Class Road Map - IT-I

L19 (1/6): Overview, Communications, Gaussian Channel
L20 (1/8):
L21 (1/13):
L22 (1/15):
– (1/20): Monday holiday
L23 (1/22):
L24 (1/27):
L25 (1/29):
L26 (2/3):
L27 (2/5):

L28 (2/10):
L29 (2/12):
– (2/17): Monday, Holiday
L30 (2/19):
L31 (2/24):
L32 (2/26):
L33 (3/3):
L34 (3/5):
L35 (3/10):
L36 (3/12):

This two-quarter course will be a thorough introduction to information theory.
Information Theory I and II

- This two-quarter course will be a thorough introduction to information theory.
- Information Theory I (EE514, Fall 2013): entropy, mutual information, asymptotic equipartition properties, data compression to the entropy limit (source coding theorem), Huffman, communication at the channel capacity limit (channel coding theorem), method of types, arithmetic coding, Fano codes, differential entropy, maximum entropy.

- Information Theory II (EE515, Winter 2014): Gaussian channel, more on compression (Lempel-Ziv, Kolmogorov complexity), spectral estimation, communication with bounded distortion and rate-distortion theory (including alternating minimization for computation of RD curve and channel capacity), some coding (e.g., convolutional codes, ECC, turbo, LDPC), non-Shannon information inequalities and properties of entropy and information measures, network information theory, information geometry, and some recent results on use of polymatroids in information theory.

Additional topics throughout will include information theory as it is applicable to pattern recognition, natural language processing, computer science and complexity, biological science, and communications.
Information Theory I and II

- This two-quarter course will be a thorough introduction to information theory.
- Information Theory I (EE514, Fall 2013): entropy, mutual information, asymptotic equipartition properties, data compression to the entropy limit (source coding theorem), Huffman, communication at the channel capacity limit (channel coding theorem), method of types, arithmetic coding, Fano codes, differential entropy, maximum entropy.
- Information Theory II (EE515, Winter 2014): Gaussian channel, more on compression (Lempel-Ziv, Kolmogorov complexity), spectral estimation, communication with bounded distortion and rate-distortion theory (including alternating minimization for computation of RD curve and channel capacity), some coding (e.g., convolutional codes, ECC, turbo, LDPC), non-Shannon information inequalities and properties of entropy and information measures, network information theory, information geometry, and some recent results on use of polymatroids in information theory.
This two-quarter course will be an thorough introduction to information theory.

Information Theory I (EE514, Fall 2013): entropy, mutual information, asymptotic equipartition properties, data compression to the entropy limit (source coding theorem), Huffman, communication at the channel capacity limit (channel coding theorem), method of types, arithmetic coding, Fano codes, differential entropy, maximum entropy.

Information Theory II (EE515, Winter 2014): Gaussian channel, more on compression (Lempel-Ziv, Kolmogorov complexity), spectral estimation, communication with bounded distortion and rate-distortion theory (including alternating minimization for computation of RD curve and channel capacity), some coding (e.g., convolutional codes, ECC, turbo, LDPC), non-Shannon information inequalities and properties of entropy and information measures, network information theory, information geometry, and some recent results on use of polymatroids in information theory.
Course Web Pages

- Class, Monday & Wednesday, 1:30-3:20, in our room (Loew Hall Room 102).

EE515: our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/) our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.

our discussion board (https://canvas.uw.edu/courses/880971/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.

Previous webpage, EE514a, fall 2013, is at our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/). Slides are available and lectures are on youtube.
Course Web Pages

- Class, Monday & Wednesday, 1:30-3:20, in our room (Loew Hall Room 102).
- EE515: our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/)

Assignment dropbox (https://canvas.uw.edu/courses/880971/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.

Discussion board (https://canvas.uw.edu/courses/880971/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.

Previous web page, EE514a, fall 2013, is at our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/). Slides are available and lectures are on youtube.

Prof. Jeff Bilmes
Course Web Pages

- Class, Monday & Wednesday, 1:30-3:20, in our room (Loew Hall Room 102).
- EE515: our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/)
- our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.
Course Web Pages

- Class, Monday & Wednesday, 1:30-3:20, in our room (Loew Hall Room 102).
- EE515: our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/)
- our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.
- our discussion board (https://canvas.uw.edu/courses/880971/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.
Course Web Pages

- Class, Monday & Wednesday, 1:30-3:20, in our room (Loew Hall Room 102).
- EE515: our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/)
- our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.
- our discussion board (https://canvas.uw.edu/courses/880971/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.
- Previous web page, EE514a, fall 2013, is at our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/). Slides are available and lectures are on youtube.
Prerequisites

- basic probability and statistics
- random processes (e.g., EE505 or a Stat 5xx class).
- Knowledge of matlab.
- EE514A - Information Theory I (or equivalent).
- The course is open to students in all UW departments.
There will be about 2-4 homeworks this quarter.
You will have approximately 1 to 2 weeks to solve the problem set.
Problem sets might also include matlab exercises, so you will need to have access to matlab, please contact the EE support staff (help@ee.washington.edu) if you currently do not have matlab access, and they should be able to help you.
The problem sets that are longer will take longer to do, so please do not wait until the night before they are due to start them.
Homework Grading Strategy

- We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.

You will receive a grade on your homework from the person(s) who is (are) grading it.

You will also assign a grade to your grader(s), for the quality of the grading (but not the grade) given to you.

This means, for each HW you will get two grades: one for your HW, and one for the quality of grading you do.

No collusion!

No grading your own HW.
Homework Grading Strategy

- We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.

- You will receive a grade on your homework from the person(s) who is (are) grading it.
We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.

You will receive a grade on your homework from the person(s) who is (are) grading it.

You will also assign a grade to your grader(s), for the quality of the grading (but not the grade) given to you.
Homework Grading Strategy

- We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.
- You will receive a grade on your homework from the person(s) who is (are) grading it.
- You will also assign a grade to your grader(s), for the quality of the grading (but not the grade) given to you.
- This means, for each HW you will get two grades: one for your HW, and one for the quality of grading you do.
Homework Grading Strategy

- We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.
- You will receive a grade on your homework from the person(s) who is (are) grading it.
- You will also assign a grade to your grader(s), for the quality of the grading (but not the grade) given to you.
- This means, for each HW you will get two grades: one for your HW, and one for the quality of grading you do.
- No collusion! 😊
Homework Grading Strategy

- We will do swap grading. I.e. you will turn in your homeworks, and I’ll send them out randomly for others to grade.
- You will receive a grade on your homework from the person(s) who is (are) grading it.
- You will also assign a grade to your grader(s), for the quality of the grading (but not the grade) given to you.
- This means, for each HW you will get two grades: one for your HW, and one for the quality of grading you do.
- No collusion! 😊
- No grading your own HW. 😊
Exams

- No exams this quarter.
Final grades will be based on a combination of the homeworks (50%) and the final presentations (50%).

Final presentations will be graded based both on your final presentation talk (how clear it is, etc.) and also a final 4-page (max) writeup.

If you are active in the class (via attendance, asking good questions, participating often in online canvas discussions, etc.), this may boost your final grade.
On Final Presentations

Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
- You also must turn in a 4-page (max) conference style writeup.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
- You also must turn in a 4-page (max) conference style writeup.
- The papers must not be ones that we covered in class, although they can be related.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
- You also must turn in a 4-page (max) conference style writeup.
- The papers must not be ones that we covered in class, although they can be related.
- You need to do your own research to find the papers yourself (i.e., that is part of the assignment).
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
- You also must turn in a 4-page (max) conference style writeup.
- The papers must not be ones that we covered in class, although they can be related.
- You need to do your own research to find the papers yourself (i.e., that is part of the assignment).
- The majority of the papers must have been published in the last 10 years (so no old or classic papers).
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.
- You also must turn in a 4-page (max) conference style writeup.
- The papers must not be ones that we covered in class, although they can be related.
- You need to do your own research to find the papers yourself (i.e., that is part of the assignment).
- The majority of the papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is, and also turning in milestone assignments.
On Final Presentations

Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory or a very related area.

You also must turn in a 4-page (max) conference style writeup.

The papers must not be ones that we covered in class, although they can be related.

You need to do your own research to find the papers yourself (i.e., that is part of the assignment).

The majority of the papers must have been published in the last 10 years (so no old or classic papers).

Your grade will be based on how clear, understandable, and accurate your presentation is, and also turning in milestone assignments.

This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work to present complex ideas in a very clear and very simple way.
Our Main Text


- It should be available at the UW bookstore, or you can get it, say, via amazon.com.

- Reading assignment: Read Chapter 9 (and do/re-do all readings listed on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)).
This will be used mainly for the last few weeks.
Other Relevant Texts


Other Relevant Texts

- “Information Theory”, Robert Ash, Dover 1965
- “An Introduction to Information Theory”, Fazlollah M. Reza, Dover 1991
Other Relevant Texts

- “Convex Optimization”, Boyd and Vandenberghe
- “Probability & Measure” Billingsley,
- “Probability with Martingales”, Williams
- “Probability, Theory & Examples”, Durrett
On Our Lecture Slides

- Slides will (hopefully) be available by the early morning before lecture.
- Slides available on our web page (http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/), as well as a printable version.
- Updated slides with typo and bug fixes will be posted as well as (buggy) ones with hand annotations/corrections/fill-ins.
- youtube lectures will again be posted not too long after lectures.
- Annotations are being made with the goodnotes app on the ipad.
Cumulative Outstanding Reading

- Read Chapter 9 in our book (Cover & Thomas, “Information Theory”).
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)).
Homework

- No homework yet.
Email me if you want to skype/google hangout rather than come to office hours.
This has been our guiding principle so far, and will continue to guide us for a while (key exception is network information theory that we will cover, where the above picture may become an arbitrary directed graph with multiple sources/receivers).

So far, channel has been discrete, but real-world channels are not discrete, we thus need differential entropy.
Shannon’s model of communication

- **Shannon’s general model of communication**

- This has been our guiding principle so far, and will continue to guide us for a while (key exception is network information theory that we will cover, where the above picture may become an arbitrary directed graph with multiple sources/receivers).
Shannon’s model of communication

- Shannon’s general model of communication

This has been our guiding principle so far, and will continue to guide us for a while (key exception is network information theory that we will cover, where the above picture may become an arbitrary directed graph with multiple sources/receivers).

- So far, channel has been discrete, but real-world channels are not discrete,
Shannon’s model of communication

- Shannon’s general model of communication

- This has been our guiding principle so far, and will continue to guide us for a while (key exception is network information theory that we will cover, where the above picture may become an arbitrary directed graph with multiple sources/receivers).

- So far, channel has been discrete, but real-world channels are not discrete, we thus need differential entropy.
Sources and Channels

<table>
<thead>
<tr>
<th>Source</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>voice</td>
<td>telephone line</td>
</tr>
<tr>
<td>text words</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>Ethernet or internet</td>
</tr>
<tr>
<td>music</td>
<td>mp3 file</td>
</tr>
<tr>
<td>human cells</td>
<td>mitosis</td>
</tr>
<tr>
<td>population of living organisms</td>
<td>life</td>
</tr>
</tbody>
</table>

- So channel can be over a typical communication link (communication over space).
Sources and Channels

<table>
<thead>
<tr>
<th>Source</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>voice</td>
<td>telephone line</td>
</tr>
<tr>
<td>text words</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>Ethernet or internet</td>
</tr>
<tr>
<td>music</td>
<td>mp3 file</td>
</tr>
<tr>
<td>human cells</td>
<td>mitosis</td>
</tr>
<tr>
<td>population of living organisms</td>
<td>life</td>
</tr>
</tbody>
</table>

- So channel can be over a typical communication link (communication over space).
- **Channel can also be storage (communication over time), or generations (biology).**
Sources and Channels

<table>
<thead>
<tr>
<th>Source</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>voice</td>
<td>telephone line</td>
</tr>
<tr>
<td>text words</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>disk</td>
</tr>
<tr>
<td>pictures</td>
<td>Ethernet or internet</td>
</tr>
<tr>
<td>music</td>
<td>mp3 file</td>
</tr>
<tr>
<td>human cells</td>
<td>mitosis</td>
</tr>
<tr>
<td>population of living organisms</td>
<td>life</td>
</tr>
</tbody>
</table>

- So channel can be over a typical communication link (communication over space).
- Channel can also be storage (communication over time), or generations (biology).
- Many applications of IT, not just communication (e.g., Kolmogorov complexity).
This is lecture 19.
This is lecture 19.

You know that the source compresses down to the entropy $H$, but no further.
This is lecture 19.

You know that the source compresses down to the entropy $H$, but no further.

You also know that the signal may be sent through the channel at a rate no more than $C$. 

![Graphs showing the relationship between error exponent and log of probability of error]$^*$
What if we want to compress $R < H$ or transmit $R > C$?
What if we want to compress $R < H$ or transmit $R > C$? $\Rightarrow$ Error.
What if we want to compress $R < H$ or transmit $R > C$? ⇒ Error.

But are all errors created equality? Are all errors as bad as others?
What if we want to compress $R < H$ or transmit $R > C$? $\Rightarrow$ Error.

But are all errors created equality? Are all errors as bad as others?

Lossy compression, where we compress to below the entropy but tolerate certain errors. Maybe errors are not too noticeable (e.g., perceptual masking), or are “good enough” for apps (e.g., cell phone quality such as CDMA or G.nnn, VoIP, jpg for images).
What if we want to compress $R < H$ or transmit $R > C$? ⇒ Error.

But are all errors created equality? Are all errors as bad as others?

Lossy compression, where we compress to below the entropy but tolerate certain errors. Maybe errors are not too noticeable (e.g., perceptual masking), or are “good enough” for apps (e.g., cell phone quality such as CDMA or G.nnn, VoIP, jpg for images).

We can measure errors with a distortion function, and we have generalization of the previously stated results.
What if we want to compress $R < H$ or transmit $R > C$? ⇒ Error.

But are all errors created equality? Are all errors as bad as others?

**Lossy compression**, where we compress to below the entropy but tolerate certain errors. Maybe errors are not too noticeable (e.g., perceptual masking), or are “good enough” for apps (e.g., cell phone quality such as CDMA or G.nnn, VoIP, jpg for images).

We can measure errors with a distortion function, and we have generalization of the previously stated results.

**Rate-distortion curves with achievable region**
Continuous Entropy
Continuous Entropy

- Differential Entropy
Continuous Entropy

- Differential Entropy
- Gaussian Channel
Continuous Entropy

- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel
Continuous Entropy
- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures
Continuous Entropy
- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures
- Generalized inequalities for entropy functions
Continuous Entropy

- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures

- Generalized inequalities for entropy functions
- Negative discrete information measures
Continuous Entropy
- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures
- Generalized inequalities for entropy functions
- Negative discrete information measures
- The general class of entropy functions, e.g., Renyí
Continuous Entropy

- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures

- Generalized inequalities for entropy functions
- Negative discrete information measures
- The general class of entropy functions, e.g., Renyí

More on Compression
Continuous Entropy

- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures

- Generalized inequalities for entropy functions
- Negative discrete information measures
- The general class of entropy functions, e.g., Renyí

More on Compression

- Lempel-Ziv compression
Continuous Entropy
- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures
- Generalized inequalities for entropy functions
- Negative discrete information measures
- The general class of entropy functions, e.g., Renyí

More on Compression
- Lempel-Ziv compression
- Kolmogorov complexity (algorithmic compression)
Continuous Entropy

- Differential Entropy
- Gaussian Channel
- Shannon Capacity for the Gaussian Channel

Information Measures

- Generalized inequalities for entropy functions
- Negative discrete information measures
- The general class of entropy functions, e.g., Renyí

More on Compression

- Lempel-Ziv compression
- Kolmogorov complexity (algorithmic compression)
- Philosophy
This is IT-II

Coding with errors
Coding with errors

- Rate distortion theory
Coding with errors

- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.
This is IT-II

Coding with errors
- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
This is IT-II

Coding with errors

- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory

- Multi-port communication
Coding with errors
   - Rate distortion theory
   - Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
   - multi-port communication
   - Slepian-wolf,
Coding with errors

- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory

- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
Coding with errors

- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory

- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory
Coding with errors

- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory

- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory

Codes on Graphs
This is IT-II

Coding with errors
- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory

Codes on Graphs
- Convolutional coding
This is IT-II

Coding with errors
- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory

Codes on Graphs
- Convolutional coding
- Turbo codes
This is IT-II

Coding with errors
- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory

Codes on Graphs
- Convolutional coding
- Turbo codes
- LDPC codes
Coding with errors
- Rate distortion theory
- Alternating minimization, Blahut-Arimoto algorithm, EM algorithm, variational EM, information geometry.

Network information theory
- multi-port communication
- Slepian-wolf,
- Multiple sensors/receivers, re-coders and node processing, and models for this.
- Polymatroidal theory and network information theory

Codes on Graphs
- Convolutional coding
- Turbo codes
- LDPC codes
- inference and LBP
And more applications in
And more applications in

- Communications
And more applications in

- Communications
- Pattern recognition and machine learning
And more applications in

- Communications
- Pattern recognition and machine learning
- **Compute Science**
And more applications in

- Communications
- Pattern recognition and machine learning
- Compute Science
- Biology
And more applications in

- Communications
- Pattern recognition and machine learning
- Compute Science
- Biology
- Philosophy (and poetry).
And more applications in

- Communications
- Pattern recognition and machine learning
- Compute Science
- Biology
- Philosophy (and poetry).
- Social science
Let's next review a bit of differential entropy from last quarter.
Let $X$ now be a continuous r.v. with cumulative distribution

$$F(x) = \Pr(X \leq x)$$ \hspace{1cm} (19.5)

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

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

**Definition 19.4.1 (differential entropy $h(X)$)**

$$h(X) = -\int_S f(x) \log f(x) \, dx$$ \hspace{1cm} (19.6)

Since we integrate over only the support set, no worries about $\log 0$.

Perhaps it is best to do some examples.
Here, $X \sim U[0, a]$ with $a \in \mathbb{R}_{++}$.

Then

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = -\log \frac{1}{a}$$

(19.5)

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.
Normal (Gaussian) distributions are very important.

We have:

\[ X \sim N(0, \sigma^2) \iff f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}x^2/\sigma^2} \] (19.5)

Lets compute the entropy of \( f \) in nats.

\[ h(X) = -\int f \ln f = -\int f(x) \left[ -\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \] (19.6)

\[ \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} + \frac{1}{2} \ln(2\pi\sigma^2) \] (19.7)

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

\[ = \frac{1}{2} \log(2\pi e\sigma^2) \text{bits} \] (19.8)

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 19.4.1**

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

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

(19.5)

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

**Definition 19.4.2**

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

Note: $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.

Thus, we have upper/lower bounds on the probability

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

(19.6)
Differential vs. Discrete Entropy

- This follows since (as expected)
  \[ \sum_i \Delta f(x_i) = \sum_i \Delta \left( \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x)dx \right) = \Delta \frac{1}{\Delta} \int f(x)dx = 1 \]

\[ (19.18) \]

- Also, as \( \Delta \to 0 \), we have \(-\log \Delta \to \infty\) and (assuming all is integrable in a Riemannian sense)
  \[ -\sum_i \Delta f(x_i) \log f(x_i) \to -\int f(x) \log f(x)dx \]

\[ (19.19) \]

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

\[ (19.20) \]

- 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

The joint differential entropy is defined as:

\[ h(X_1, X_2, \ldots, X_n) = - \int f(x_1:n) \log f(x_1:n) dx_1:n \] (19.1)

Conditional differential entropy

\[ h(X|Y) = - \int f(x, y) \log f(x|y) dx dy = h(X, Y) - h(Y) \] (19.2)
Entropy of a Multivariate Gaussian

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

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \quad (19.1)$$

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

$$h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma|\right] \text{ bits} \quad (19.2)$$

- 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 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.
Relative Entropy/KL-Divergence & Mutual Information

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

\[
D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \geq 0 \tag{19.8}
\]

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

- Mutual Information:

\[
D(f(X,Y)||f(X)f(Y)) = I(X;Y) = h(X) − h(X|Y) \tag{19.9}
\]
\[
= h(Y) − h(Y|X) \geq 0 \tag{19.10}
\]

- 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

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

(19.8)

- And bounds of the form

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

(19.9)

- 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

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

(19.10)

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.
Theorem 19.4.1

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 $EX = 0$ and $EXX^\top = K$. Then

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

(19.9)

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

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

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 $X$. This is quite hard to analyze 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...
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.
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 \( X \).
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 $X$.
- This is quite hard to analyze 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 $x \in X$, $p(y|x)$ 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 $X$.
- This is quite hard to analyze so we may consider only additive noise $Y = X + Z$ where $Z$ is a random variable.
- We further simplify by saying that $Z \perp X$.
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 $X$.
- This is quite hard to analyze 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 $Z_i \perp \!\!\!\!\perp X_i$. 

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 $\sigma^2 > 0$, what is the capacity?

Then, 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.
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 $[0, 1)$), number of bits per channel use is $\infty$.

If $\sigma^2 > 0$, what is the capacity?

Then, 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.
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 $[0, 1)$), number of bits per channel use is $\infty$. 

If input power is constrained as well (which is also more practical and realistic), then the problem becomes interesting.
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 \( Z_i \perp \perp X_i \).
- 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 \( \sigma^2 > 0 \), what is the capacity?
Gaussian Channel

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

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

- Above is our model, where \( Y_i = X_i + Z_i \), with \( Z_i \sim \mathcal{N}(0, \sigma^2) \) 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 \([0, 1)\)), number of bits per channel use is \( \infty \).

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

- 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 \( \sigma^2 > 0 \), what is the capacity? Then, 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.
Power constraint

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

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

(19.1)

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

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

$$
\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P \tag{19.1}
$$

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

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P$$  \hspace{1cm} (19.1)

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 $\approx E X^2$.

- 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 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

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

(19.1)

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 time window size and bound the maximum of the averages within each window).

- But let's stick with the one above in Equation (19.1).
Example

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

- 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?
Example

- 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 $Y < 0$. 

Example

- 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:

\[
(19.3)
\]
Example

- 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:

\[ P_e \]

(19.3)
Example

- 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 \( Y < 0 \).
- Error:

\[
P_e = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P}) \quad (19.2)
\]

(19.3)
Example

- 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:

\[
P_e = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P}) \tag{19.2}
\]

\[
= \frac{1}{2} \Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2} \Pr(Z > \sqrt{P}|X = -\sqrt{P}) \tag{19.3}
\]
Example

- 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 \( Y < 0 \).
- Error:

\[
P_e = \frac{1}{2} \Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2} \Pr(Y > 0|X = -\sqrt{P})
\] (19.2)

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

\[
= \Pr(Z > \sqrt{P})
\] (19.3)
Example

- The two separate error types

\[ \text{or} \]

\[ \text{or} \]
Example

- The two separate error types

\[ 0 \pm \sqrt{P} \]

or

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

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

\[ \text{Pr}(Z > \pm \sqrt{P}) = 1 - \Phi(\frac{\pm \sqrt{P}}{\sigma}) \] (19.4)

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

\[ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt \] (19.5)
Example

- The two separate error types

  \[
  \begin{array}{c}
  \text{0} + \sqrt{P} \\
  \text{or} \\
  \text{0} - \sqrt{P}
  \end{array}
  \]

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

  \[
  \begin{array}{c}
  -\sqrt{P} \\
  \text{0} + \sqrt{P}
  \end{array}
  \]

- We have that

  \[
  \Pr(Z > \sqrt{P}) = 1 - \Phi\left(\frac{\sqrt{P}}{\sigma^2}\right) \quad (19.4)
  \]

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

  \[
  \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad (19.5)
  \]
In fact, we have essentially just turned a Gaussian channel into a discrete BSC:

\[
\begin{array}{c}
X \xrightarrow{p} Y \\
0 \quad 1 \\
1 \xrightarrow{1-p} 0
\end{array}
\]

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

\[
\begin{array}{ccc}
0 & \rightarrow & 0 \\
X &  & Y \\
1 & \rightarrow & 1
\end{array}
\]

\[
\begin{array}{ccc}
0 & \rightarrow & 0 \\
X &  & Y \\
1 & \rightarrow & 1
\end{array}
\]

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

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

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

\[
\begin{array}{c}
\text{X} \\
0 \\
1
\end{array}
\begin{array}{c}
\text{Y} \\
0 \\
1
\end{array}
\begin{array}{c}
1 - p \\
p \\
1 - p
\end{array}
\begin{array}{c}
1 - p \\
p \\
1 - p
\end{array}
\]

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.

\[ C = \max_{p(x)} \mathbb{E} X^2 \leq P \]

\[ I(X;Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) = \mathbb{H}(Y) - \mathbb{H}(Z|X) \] (19.7)

Note, when \( X \perp Z \), \( \mathbb{H}(X+Z) \geq \mathbb{H}(Z) \), in general entropy of sums of r.v.s is tricky to characterize.
We need a capacity notion, but here under a power constraint.

**Definition 19.5.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (19.6)$$
We need a capacity notion, but here under a power constraint.

**Definition 19.5.1**

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} (19.6)

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.

**Definition 19.5.1**

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

$$C = \max_{p(x) : E X^2 \leq P} I(X; Y) \text{ bits} \quad (19.6)$$

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

$$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(Z|X) \quad (19.8)$$

Note, when $X \perp Z$, $h(X+Z) \geq h(Z)$, in general entropy of sums of r.v.s is tricky to characterize.

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

**Definition 19.5.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (19.6)$$

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

$$I(X; Y) \quad (19.8)$$
We need a capacity notion, but here under a power constraint.

**Definition 19.5.1**

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} (19.6)

- 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

  $$I(X; Y) = h(Y) - h(Y | X)$$  \hspace{1cm} (19.8)
We need a capacity notion, but here under a power constraint.

**Definition 19.5.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X;Y) \text{ bits} \quad (19.6)$$

- 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
  $$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X) \quad (19.7)$$
  $$= h(Y) - h(Z|X) \quad (19.8)$$

Note, when $X \perp Z$, $h(X+Z) \geq h(Z)$, in general entropy of sums of r.v.s is tricky to characterize.
Capacity of Gaussian Channel

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

**Definition 19.5.1**

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

$$C = \max_{p(x): EX^2 \leq P} I(X; Y) \text{ bits} \quad (19.6)$$

- 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

  $$I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (19.7)$$

  $$= h(Y) - h(Z|X) \quad (19.8)$$
Capacity of Gaussian Channel

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

**Definition 19.5.1**

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

$$C = \max_{p(x) : EX^2 \leq P} I(X;Y) \text{ bits} \quad (19.6)$$

- 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

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) \quad (19.7)$$

$$= h(Y) - h(Z|X) = h(Y) - h(Z) \quad (19.8)$$
We need a capacity notion, but here under a power constraint.

**Definition 19.5.1**

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

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

(19.6)

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

(19.7)

(19.8)

- Note, when $X \perp Z$, $h(X + Z) \geq h(Z)$, in general entropy of sums of r.v.s is tricky to characterize.
Capacity of Gaussian Channel

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

**Definition 19.5.1**

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} (19.6)

- 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} (19.7)
  
  $$= h(Y) - h(Z|X) = h(Y) - h(Z)$$  \hspace{1cm} (19.8)

- Note, when $X \perp \! \! \perp Z$, $h(X+Z) \geq h(Z)$, in general entropy of sums of r.v.s is tricky to characterize.

- **Strategy for finding $C$** same as before: 1) upper bound $I(X;Y)$, and then; 2) find a (not nec. unique) $p(x)$ achieving the upper bound.
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 $EZ = 0$. 

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

$X_i \rightarrow \oplus \rightarrow Y_i$
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

\[
h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \tag{19.9}
\]
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

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

  (19.9)

- Also,

  $$EY^2$$

  (19.11)
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

$$h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|]$$  \hspace{1cm} (19.9)

- Also,

$$EY^2 = E(X + Z)^2$$  \hspace{1cm} (19.11)
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

$$h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \quad (19.9)$$

Also,

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2$$

since $X \perp Z$
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

$$h(X) \leq \frac{1}{2} \log[(2\pi e)^2 |K|] \quad (19.9)$$

Also,

$$EY^2 = E(X + Z)^2 = EX^2 + 2EXEZ + EZ^2$$

since $X \perp Z$

$$= P + \sigma^2$$

signal power, $EX^2$

noise power, $EZ^2$
Capacity of Gaussian Channel

- Thus, we can upper bound the mutual information as follows:

\[ I(X; Y) \]

(19.13)
Capacity of Gaussian Channel

- Thus, we can upper bound the mutual information as follows:

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

(19.13)
Thus, we can upper bound the mutual information as follows:

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

(19.12)

(19.13)

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

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

(19.14)
Thus, we can upper bound the mutual information as follows:

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

(19.12)

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

(19.13)

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

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

(19.12)

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

(19.13)
Capacity of Gaussian Channel

- Thus, we can upper bound the mutual information as follows:

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

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

where SNR is the signal to noise ratio.
Capacity of Gaussian Channel

- Thus, we can upper bound the mutual information as follows:

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

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

(19.12) (19.13)

where \( \text{SNR} \) 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:

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

(19.12)

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

(19.13)

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

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

(19.14)
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} (19.15)
The capacity of the Gaussian channel is

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

(19.15)

- So, the maximum capacity achieved when \( X \sim \mathcal{N}(0, P) \).
The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (19.15)

- So, the maximum capacity achieved 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).
The capacity of the Gaussian channel is

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

(19.15)

- So, the maximum capacity achieved 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).
- Without a power constraint, capacity is infinite.
Ex: PCM sampled and quantized audio, 6dB SNR/bit

The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \quad (19.16) \]
Ex: PCM sampled and quantized audio, 6dB SNR/bit

The capacity of the Gaussian channel is

$$C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \quad (19.16)$$

- In fact, from this we get the standard 6.02dB SNR/bit for PCM sampled and quantized audio. I.e.,

$$16\text{bits/channel use} = \frac{1}{2} \log(1 + \text{SNR}) \quad (19.17)$$

or $2^{32} = 1 + \text{SNR}$ or $\text{SNR} = 2^{32} - 1$. 


Ex: PCM sampled and quantized audio, 6dB SNR/bit

The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (19.16)

- In fact, from this we get the standard 6.02dB SNR/bit for PCM sampled and quantized audio. I.e.,

\[ 16\text{bits/channel use} = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (19.17)

or \[ 2^{32} = 1 + \text{SNR} \text{ or } \text{SNR} = 2^{32} - 1. \]

- We get standard 96dB SNR from 16bit PCM samples via:

\[ 10 \log_{10}(\text{SNR}) = 10 \times 32 \log_{10}(2) \approx 96.33\text{dB} \]  \hspace{1cm} (19.18)
The capacity of the Gaussian channel is

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

(19.16)

- In fact, from this we get the standard 6.02dB SNR/bit for PCM sampled and quantized audio. I.e.,

\[
\text{16 bits/channel use} = \frac{1}{2} \log(1 + \text{SNR}) \]  

(19.17)

or \(2^{32} = 1 + \text{SNR}\) or \(\text{SNR} = 2^{32} - 1\).

- We get standard 96dB SNR from 16bit PCM samples via:

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

(19.18)

- And 96.33/16 \(\approx 6.02\text{dB/bit}\).
Ex: PCM sampled and quantized audio, 6dB SNR/bit

The capacity of the Gaussian channel is

\[ C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (19.16)

- In fact, from this we get the standard 6.02dB SNR/bit for PCM sampled and quantized audio. I.e.,

\[ 16\text{bits/channel use} = \frac{1}{2} \log(1 + \text{SNR}) \]  \hspace{1cm} (19.17)

or \[ 2^{32} = 1 + \text{SNR} \text{ or } \text{SNR} = 2^{32} - 1. \]

- We get standard 96dB SNR from 16bit PCM samples via:

\[ 10 \log_{10}(\text{SNR}) = 10 \times 32 \log_{10}(2) \approx 96.33\text{dB} \]  \hspace{1cm} (19.18)

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

- Every additional bit (PCM sampled/quantized audio, e.g., audio compact disk) adds 6.02 dB of SNR.
Capacity of the channel definitions

Definition 19.5.2

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

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

(19.19)

3. Decoding function

\[g : \mathcal{Y}^n \rightarrow \{1, \ldots, M\}\]

(19.20)
### Capacity of the channel definitions

**Definition 19.5.3**

The rate is

\[ R = \frac{\log M}{n} \text{ bits per channel use} \quad (19.21) \]

**Definition 19.5.4**

A rate \( R \) is achievable if \( \exists \) a sequence of \( (2^{nR}, n) \) codes satisfying the power constraint \( P \) s.t. \( \lambda^{(n)} \to 0 \) as \( n \to \infty \).

**Definition 19.5.5**

The capacity of a Gaussian channel is the supremum of the achievable rates (i.e., the largest possible achievable rate).

Compare with Lecture 14 on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/).
The Capacity of a Gaussian channel with input power constraint $P$ and noise variance $\sigma^2$ is

$$C = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) \text{ bits per channel use}$$ (19.22)

proof sketch.

- Typical $x$ set $A_{\epsilon}^{(n)}$ has volume $\leq 2^{n(h(X)+\epsilon)}$
The Capacity of a Gaussian channel with input power constraint $P$ and noise variance $\sigma^2$ is

$$C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) \text{ bits per channel use} \quad (19.22)$$

proof sketch.

- Typical $x$ set $A^{(n)}_\epsilon$ has volume $\leq 2^{n(h(X)+\epsilon)}$
- conditional typical $Y$ volume $\leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)}$. 
The Capacity of a Gaussian channel with input power constraint $P$ and noise variance $\sigma^2$ is

$$C = \frac{1}{2} \log(1 + \frac{P}{\sigma^2}) \text{ bits per channel use} \quad (19.22)$$

**proof sketch.**

- Typical $x$ set $A^{(n)}_\epsilon$ has volume $\leq 2^{n(h(X)+\epsilon)}$
- Conditional typical $Y$ volume $\leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)}$.
- Unconditional typical set volume of $Y$ is $\leq 2^{n(h(Y)+\epsilon)}$ but $h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)]$ and $h(Z) = \frac{1}{2} \log[2\pi e\sigma^2]$
Capacity of the channel definitions

**proof sketch.**

How many $X$-conditional volumes can we pack into total available volume? Ratio of total available volume to $X$-conditional volume.

\[ \leq \frac{2^{n \cdot h(Y)}}{2^{n \cdot h(Z)}} = \frac{2^{n \frac{1}{2} \log[2\pi e(P + \sigma^2)]}}{2^{n \frac{1}{2} \log[2\pi e\sigma^2]}} \approx 2^{\frac{n}{2} \log \frac{P + \sigma^2}{\sigma^2}} = \left[\left(\frac{P + \sigma^2}{\sigma^2}\right)\right]^{n/2} \]
How many $X$-conditional volumes can we pack into total available volume? Ratio of total available volume to $X$-conditional volume.

\[
\leq \frac{2^{nh(Y)}}{2^{nh(Z)}} = \frac{2^{n \frac{1}{2} \log[2\pi e(P+\sigma^2)]}}{2^{n \frac{1}{2} \log[2\pi e\sigma^2]}} \approx 2^{n \frac{1}{2} \log \frac{P+\sigma^2}{\sigma^2}} = \left[\frac{(P + \sigma^2)/\sigma^2}{\sigma^2}\right]^{n/2}
\]

Assuming no overlap of volumes which is best (max rate) we can do.
Capacity of the channel definitions

proof sketch.

- How many $X$-conditional volumes can we pack into total available volume? Ratio of total available volume to $X$-conditional volume.

$$\leq \frac{2^{nh(Y)}}{2^{nh(Z)}} = \frac{2^n \frac{1}{2} \log[2\pi e (P+\sigma^2)]}{2^n \frac{1}{2} \log[2\pi e \sigma^2]} \approx 2^{\frac{n}{2} \log \frac{P+\sigma^2}{\sigma^2}} = [(P + \sigma^2)/\sigma^2]^{n/2}$$

- Assuming no overlap of volumes which is best (max rate) we can do.

- The above is measured in counts for $n$ channel usages. To convert it into bits per channel use, we take log and divide by $n$ to get

$$2^{\frac{n}{2} \log \frac{P+\sigma^2}{\sigma^2}} = 2^n R$$ meaning that $$R = \frac{1}{2} \log(1 + P/\sigma^2) \quad (19.23)$$
Capacity of the channel definitions

- If everything is jointly Gaussian i.i.d., typical volumes will be spheres.
Capacity of the channel definitions

- If everything is jointly Gaussian i.i.d., typical volumes will be spheres.
- Relate sphere volume to typical set to get the radius (in $n$-D)

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{\frac{n}{2}} \log[2\pi e\sigma^2] = (2\pi er^2)^{n/2} \quad (19.24)$$

and we get

$$r\sigma^2 = \Gamma^{1/2}\left(\frac{n}{2} + 1\right)(2e\sigma^2)^{1/2} \quad (19.25)$$
Capacity of the channel definitions

- If everything is jointly Gaussian i.i.d., typical volumes will be spheres.
- Relate sphere volume to typical set to get the radius (in $n$-D)

\[
V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{\frac{n}{2}} \log[2\pi e\sigma^2] = (2\pi e r^2)^{n/2}
\]  

(19.24)

and we get

\[
r_\sigma^2 = \Gamma^{1/2}\left(\frac{n}{2} + 1\right)(2e\sigma^2)^{1/2}
\]

(19.25)

- Bin packing figure, how many small spheres fit in large one:
proof $\Delta$ from discrete proof.

**proof sketch $\Delta$ continued.**

- We need to show that if $R < C$, $\exists$ a code with $P^n_e \to 0$ with $n \to \infty$.

- We do random codeword generation (like in discrete case) but in this case from Gaussians with $EX^2 = P - \epsilon$ so that

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \to P - \epsilon \text{ as } n \to \infty \quad (19.26)$$

- Also have an additional source of possible error

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i^2(1) > P \right\} \quad (19.27)$$
Proof $\Delta$ from discrete proof.

**proof sketch $\Delta$.**

- We add $E_0$ to the other errors we previously had in discrete case (note again we use same trick to show that considering only message 1 is sufficient due to symmetry).

- By the weak law of large numbers, $E_0 \rightarrow 0$ also as $n \rightarrow \infty$ as do the other sources of errors.