Elements of Information Theory

Lecture 6
Differential Entropy and The Gaussian Channel

Instructor: Yichen Wang

Ph.D./Professor

School of Information and Communications Engineering
Division of Electronics and Information
Xi'an Jiaotong University

Outlines

- Differential Entropy
- AEP for Continuous Random Variable
- Mutual Information
- Gaussian Channel and Channel Capacity

Differential Entropy

Differential Entropy for Continuous Random Variable

The differential entropy h(X) of a continuous random variable X with density f(x) is defined as

$$h(X) = -\int_S f(x) \log f(x) dx$$

where S is the support set of the random variable.

Discussions:

1. ***As differential entropy involves an integral and a density, we should include the statement if it exists.***
2. ***Is differential entropy also nonnegative?***

Differential Entropy

Example (Uniform Distribution)

Consider a random variable distributed uniformly from 0 to a so that its density is 1/a from 0 to a and 0 elsewhere. Then its differential entropy is

$$\begin{aligned} h(X) &= -\int_S f(x)\log f(x)dx \\ &= -\int_0^a \frac{1}{a}\log\left(\frac{1}{a}\right)dx = \log a \text{ bits} \end{aligned}$$

Differential Entropy

Example (Normal Distribution)

Let random variable X follows normal distribution, i.e.,

$$X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

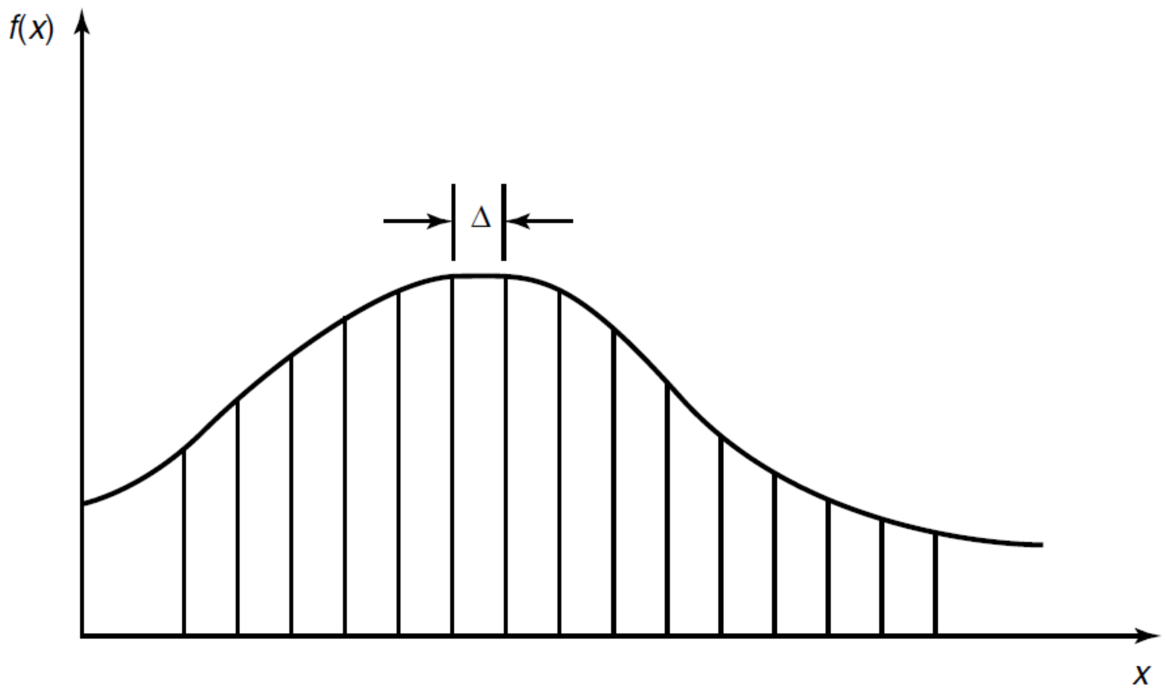
Then calculating differential entropy in nats, we obtain

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \phi(x) \ln\phi(x) dx = -\int_{-\infty}^{\infty} \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2} \right] dx \\ &= \frac{\mathbb{E}\left\{X^2\right\}}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 = \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2}\ln 2\pi e\sigma^2 \text{ nats} = \frac{1}{2}\log 2\pi e\sigma^2 \text{ bits} \end{aligned}$$

Differential Entropy

Relation of Differential Entropy to Discrete Entropy

- ***Consider a random variable X with density f(x)***
- ***Divide the range of X into bins of length Δ***

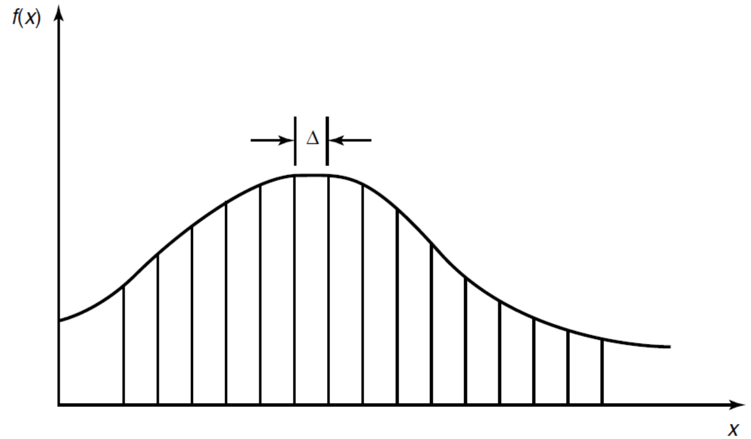


Differential Entropy

Relation of Differential Entropy to Discrete Entropy

- ***The mean value theorem tells us that***

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$$



- ***Construct the quantized random variable*** X^Δ ***:***

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

- ***The probability that*** $X^\Delta = x_i$ ***is***

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

Differential Entropy

Relation of Differential Entropy to Discrete Entropy

- ***The entropy of the quantized random variable is***

$$
\begin{aligned}
H\left(X^{\Delta}\right) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i \\
&= -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \cdot \log\Big(f(x_i)\Delta\Big) \\
&= -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \cdot \log f(x_i) - \sum_{i=-\infty}^{\infty} f(x_i)\Delta \cdot \log\Delta \\
&= \boxed{-\sum_{i=-\infty}^{\infty} f(x_i)\Delta \cdot \log f(x_i)} - \log\Delta
\end{aligned}
$$

Differential Entropy

Theorem

If the density f(x) of the random variable X is Riemann integrable, then

$$H\left(X^{\Delta}\right) + \log\Delta \quad \longrightarrow \quad h(f) = h(X), \quad \text{as } \Delta \ \rightarrow \ 0.$$

Thus, the entropy of an n-bit quantization of a continuous random variable X with $\Delta=2^{-n}$ is approximately h(X)+n.

⇩

Differential Entropy: $h(X) = \int_{-\infty}^{\infty} f(x)\log\frac{1}{f(x)}dx$

Differential Entropy

Something not good:

- ***h(X) does not give the amount of information for X***
- ***h(X) is not necessarily positive***

Something we expect:

- ✓ ***Compare the uncertainty of two continuous random variables (quantized to the same precision)***
- ✓ ***Mutual information still works***

Theorem $$h(aX) = h(X) + \log|a|$$

Theorem $$h(X + c) = h(X)$$

Differential Entropy

Theorem

If we have the constraints that $\mathbb{E}\{X\} = 0$ ***and*** $\mathbb{E}\{X^2\} = \sigma^2$***, the Gaussian (normal) distribution have the maximum differential entropy.***

$$\text{Let } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \Longrightarrow h\big(X, p(x)\big) = \frac{1}{2}\log 2\pi e\sigma^2$$

Suppose $q(x)$ be another probability density function

$$-\int q(x)\log p(x)dx = -\int q(x)\log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\right]dx = \frac{1}{2}\log 2\pi e\sigma^2 = h\big(X, p(x)\big)$$

$$h\big(X, q(x)\big) - \int q(x)\log\frac{1}{p(x)}dx = \int q(x)\log\frac{p(x)}{q(x)}dx \le \log\int q(x)\frac{p(x)}{q(x)}dx = 0$$

$$h\big(X, q(x)\big) \le h\big(X, p(x)\big)$$

Differential Entropy

Definition (Joint Differential Entropy)

The differential entropy of a set $X_1, X_2, \ldots, X_n$ of random variables with density $f(x_1, x_2, \ldots, x_n)$ is defined as

$$h(X_1, X_2, \cdots, X_n) = -\int f(x^n) \log f(x^n) dx^n.$$

Definition (Conditional Differential Entropy)

If X, Y have a joint density function f(x, y), we can define the conditional differential entropy h(X|Y) as

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

Moreover, we have the following relationship:

$$h(X|Y) = h(X, Y) - h(Y).$$

Differential Entropy

Theorem

(Entropy of a Multivariate Normal Distribution)

Let $X_1, X_2, \dots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K. Then

$$h(X_1, X_2, \cdots, X_n) = h\left(\mathcal{N}_n(\mu, K)\right) = \frac{1}{2}\log\left(2\pi e\right)^n |K| \ \text{ bits},$$

where $|K|$ denotes the determinant of K.

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left|K\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$$

Differential Entropy

$$
\begin{aligned}
h(f) &= -\int f(\mathbf{x}) \ln f(\mathbf{x}) d\mathbf{x} \\
&= -\int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1} (\mathbf{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}} \right] d\mathbf{x} \\
&= \frac{1}{2}\mathbb{E}\left\{ \sum_{i,j} (X_i - \mu_i)\left(K^{-1}\right)_{ij}(X_j - \mu_j) \right\} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\sum_{i,j} \mathbb{E}\left\{ (X_i - \mu_i)(X_j - \mu_j) \right\} \left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\sum_{j}\sum_{i} K_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n |K| \\
&= \frac{1}{2}\ln(2\pi e)^n |K| \ \text{ nats} \\
&= \frac{1}{2}\log(2\pi e)^n |K| \ \text{ bits}
\end{aligned}
$$

Outlines

- Differential Entropy
- AEP for Continuous Random Variable
- Mutual Information
- Gaussian Channel and Channel Capacity

AEP for Continuous R. V.

Theorem (AEP for Continuous Random Variables)

Let X_1, X_2, … , X_n be a sequence of random variables drawn i.i.d. according to the density f(x). Then, we have

$$-\frac{1}{n}\log f\left(X_1,\cdots,X_n\right) \to \mathbb{E}\left\{-\log f(X)\right\} = h(X) \quad \text{in probability.}$$

Definition (Typical Set)

For ε>0 and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to f(x) as follows:

$$A_\epsilon^{(n)} = \left\{(x_1,\cdots,x_n) \in S^n : \left|-\frac{1}{n}\log f\left(X_1,\cdots,X_n\right) - h(X)\right| \le \epsilon\right\},$$

where $f(x_1,x_2,\cdots,x_n) = \prod_{i=1}^{n} f(x_i)$.

AEP for Continuous R. V.

The analog of the cardinality of the typical set for the discrete case is the volume of the typical set for continuous random variables.

Definition (Volume)

The volume* Vol(A) *of the set* $A \subset \mathcal{R}^n$ *is defined as

$$\mathrm{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n.$$

AEP for Continuous R. V.

Theorem

The typical set $A_\epsilon^{(n)}$ ***has the following properties:***

1. $\Pr\left\{A_\epsilon^{(n)}\right\} > 1-\epsilon$ ***for n sufficiently large.***
2. $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \le 2^{n[h(X)+\epsilon]}$ ***for all n.***
3. $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \ge (1-\epsilon)2^{n[h(X)-\epsilon]}$ ***for n sufficiently large.***

Theorem

The set $A_\epsilon^{(n)}$ ***is the smallest volume set with probability*** ***≥ 1-ε, to first order in the exponent.***

AEP for Continuous R. V.

Proof for Property 2

$$\begin{aligned} 1 &= \int_{S^n} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n[h(X)+\epsilon]} dx_1 dx_2 \cdots dx_n = 2^{-n[h(X)+\epsilon]} \mathrm{Vol}\left(A_\epsilon^{(n)}\right) \end{aligned}$$

Proof for Property 3

$$\begin{aligned} 1-\epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n \\ &\leq \int_{A_\epsilon^{(n)}} 2^{-n[h(X)-\epsilon]} dx_1 dx_2 \cdots dx_n = 2^{-n[h(X)-\epsilon]} \mathrm{Vol}\left(A_\epsilon^{(n)}\right) \end{aligned}$$

Outlines

- Differential Entropy
- AEP for Continuous Random Variable
- Mutual Information
- Gaussian Channel and Channel Capacity

Mutual Information

Definition

(Mutual Information for Continuous R.V.)

The mutual information I(X;Y) between two random variables with joint density f(x,y) is defined as

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y) \end{aligned}$$

Mutual Information

Question:

What can mutual information between two random variables be viewed as?

The limit of the mutual information between their quantized versions.

$$
\begin{aligned}
I\left(X^{\Delta};Y^{\Delta}\right) &= H\left(X^{\Delta}\right) - H\left(X^{\Delta}|Y^{\Delta}\right) \\
&\approx h(X) - \log\Delta - \left[h(X|Y) - \log\Delta\right] = I(X;Y)
\end{aligned}
$$

$I(X;Y) \geq 0$ with equality iff X and Y are independent.

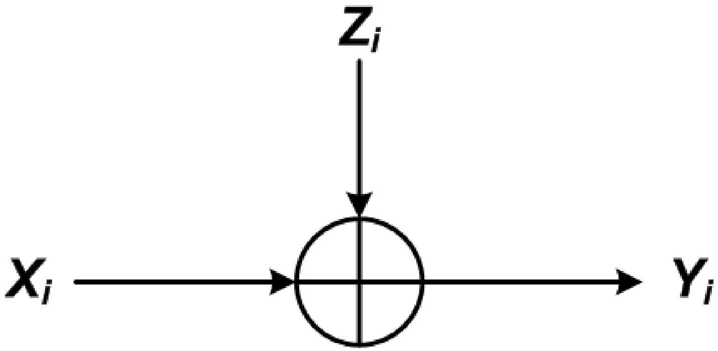
$h(X|Y) \leq h(X)$ with equality iff X and Y are independent.

Outlines

- **Differential Entropy**
- **AEP for Continuous Random Variable**
- **Mutual Information**
- **Gaussian Channel and Channel Capacity**

Gaussian Channel and Capacity

The most important continuous channel is the Gaussian channel.



$$Y_i = \underline{X_i + Z_i}, \quad Z_i \sim \mathcal{N}(0, N)$$

Z_i is assumed to be independent of signal X_i

If the noise variance is zero or the input is unconstrained, the capacity of the channel is infinite.

⇩

How about the capacity with input power constraint?

Gaussian Channel and Capacity

<u>Definition (Information Capacity)</u>

The information capacity of the Gaussian channel with power constraint P is

$$C = \max_{f(x):\mathbb{E}\{X^2\}\le P} I(X;Y)$$

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) = h(Y) - h(Z) \end{aligned}$$

$$Z \sim \mathcal{N}(0,N) \quad \longrightarrow \quad h(Z) = \frac{1}{2}\log 2\pi e N$$

$$\mathbb{E}\{Y^2\} = \mathbb{E}\{(X+Z)^2\} = \mathbb{E}\{X^2\} + 2\mathbb{E}\{X\}\mathbb{E}\{Z\} + \mathbb{E}\{Z^2\} = P + N$$

Gaussian Channel and Capacity

$$
\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) \\
&= h(Y) - h(X+Z|X) = h(Y) - h(Z) \\
&\le \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi e N = \frac{1}{2}\log\left(1+\frac{P}{N}\right)
\end{aligned}
$$

Why?

The information capacity of the Gaussian channel is

$$C = \max_{f(x):\mathbb{E}\{X^2\}\le P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right),$$

and the maximum is attained when $X \sim \mathcal{N}(0,P)$ **.**

Gaussian Channel and Capacity

Definition

An (M,n) code for the Gaussian channel with power constraint P consists of the following:

1. ***An index set {1, 2, … , M}.***
2. ***An encoding function x: {1, 2, … , M}*** $\rightarrow \chi^n$***, yielding codewords*** $x^n(1), x^n(2), \ldots, x^n(M)$***, satisfying the power constraint P; that is, for every codeword***

$$\sum_{i=1}^{n} x_i^2(w) \leq nP, \quad w = 1, 2, \cdots, M.$$

3. ***A decoding function***

$$g : \mathcal{Y}^n \longrightarrow \{1, 2, \cdots, M\}.$$

Gaussian Channel and Capacity

Definition

A rate R is said to be achievable for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates.

Theorem

The capacity of a Gaussian channel with power constraint P and noise variance N is

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right) \quad \text{bits per transmission.}$$

Gaussian Channel and Capacity

Proof:

1. Generation of the codebook.

- *We wish to generate a codebook in which all the codewords satisfy the power constraint.*
- *Generate the codewords with each element i.i.d. according to a normal distribution with variance P-ε.*
- *Let $X_i(w)$, $i = 1, 2, \ldots, n$, $w = 1, 2, \ldots, 2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P-\varepsilon)$, forming codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR}) \in R^n$.*

Gaussian Channel and Capacity

Proof:

2. Encoding.

- ***The codebook is revealed to both the sender and the receiver.***
- ***To send the message index w, the transmitter sends the wth codeword $X^n(w)$ in the codebook.***

3. Decoding.

- ***The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector.***
- ***If there is one and only one such codeword $X^n(w)$, the receiver declares $\hat{W} = w$ to be the transmitted codeword.***
- ***Otherwise, the receiver declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.***

Gaussian Channel and Capacity

Proof:

4. Probability of error.

- ***Assume that codeword 1 was sent. Thus, we have***

$$Y^n = X^n(1) + Z^n$$

- ***Define the following events:***

$$E_0 = \left\{ \frac{1}{n} \sum_{j=1}^{n} X_j^2(1) > P \right\}$$

$$E_i = \left\{ \left(X^n(i), Y^n \right) \text{ is in } A_\epsilon^{(n)} \right\}$$

- ***An error occurs if E_0 occurs or E_1^c occurs or $E_2 \cup \cdots \cup E_{2^{nR}}$ occurs***

Gaussian Channel and Capacity

Proof:

➢ ***Let $\mathcal{E}$ denote the event $\hat{W} \neq W$.***

$$\begin{aligned}
\Pr\{\mathcal{E}|W=1\} &= P\left(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}\right) \\
&\leq P\left(E_0\right) + P\left(E_1^c\right) + \sum_{i=2}^{2^{nR}} P\left(E_i\right) \\
&\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\
&= 2\epsilon + \left(2^{nR}-1\right) 2^{-n(I(X;Y)-3\epsilon)} \\
&\leq 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \\
&\leq 3\epsilon
\end{aligned}$$

For sufficiently large n and $R < I(X;Y) - 3\epsilon$

Gaussian Channel and Capacity

Proof:

- ***Choosing a good codebook and deleting the worst half of the codewords, we obtain a code with low maximal probability of error.***
- ***The power constraint is satisfied by each of the remaining codewords.***
- ***We have constructed a code that achieves a rate arbitrarily close to capacity.***

Gaussian Channel and Capacity

For the baseband model of realistic wireless communications systems, we assume that the signal is bandlimited to W.

We can reconstruct the bandlimited signal from samples under the sampling rate 1/2W.

Consider the time interval [0, T]. The energy per signal sample is

$$\frac{PT}{2WT} = \frac{P}{2W}$$

The power spectral density of AWGN is $N_0/2$. Then, the energy per noise sample is

$$\frac{N_0}{2} 2W \frac{T}{2WT} = \frac{N_0}{2}$$

Gaussian Channel and Capacity

The capacity per sample is

$$C = \frac{1}{2}\log\left(1 + \frac{\frac{P}{2W}}{\frac{N_0}{2}}\right) = \frac{1}{2}\log\left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per sample.}$$

The capacity of the channel is

$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per second.}$$

When the bandwidth is small,

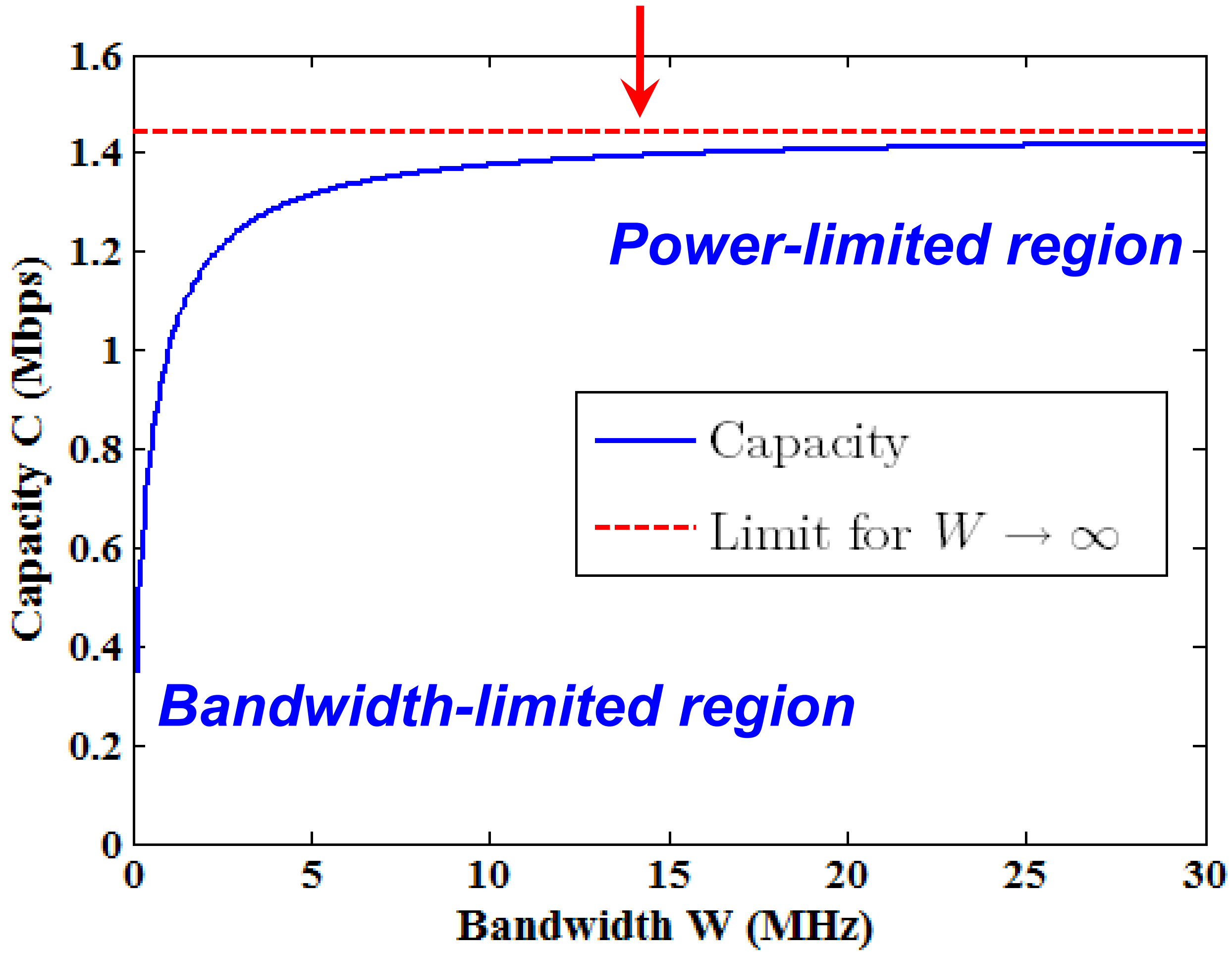
$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \approx W\log\left(\frac{P}{N_0 W}\right) \Longrightarrow \text{bandwidth} - \text{limited regime}$$

When the bandwidth is large,

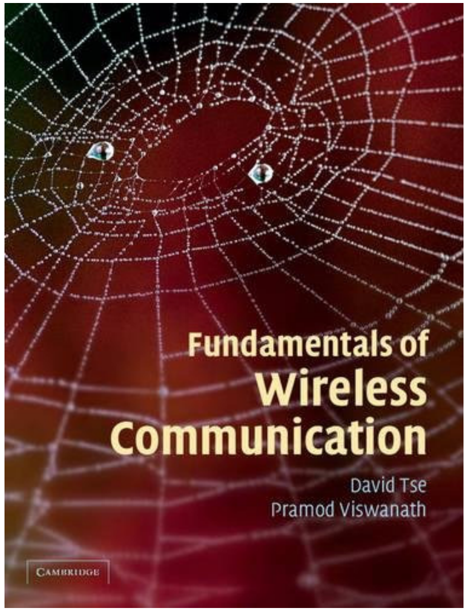
$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \approx W\left(\frac{P}{N_0 W}\right)\log e = \frac{P}{N_0}\log e \Longrightarrow \text{power} - \text{limited regime}$$

Gaussian Channel and Capacity

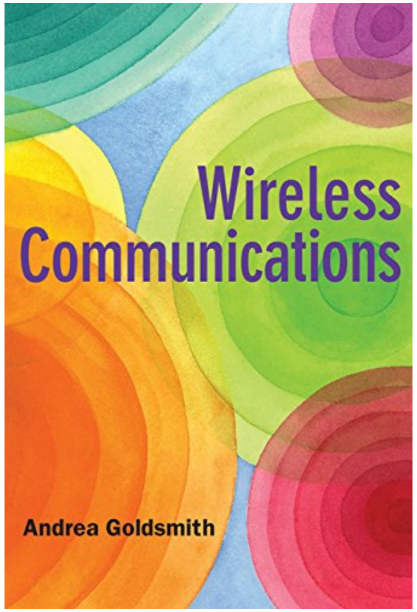
$$C_\infty = \frac{P}{N_0}\log e$$



Gaussian Channel and Capacity



David Tse and Pramod Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.



Andrea Goldsmith, Wireless Communications, Cambridge University Press, 2005.