matrix $\Sigma$ and is not at all dependent on the mean $\mu$. Could solve for $|\Sigma|$ as a function of entropy, get $\propto 2^{h(X)}$. The determinant is a form of spread, or dispersion of the distribution.
When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)}$$  \hspace{1cm} (18.3)$$

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} (18.4)
Entropy of a Multivariate Gaussian

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

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)} \quad (18.3)$$

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

$$h(X) = \frac{1}{2} \log \left( (2\pi e)^n |\Sigma| \right) \text{ bits} \quad (18.4)$$

- 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.,

$$X \sim \mathcal{N} (\mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu)}$$  \hspace{1cm} (18.3)$$

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} (18.4)$$

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

Could solve for $|\Sigma|$ as a function of entropy, get $\propto 2^{h(X)}$. 

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

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi \Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)}$$  \hfill (18.3)

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

$$h(X) = \frac{1}{2} \log \left[ (2\pi e)^n |\Sigma| \right] \text{ bits}$$  \hfill (18.4)

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

Could solve for $|\Sigma|$ as a function of entropy, get $\propto 2^{h(X)}$.

The determinant is a form of spread, or dispersion of the distribution.
Entropy of a Multivariate Gaussian: Derivation

\[ h(X) \]

\[(18.11) \]
Entropy of a Multivariate Gaussian: Derivation

\[ h(X) = - \int f(x) \left[ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) - \ln \left( (2\pi)^{n/2} |\Sigma|^{1/2} \right) \right] \]  
(18.5)

\[ = \frac{1}{2} \text{tr} \left( EE \right) + \frac{1}{2} \ln \left( (2\pi)^n |\Sigma| \right) \]  
(18.10)

\[ = \frac{n}{2} + \frac{1}{2} \ln \left( (2\pi e)^n |\Sigma| \right) \]  
(18.11)
Entropy of a Multivariate Gaussian: Derivation

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

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

(18.5)

(18.6)

(18.11)
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

\[ \alpha = \text{tr}(\alpha) \]

\[ \text{tr} (ABC) = \text{tr} (BCA) \]

(18.11)


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

\[(18.5)\]

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

\[(18.6)\]

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

\[(18.7)\]

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

\[(18.8)\]

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

\[(18.10)\]

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

\[(18.11)\]
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

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

(18.8)

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

(18.9)

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

(18.11)
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

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

(18.8)

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

(18.9)

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

(18.10)

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

(18.11)
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

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

(18.8)

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

(18.9)

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

(18.10)

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

(18.11)

\[ = \frac{n}{2} + \frac{1}{2} \ln \left[ (2\pi)^n |\Sigma| \right] \]
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

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

(18.8)

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

(18.9)

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

(18.10)

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

(18.11)
Entropy of a Multivariate Gaussian: Derivation

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

(18.5)

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

(18.6)

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

(18.7)

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

(18.8)

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

(18.9)

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

(18.10)

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

(18.11)

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

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

(18.12)

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

Mutual Information:

\[ I(X;Y) = h(X) - h(X|Y) \]  

(18.13)

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

(18.14)

Thus, since \( I(X;Y) \geq 0 \) we have again that conditioning reduces entropy, i.e., \( h(Y) \geq h(Y|X) \).
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx \geq 0$$  \hspace{1cm} (18.12)\

We can, like in the discrete case, use Jensen’s inequality to prove the non-negativity of $D(f\|g)$. 
Relative Entropy/KL-Divergence & Mutual Information

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

\[ D(f \| g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \quad (18.12) \]

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

- Mutual Information:

\[
D(f(X,Y) \| f(X)f(Y)) = I(X;Y) = h(X) - h(X|Y) \quad (18.13)
= h(Y) - h(Y|X) \geq 0 \quad (18.14)
\]
The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

\[ D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \quad (18.12) \]

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

Mutual Information:

\[ D(f(X,Y)||f(X)f(Y)) = I(X;Y) = h(X) - h(X|Y) \quad (18.13) \]
\[ = h(Y) - h(Y|X) \geq 0 \quad (18.14) \]

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

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

(18.15)
Chain rules and more

- We still have chain rules
  \[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  
  (18.15)

- And bounds of the form
  \[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \]  
  (18.16)

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
\[ f(A) = H(X_A) \]  
(18.17)
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.

Similarly, when some variables independent, adding independent variables with negative entropy can decrease overall entropy.
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} (18.15)

- 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} (18.16)

- 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
  \[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i | X_{1:i-1}) \]  
  (18.15)

- And bounds of the form
  \[ \sum_i h(X_i | X_{1:n \setminus \{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \]  
  (18.16)

- 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
  \[ f(A) = H(X_A) \]
  (18.17)

  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} (18.15)

- 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} (18.16)

- 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} (18.17)

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

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

(18.15)

- And bounds of the form

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

(18.16)

- 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

\[ f(A) = H(X_A) \]  

(18.17)

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.
Chain rules and more

- We still have chain rules

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

(18.15)

- And bounds of the form

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

(18.16)

- 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

\[ f(A) = H(X_A) \]  

(18.17)

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

\[ \log |K| \leq n \prod_{i=1}^{n} K_{ii} \] (18.18)

whenever \( K \) is positive semi-definite (a result known as Hadamard’s inequality).
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:

$$|K| \leq n \prod_{i=1}^{n} K_{ii}$$

whenever $K$ is positive semi-definite (a result known as Hadamard’s inequality).
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:

\[
|K| \leq \prod_{i=1}^{n} K_{ii}
\] (18.18)

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 $\mu = E[X]$. 
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 $C_{ij} = E[X_iX_j]$. 

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 $C_{ijkl} = E[X_iX_jX_kX_\ell]$, 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 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.
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.
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.
- It can be shown, moreover, that the multivariate Gaussian only has first and 2nd order moments, but all higher order moments are zero (in fact, the Gaussian can be parameterized exactly via its first and second order moments).
Now first the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (a matrix) consider all distributions that have these first and second order moments.

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

Consider what do moments do? They further constrain the possible set of distributions once their value is set.

Perhaps it then makes sense to intuitively consider that the highest entropy would be granted to the distributions with zero higher 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:
Moments of Random Vectors

- Now first the first and second order moments $C'(1)$ (a vector) and $C'(2)$ (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

- Now first the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (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 further constrain the possible set of distributions once their value is set.
Moments of Random Vectors

- Now first the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (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 further constrain the possible set of distributions once their value is set.
- Perhaps it then makes sense to intuitively consider that the highest entropy would be granted to the distributions with zero higher order moments.
Moments of Random Vectors

Now first the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (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 further constrain the possible set of distributions once their value is set.

Perhaps it then makes sense to intuitively consider that the highest entropy would be granted to the distributions with zero higher 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

Now first the first and second order moments $C^{(1)}$ (a vector) and $C^{(2)}$ (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 further constrain the possible set of distributions once their value is set.

Perhaps it then makes sense to intuitively consider that the highest entropy would be granted to the distributions with zero higher 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:
Theorem 18.3.1

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 $E X = 0$ and $E X X^\top = K$. Then

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

(18.19)

with equality when $X \sim \mathcal{N}(0, K)$. 
proof of Theorem 18.3.1.

- Let $g(X)$ be such that $\int g(x)XX^T dx = K$ (the covariance matrix).
proof of Theorem 18.3.1.

- 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 18.3.1.

- 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,

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

(18.20)
proof of Theorem 18.3.1.

- 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,

\[
\int g(x) \log \eta(x) = -\frac{1}{2} x^\top K^{-1} x - \frac{1}{2} \ln[(2\pi)^n|K|] \]  

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

\[
0 \leq D(g||\eta) = \int g(x) \log g(x)/\eta(x) dx = -h(g) - \int g \log \eta \]  

\[
= -h(g) - \int \eta \log \eta = -h(g) + h(\eta) \]
proof of Theorem 18.3.1.

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: 

$$p(x) \propto \exp(\lambda f(x)) \quad (18.23)$$

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

If $f(X) = (X_1, \ldots, X_n, \{X_i X_j\}_{ij})$, we get back the Gaussian.
proof of Theorem 18.3.1.

- Thus, we get \( h(g) \leq h(\eta) \).
- Finally, recall that the Gaussian achieves this entropy.
proof of Theorem 18.3.1.

- 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 18.3.1.

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 18.3.1.

- 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

\[
p(x) \propto \exp(\lambda f(x))
\]  

(18.23)

where \( \lambda \) a vector of parameters (Lagrange multipliers).
proof of Theorem 18.3.1.

- 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:

\[
p(x) \propto \exp(\lambda f(x))
\]

(18.23)

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 $p(y|x)$. 

---

Prof. Jeff Bilmes
EE514a/Fall 2013/Information Theory I – Lecture 18 - Dec 3rd, 2013
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.
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 $Y = \nu(X)$ where $\nu$ 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 $x \in X$, $p(y|x)$ 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 $Y = \nu(X)$ where $\nu$ is a random function that may or may not be dependent on $X$.
- This is quite hard to analyze so we may consider only additive noise $Y = X + \nu$ where $\nu$ 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 $Y = \nu(X)$ where $\nu$ is a random function that may or may not be dependent on $X$.
- This is quite hard to analyze so we may consider only additive noise $Y = X + \nu$ where $\nu$ is a random variable.
- We further simplify by saying that $\nu \perp\!\!\!\perp X$
Continuous Channels

- So far, we have considered discrete channels which are modeled by conditional probability distributions $p(y|x)$.
- That is, for a given $x \in X$, $p(y|x)$ 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 $Y = \nu(X)$ where $\nu$ is a random function that may or may not be dependent on $X$.
- This is quite hard to analyze so we may consider only additive noise $Y = X + \nu$ where $\nu$ is a random variable.
- We further simplify by saying that $\nu \perp \perp X$
- and moreover that $\nu$ 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 $Z_i \perp \perp X_i$. 
Gaussian Channel

\[ \text{Above is our model, where } Y_i = X_i + Z_i, \text{ with } Z_i \sim N(0, \sigma^2) \text{ and } Z_i \perp X_i. \]

\[ \text{If } \sigma^2 = 0, \text{ what is the capacity of this channel?} \]
Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim N(0, \sigma^2) \) and \( Z_i \perp X_i \).

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 \([0, 1)\)).
Above is our model, where $Y_i = X_i + Z_i$, with $Z_i \sim N(0, \sigma^2)$ and $Z_i \perp \perp X_i$.

- 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 $[0, 1)$).
- If $\sigma^2 > 0$, 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 \( Z_i \perp X_i \).
- 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 \([0, 1)\)).
- If \( \sigma^2 > 0 \), what is the capacity?
- If \( \sigma^2 > 0 \), capacity is still infinite, 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 $Z_i \perp X_i$.

- 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 $[0, 1)$).
- If $\sigma^2 > 0$, what is the capacity?
- If $\sigma^2 > 0$, capacity is still infinite, 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.
Average power constraint: for any codeword of length $n$, we require that

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

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

This one allows a balance/tradeoff. I.e., we can use one large value other 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 window size and bound the maximum of the averages within each window). But let's stick with the one above in Equation (18.24).
Power constraint

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

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

(18.24)

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

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

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

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

(18.24)

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

- This one allows a balance/tradeoff. I.e., we can use one large value of 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

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

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

- This one allows a balance/tradeoff. I.e., we can use one large value of 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 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} (18.24)

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

- This one allows a balance/tradeoff. I.e., we can use one large value of 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 window size and bound the maximum of the averages within each window).

- But lets stick with the one above in Equation (18.24).
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
- \( X \in \{+\sqrt{P}, -\sqrt{P}\} \) means that \( EX^2 = P \), so this satisfies the constraint.
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).

- \( X \in \{ +\sqrt{P}, -\sqrt{P} \} \) means that \( EX^2 = P \), so this satisfies the constraint.

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

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
- \( X \in \{+\sqrt{P}, -\sqrt{P}\} \) means that \( EX^2 = P \), so this satisfies the constraint.
- For a uniform source distribution, decode as \(+\sqrt{P}\) if \( Y > 0 \) and \(-\sqrt{P}\) if \( Y < 0 \).
- Error:

\[
(18.26)
\]
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
- $X \in \{+\sqrt{P}, -\sqrt{P}\}$ means that $E X^2 = P$, so this satisfies the constraint.
- For a uniform source distribution, decode as $+\sqrt{P}$ if $Y > 0$ and $-\sqrt{P}$ if $Y < 0$.
- Error:

$$P_e$$

(18.26)
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
- \( X \in \{ +\sqrt{P}, -\sqrt{P} \} \) means that \( EX^2 = P \), so this satisfies the constraint.
- For a uniform source distribution, decode as \( +\sqrt{P} \) if \( Y > 0 \) and \( -\sqrt{P} \) if \( Y < 0 \).
- Error:

\[
P_e = \frac{1}{2} \Pr(Y < 0 | X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0 | X = -\sqrt{P}) \tag{18.25}
\]

(18.26)
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).
- \( X \in \{+\sqrt{P}, -\sqrt{P}\} \) means that \( EX^2 = P \), so this satisfies the constraint.
- For a uniform source distribution, decode as \(+\sqrt{P}\) if \( Y > 0 \) and \(-\sqrt{P}\) if \( Y < 0 \).
- Error:

\[
P_e = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P}) \quad (18.25)
\]
\[
= \frac{1}{2} \Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P}|X = -\sqrt{P})
\]
\[
= \Pr(Z > \sqrt{P}) \quad (18.26)
\]
Example

- Send 1 bit over channel at a time (obviously sub-optimal use of the channel).

  \[ X \in \left\{ +\sqrt{P}, -\sqrt{P} \right\} \]
  means that \( EX^2 = P \), so this satisfies the constraint.

- 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 \mid X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0 \mid X = -\sqrt{P}) \quad (18.25)
  
  = \frac{1}{2} \Pr(Z < -\sqrt{P} \mid X = +\sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P} \mid X = -\sqrt{P})
  
  = \Pr(Z > \sqrt{P}) \quad (18.26)
  \]
Example

- The two separate error types

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

- The two separate error types

\[ +\sqrt{P_0} \quad \text{or} \quad -\sqrt{P_0} \]

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

- The two separate error types

\[ \begin{align*}
0 &+\sqrt{P} \\
\text{or} \\
-\sqrt{P} &0
\end{align*} \]

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

\[ \begin{align*}
\sqrt{P} &0 \\
0 &+\sqrt{P}
\end{align*} \]

- We have that

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

(18.27)

where \(\Phi\) is cumulative normal distribution, i.e.,

\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt
\]

(18.28)
Example

In fact, we have essentially just turned a Gaussian channel into a discrete BSC:

\[
\begin{align*}
X & \quad Y \\
0 & \quad 0 \\
1 & \quad 1 \\
\end{align*}
\]

\[
\begin{align*}
& 1 - p \\
p & \quad p \\
& 1 - p \\
\end{align*}
\]

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

![Diagram of a discrete BSC with transition probabilities](image)

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:

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

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

We get

\[ C = \max_p \mathbb{E} X^2 \leq P \]

\[ I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X) \]

\[ = h(Y) - h(Z|X) = h(Y) - h(Z) \]
We need a capacity notion, but here under a power constraint.

We get

**Definition 18.4.1**

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

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

(18.29)
Capacity of Gaussian Channel

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

**Definition 18.4.1**

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

$$C = \max_{p(x):EX^2 \leq P} I(X;Y) \text{ bits}$$ (18.29)

- Like in discrete case, have not (yet) said anything about transmission rate and/or if we can communicate at that rate.
We need a capacity notion, but here under a power constraint.

We get

**Definition 18.4.1**

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

$$C = \max_{p(x):EX^2 \leq P} I(X;Y) \text{ bits}$$ (18.29)

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

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z|X)$$ (18.31)
We need a capacity notion, but here under a power constraint. We get

**Definition 18.4.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits}$$  \hspace{1cm} (18.29)

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

$$I(X; Y)$$  \hspace{1cm} (18.31)
Capacity of Gaussian Channel

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

**Definition 18.4.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (18.29)$$

- 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

$$I(X; Y) = h(Y) - h(Y|X)$$

(18.31)
We need a capacity notion, but here under a power constraint.

We get

**Definition 18.4.1**

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

$$C = \max_{p(x) : E X^2 \leq P} I(X; Y) \text{ bits} \quad (18.29)$$

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

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (18.30)$$

$$\quad (18.31)$$
We need a capacity notion, but here under a power constraint.

We get

**Definition 18.4.1**

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

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

(18.29)

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

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

(18.30)

$$= h(Y) - h(Z|X)$$

(18.31)
We need a capacity notion, but here under a power constraint.

We get

Definition 18.4.1

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (18.29)$$

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

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (18.30)$$

$$= h(Y) - h(Z|X) = h(Y) - h(Z) \quad (18.31)$$
But since $Z$ is Gaussian, $h(Z) = \frac{1}{2} \log(2\pi e\sigma^2)$ where $\sigma^2$ is the noise power, $EZ^2 = \sigma^2 = N$, with $EZ = 0$. 

$Z_i \sim N(0, \sigma^2)$

$X_i \longrightarrow \oplus \longrightarrow Y_i$
Capacity of Gaussian Channel

- But since $Z$ is Gaussian, $h(Z) = \frac{1}{2} \log(2\pi e \sigma^2)$ where $\sigma^2$ is the noise power, $EZ^2 = \sigma^2 = N$, with $EZ = 0$.

- We also saw earlier, since Gaussians have maximum entropy for a given 2nd moment, that if $EX = 0$, $\text{var}(X) = K$, then

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

(18.32)
Capacity of Gaussian Channel

- But since $Z$ is Gaussian, $h(Z) = \frac{1}{2} \log(2\pi e \sigma^2)$ where $\sigma^2$ is the noise power, $EZ^2 = \sigma^2 = N$, with $EZ = 0$.

- We also saw earlier, since Gaussians have maximum entropy for a given 2nd moment, that if $EX = 0$, $\text{var}(X) = K$, then

$$h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \quad (18.32)$$

- Also,

$$EY^2 \quad (18.34)$$
Capacity of Gaussian Channel

But since $Z$ is Gaussian, $h(Z) = \frac{1}{2} \log(2\pi e\sigma^2)$

where $\sigma^2$ is the noise power, $EZ^2 = \sigma^2 = N$, with $EZ = 0$.

We also saw earlier, since Gaussians have maximum entropy for a given 2nd moment, that if $EX = 0$, $\text{var}(X) = K$, then

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

Also,

$$EY^2 = E(X + Z)^2$$
Capacity of Gaussian Channel

But since \( Z \) is Gaussian, \( h(Z) = \frac{1}{2} \log(2\pi e \sigma^2) \)
where \( \sigma^2 \) is the noise power, \( EZ^2 = \sigma^2 = N \),
with \( EZ = 0 \).

We also saw earlier, since Gaussians have maximum entropy for a
given 2nd moment, that if \( EX = 0 \), \( \text{var}(X) = K \), then

\[
h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \tag{18.32}
\]

Also,

\[
EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2
\text{ since } X \perp Z
\]

\[
\tag{18.33}
\]

\[
\tag{18.34}
\]
But since $Z$ is Gaussian, $h(Z) = \frac{1}{2} \log(2\pi e \sigma^2)$ where $\sigma^2$ is the noise power, $EZ^2 = \sigma^2 = N$, with $EZ = 0$.

We also saw earlier, since Gaussians have maximum entropy for a given 2nd moment, that if $EX = 0$, $\text{var}(X) = K$, then

$$h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \quad (18.32)$$

Also,

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2 \quad \text{since } X \perp \perp Z$$

$$= P + \sigma^2$$

(18.34)
Thus, we can upper bound the mutual information as follows:

\[ I(X; Y) \] 

(18.36)

where SNR is the signal to noise ratio.
Thus, we can upper bound the mutual information as follows:

\[ I(X;Y) = h(Y) - h(Z) \]

(18.36)
Thus, we can upper bound the mutual information as follows:

\[ I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2) \]

\[ (18.35) \]

\[ (18.36) \]
Thus, we can upper bound the mutual information as follows:

\[ I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2) \]

(18.35)

\[ = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) \]

(18.36)
Thus, we can upper bound the mutual information as follows:

\[
I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2) 
\]

\[
= \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR})
\]
Thus, we can upper bound the mutual information as follows:

\[
I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2) 
\]

(18.35)

\[
= \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) 
\]

(18.36)

where SNR is the signal to noise ratio.
Thus, we can upper bound the mutual information as follows:

$$I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2)$$

(18.35)

$$= \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR})$$

(18.36)

where SNR is the signal to noise ratio.

We can achieve equality in the bound on $h(Y)$ by ensuring $Y$ is Gaussian, and this is the case if $X$ is Gaussian (sums of Gaussians are Gaussian).
Thus, we can upper bound the mutual information as follows:

\[ I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log(2\pi e(P + \sigma^2)) - \frac{1}{2} \log(2\pi e\sigma^2) \]

\[ = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) = \frac{1}{2} \log(1 + \text{SNR}) \]

where SNR is the signal to noise ratio.

We can achieve equality in the bound on \( h(Y) \) by ensuring \( Y \) is Gaussian, and this is the case if \( X \) is Gaussian (sums of Gaussians are Gaussian).

The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) = \frac{1}{2} \log(1 + \text{SNR}) \]
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  

(18.38)
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) = \frac{1}{2} \log(1 + \text{SNR}) \]  

(18.38)

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  

(18.38)

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).
- Rate depends on SNR - if signal level is allowed to be much larger than noise, then rate should increase (log when information measured in bits).
Capacity of Gaussian Channel

The capacity of the Gaussian channel is

\[
C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \tag{18.38}
\]

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).
- Rate depends on SNR - if signal level is allowed to be much larger than noise, then rate should increase (log when information measured in bits).
- In fact, from this we get the standard 6.02dB SNR/bit for audio. i.e., \( 16 = 1/2 \log(1 + \text{SNR}) \), or \( 2^{32} = 1 + \text{SNR} \) or \( \text{SNR} = 2^{32} - 1 \).
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  

(18.38)

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).

- Rate depends on SNR - if signal level is allowed to be much larger than noise, then rate should increase (log when information measured in bits).

- In fact, from this we get the standard 6.02dB SNR/bit for audio. I.e., \( 16 = \frac{1}{2} \log(1 + \text{SNR}) \), or \( 2^{32} = 1 + \text{SNR} \) or \( \text{SNR} = 2^{32} - 1 \).

- \( 10 \log_{10}(\text{SNR}) = 10 \times 32 \log_{10}(2) \approx 96.33 \text{dB} \).
Capacity of Gaussian Channel

The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (18.38)

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).
- Rate depends on SNR - if signal level is allowed to be much larger than noise, then rate should increase (log when information measured in bits).
- In fact, from this we get the standard 6.02dB SNR/bit for audio. I.e., \( 16 = \frac{1}{2} \log(1 + \text{SNR}) \), or \( 2^{32} = 1 + \text{SNR} \) or \( \text{SNR} = 2^{32} - 1 \).
- \( 10 \log_{10}(\text{SNR}) = 10 \times 32 \log_{10}(2) \approx 96.33\text{dB} \).
- And \( 96.33/16 \approx 6.02\text{dB}/\text{bit} \).
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \quad (18.38) \]

- Makes sense: the maximum transmission rate obtained when \( X \sim \mathcal{N}(0, P) \).

- Rate depends on SNR - if signal level is allowed to be much larger than noise, then rate should increase (log when information measured in bits).

- In fact, from this we get the standard 6.02dB SNR/bit for audio. I.e., \( 16 = \frac{1}{2} \log(1 + \text{SNR}) \), or \( 2^{32} = 1 + \text{SNR} \) or \( \text{SNR} = 2^{32} - 1 \).

- \( 10 \log_{10}(\text{SNR}) = 10 \times 32 \log_{10}(2) \approx 96.33 \text{dB} \).

- And \( 96.33/16 \approx 6.02 \text{dB/bit} \).

- Every additional bit (on an audio CD) adds 6.02 dB of SNR.
An \((M, n)\) code for the Gaussian channel, with power constraint \(P\), includes

1. index set \(\{1, 2, \ldots, M\}\)

2. Encoding function \(X : \{1, \ldots, M\} \rightarrow \mathcal{X}^n\) giving codewords \(X^n(1), X^n(2), \ldots, X^n(M)\) with

\[
\frac{1}{n} \sum_{i=1}^{n} X_i^2(\omega) \leq P \quad \forall \omega \in \{1, \ldots, M\} \quad (18.39)
\]

3. Decoding function

\[g : \mathcal{Y}^n \rightarrow \{1, \ldots, M\} \quad (18.40)\]
Definition 18.4.3

The rate is

\[ R = \frac{\log M}{n} \]  

bits per channel use  

(18.41)