Logistics Review 

$R(D) = R^{(T)}(D)$

Geometry Scratch

EE515A – Information Theory II
Spring 2012

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Spring Quarter, 2012
http://j.ee.washington.edu/~bilmes/classes/ee515a_spring_2012/

Lecture 25 - April 20th, 2012
Outstanding Reading

- Read all chapters assigned from IT-I (EE514, Winter 2012).
- Read chapter 8 in the book.
- Read chapter 9 in the book.
- Read chapter 10 in the book (chapter on rate distortion theory).
Please do use our discussion board (https://catalyst.uw.edu/gopost/board/bilmes/27386/) for all questions, comments, so that all will benefit from them being answered.
The rate-distortion function $R(D)$ for Bernoulli($p$) with $d(x, \hat{x}) = 1\{x \neq \hat{x}\}$ (Hamming distortion) has the following form:

$$R(D) = \begin{cases} H(p) - H(D) & \text{if } 0 \leq D \leq \min \{p, 1-p\} \\ 0 & \text{if } D > \min \{p, 1-p\} \end{cases} \quad (1)$$
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?
Distortion vs. Error

- Is it always the case that \( R(D) = H \) at \( D = 0 \)?
- No. If \( D = 0 \) does not require \( P_e = 0 \), then we can compress below the entropy with zero distortion but non-zero error.

![Diagram showing achievable and unachievable regions in rate-distortion analysis.](image-url)
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?
- No. If $D = 0$ does not require $P_e = 0$, then we can compress below the entropy with zero distortion but non-zero error.

Why is $R(0) = H(p)$ in $X \sim \text{Bernoulli}(p)$ r.v. case above?
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?
- No. If $D = 0$ does not require $P_e = 0$, then we can compress below the entropy with zero distortion but non-zero error.

Why is $R(0) = H(p)$ in $X \sim \text{Bernoulli}(p)$ r.v. case above?
- Since Hamming distortion is such that $\{D = 0\} \Leftrightarrow P_e = 0$. 
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?

- No. If $D = 0$ does not require $P_e = 0$, then we can compress below the entropy with zero distortion but non-zero error.

- Why is $R(0) = H(p)$ in $X \sim \text{Bernoulli}(p)$ r.v. case above?

- Since Hamming distortion is such that $\{D = 0\} \iff P_e = 0$.

- Key point (again): distortion not necessarily the same as error.
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?
- No. If $D = 0$ does not require $P_e = 0$, then we can compress below the entropy with zero distortion but non-zero error.

```
\begin{align*}
\text{Rate } R & \\
\text{Achievable region} & \\
\text{Unachievable region} & \\
0 & \text{Distortion} \\
D_{\text{max}} & \\
\end{align*}
```

- Why is $R(0) = H(p)$ in $X \sim \text{Bernoulli}(p)$ r.v. case above?
- Since Hamming distortion is such that $\{D = 0\} \iff P_e = 0$.
- Key point (again): distortion not necessarily the same as error.
- Achievable rate distortion region is “up-right”-closed. Why?
Distortion vs. Error

- Is it always the case that \( R(D) = H \) at \( D = 0 \)?
- No. If \( D = 0 \) does not require \( P_e = 0 \), then we can compress below the entropy with zero distortion but non-zero error.

\[
\begin{array}{c}
\text{Rate} R \\
\text{H} \\
\text{achievable region} \\
\text{unachievable region} \\
0 \quad \quad \quad \quad \quad D_{\text{max}} \\
\end{array}
\]

- Why is \( R(0) = H(p) \) in \( X \sim \text{Bernoulli}(p) \) r.v. case above?
- Since Hamming distortion is such that \( \{D = 0\} \iff P_e = 0 \).
- Key point (again): distortion not necessarily the same as error.
- Achievable rate distortion region is “up-right”-closed. Why?
- We don’t know if it is always convex yet. Give example of non-convex up-right closed region.
Distortion vs. Error

- Is it always the case that $R(D) = H$ at $D = 0$?
- No. If $D = 0$ does not require $P_e = 0$, then we can compress below the entropy with zero distortion but non-zero error.

Why is $R(0) = H(p)$ in $X \sim \text{Bernoulli}(p)$ r.v. case above?
- Since Hamming distortion is such that $\{D = 0\} \iff P_e = 0$.
- Key point (again): distortion not necessarily the same as error.
- Achievable rate distortion region is “up-right”-closed. Why?
- We don’t know if it is always convex yet. Give example of non-convex up-right closed region. A: staircase down to the right.
Theorem 2.2

Let $R(D)$ be the rate-distortion function and let $R^{(I)}(D)$ be the information rate distortion function. Then

$$R(D) = R^{(I)}(D)$$

(2)

- This means that the minimum coding rate for achieving distortion $D$ is, perhaps now unsurprisingly, $R^{(I)}(D)$. 
Key Theorem

**Theorem 2.2**

Let $R(D)$ be the rate-distortion function and let $R^{(I)}(D)$ be the information rate distortion function. Then

$$R(D) = R^{(I)}(D)$$

(2)

- This means that the minimum coding rate for achieving distortion $D$ is, perhaps now unsurprisingly, $R^{(I)}(D)$.
- Two things to prove: (1) that if $(R, D)$ is achievable, than $R \geq R^{(I)}(D)$, and (2) if $R \geq R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$. 
Key Theorem

Theorem 2.2

Let \( R(D) \) be the rate-distortion function and let \( R^{(I)}(D) \) be the information rate distortion function. Then

\[
R(D) = R^{(I)}(D)
\]  

(2)

- This means that the minimum coding rate for achieving distortion \( D \) is, perhaps now unsurprisingly, \( R^{(I)}(D) \).
- Two things to prove: (1) that if \((R, D)\) is achievable, then \( R \geq R^{(I)}(D) \), and (2) if \( R \geq R^{(I)}(D) \), then there exists a sequence of codes that can achieve rate-distortion pair \((R, D)\).
- For now, let's look at Gaussian sources.
Gaussian Channels

Theorem 2.3

For Gaussian sources $X \sim \mathcal{N}(0, \sigma^2)$ with a squared-error distortion, we have a rate distortion function of the form:

$$R^{(I)}(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & \text{if } 0 \leq D \leq \sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

(3)

Thus, $R^{(I)}(D)$ has the same plot profile that we have seen.
Gaussian Channels

Theorem 2.3

For Gaussian sources \( X \sim \mathcal{N}(0, \sigma^2) \) with a squared-error distortion, we have a rate distortion function of the form:

\[
R^{(I)}(D) = \begin{cases} 
\frac{1}{2} \log \frac{\sigma^2}{D} & \text{if } 0 \leq D \leq \sigma^2 \\
0 & \text{otherwise.}
\end{cases}
\] (3)

Thus, \( R^{(I)}(D) \) has the same plot profile that we have seen.

What happens when \( D \) gets very close to zero and why?
Example: Multiple Gaussians Unequal Noise

- What would be the rate for multiple Gaussians with different noise? I.e., given $X_1:m$ with $X_i \sim \mathcal{N}(0, \sigma_i^2)$ and with $X_i \perp \perp X_j$ for all $i \neq j$, and no requirement for the $\{\sigma_i^2\}$'s to be equal.
Example: Multiple Gaussians Unequal Noise

- What would be the rate for multiple Gaussians with different noise? I.e., given $X_{1:m}$ with $X_i \sim \mathcal{N}(0, \sigma_i^2)$ and with $X_i \perp \perp X_j$ for all $i \neq j$, and no requirement for the $\{\sigma_i^2\}$'s to be equal.

- Overall distortion is of the form $d(x_{1:m}, \hat{x}_{1:m}) = \sum_{i=1}^{m} (x_i - \hat{x}_i)^2$ with $E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D$ where $D$ is overall distortion constraint.

Information rate distortion function has form:

$$R(D) = R(D)$$
Example: Multiple Gaussians Unequal Noise

- What would be the rate for multiple Gaussians with different noise? I.e., given $X_{1:m}$ with $X_i \sim \mathcal{N}(0, \sigma_i^2)$ and with $X_i \perp \perp X_j$ for all $i \neq j$, and no requirement for the $\{\sigma_i^2\}_i$’s to be equal.

- Overall distortion is of the form $d(x_{1:m}, \hat{x}_{1:m}) = \sum_{i=1}^{m} (x_i - \hat{x}_i)^2$ with $E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D$ where $D$ is overall distortion constraint.

- Information rate distortion function has form:

$$R(D) = \min_{f(\hat{x}_{1:m}|x_{1:m}): E[d(X_{1:m}, \hat{X}_{1:m})] \leq D} I(X_{1:m}; \hat{X}_{1:m}) \tag{4}$$
Example: Multiple Gaussians Unequal Noise

- What would be the rate for multiple Gaussians with different noise? I.e., given \(X_{1:m}\) with \(X_i \sim \mathcal{N}(0, \sigma_i^2)\) and with \(X_i \perp \perp X_j\) for all \(i \neq j\), and no requirement for the \(\{\sigma_i^2\}\)'s to be equal.

- Overall distortion is of the form
  \[
  d(x_{1:m}, \hat{x}_{1:m}) = \sum_{i=1}^{m} (x_i - \hat{x}_i)^2
  \]
  with \(E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D\) where \(D\) is overall distortion constraint.

- Information rate distortion function has form:
  \[
  R(D) = \min_{f(\hat{x}_{1:m}|x_{1:m})} \text{I}(X_{1:m}; \hat{X}_{1:m}) \quad (4)
  \]

- We need to know how many bits to allocate to each source symbol (and how much “local distortion to use”) to achieve given overall distortion \(D\). Any guesses?
Example: Multiple Gaussians Unequal Noise

- In general, we need to use KKT conditions to get final distortions, very similar to what we did for multiple Gaussian channel uses.
Example: Multiple Gaussians Unequal Noise

- In general, we need to use KKT conditions to get final distortions, very similar to what we did for multiple Gaussian channel uses.
- We get
Example: Multiple Gaussians Unequal Noise

- In general, we need to use KKT conditions to get final distortions, very similar to what we did for multiple Gaussian channel uses.
- We get

**Theorem 2.4**

Given parallel Gaussian source $X_i \sim \mathcal{N}(0, \sigma_i^2)$ i.i.d., under squared loss $d(x_{1:m}, \hat{x}_{1:m}) = \sum_i (x_i - \hat{x}_i)^2$, we have

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i} = \sum_{i=1}^{m} R_i$$  \hspace{1cm} (5)$$

where

$$D_i = \begin{cases} \lambda & \text{if } \lambda < \sigma_i^2 \ (\Rightarrow R_i > 0) \\ \sigma_i^2 & \text{if } \lambda \geq \sigma_i^2 \ (\Rightarrow R_i = 0) \end{cases} = \min(\lambda, \sigma_i^2)$$  \hspace{1cm} (6)$$

and where $\lambda$ is chosen so that $\sum_i D_i = D$. 
Example: Multiple Gaussians Unequal Noise

Thus, if $\sigma_i^2$ is too small (so that $\lambda > \sigma_i^2$, we allocate no bits to that source symbol.
Example: Multiple Gaussians Unequal Noise

- Thus, if $\sigma_i^2$ is too small (so that $\lambda > \sigma_i^2$, we allocate no bits to that source symbol.
- If $\sigma_i^2$ is sufficiently large, we allocate $R_i = \frac{1}{2} \log \frac{\sigma_i^2}{\lambda}$ bits.
Example: Multiple Gaussians Unequal Noise

- Thus, if $\sigma_i^2$ is too small (so that $\lambda > \sigma_i^2$, we allocate no bits to that source symbol.
- If $\sigma_i^2$ is sufficiently large, we allocate $R_i = \frac{1}{2} \log \frac{\sigma_i^2}{\lambda}$ bits.
- This is the well known reverse water filling argument (or reverse gravity water filling of tanks hanging from a ceiling).
Example: Multiple Gaussians Unequal Noise

- Thus, if $\sigma_i^2$ is too small (so that $\lambda > \sigma_i^2$), we allocate no bits to that source symbol.
- If $\sigma_i^2$ is sufficiently large, we allocate $R_i = \frac{1}{2} \log \frac{\sigma_i^2}{\lambda}$ bits.
- This is the well known reverse water filling argument (or reverse gravity water filling of tanks hanging from a ceiling).
- Let $\hat{\sigma}_i^2 = \sigma_i^2 - D_i$. Water fills tanks hanging from ceiling in reverse gravity, current water line defines $\lambda$ which descents and pushes down any $D_i$ with it. This happens until $\sum_i D_i = D$. 
Converse of Theorem 2.2 states that if \( \{X_i\}_i \) is an i.i.d. source with probability distribution \( X_i \sim p(x) \), and \( d(x, \hat{x}) \) is a distortion measure, than any \((2^{nR}, n)\) code with average distortion

\[
E[d(X^n, \hat{X}^n)] = \frac{1}{n} \sum_{i=1}^{n} E[d(X_i, \hat{X}_i)] \leq D
\]  

has rate \( R \geq R^{(I)}(D) \).
Rate-Distortion Theorem: Converse

- Converse of Theorem 2.2 states that if \( \{X_i\}_i \) is an i.i.d. source with probability distribution \( X_i \sim p(x) \), and \( d(x, \hat{x}) \) is a distortion measure, than any \((2^nR, n)\) code with average distortion

\[
E[d(X^n, \hat{X}^n)] = \frac{1}{n} \sum_{i=1}^{n} E[d(X_i, \hat{X}_i)] \leq D
\]  

has rate \( R \geq R(I)(D) \)

- Alternatively, for any achievable \((R, D)\) pair, we have that \( R \geq R(I)(D) \).
Rate-Distortion Theorem: Converse

Converse of Theorem 2.2 states that if \( \{X_i\}_i \) is an i.i.d. source with probability distribution \( X_i \sim p(x) \), and \( d(x, \hat{x}) \) is a distortion measure, than any \((2^nR, n)\) code with average distortion

\[
E[d(X^n, \hat{X}^n)] = \frac{1}{n} \sum_{i=1}^{n} E[d(X_i, \hat{X}_i)] \leq D
\]  

(7)

has rate \( R \geq R^{(I)}(D) \)

Alternatively, for any achievable \((R, D)\) pair, we have that \( R \geq R^{(I)}(D) \).

This is analogous to saying that if \( P_e \to 0 \), we can’t compress lower than the entropy.
Lemma 2.5

$R^{(I)}(D)$ is: (1) non-increasing in $D$, and (2) convex in $D$. 
Main Theorem: Achievability

Theorem 3.1 (Achievability in 2.2)

Given $X_i$, for $i = 1, \ldots, n$ i.i.d., $\sim p(x)$, and given distortion $d(x, \hat{x})$ and $R(I)(D)$, for any $D$ and any $R > R(I)(D)$, then $(R, D)$ is achievable. I.e. there exists a sequence of $(2^{nR}, n)$ rate-distortion codes with rate $R$ and asymptotic distortion $D$. 
Definition 3.2 (distortion $\epsilon$-typical)

Let $p(x, \hat{x})$ be a joint distortion, $d(x, \hat{x})$ a distortion. For any $\epsilon > 0$, $(x^n, \hat{x}^n)$ (a pair of sequences) is distortion $\epsilon$-typical if all four of the below are true:

\[
\left| - \frac{1}{n} p(x^n) - H(X) \right| < \epsilon \quad \text{$x$-typical} \quad \text{(8)}
\]

\[
\left| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) \right| \leq \epsilon \quad \text{new, "distortion typical"} \quad \text{(11)}
\]
Typicality lives

Definition 3.2 (distortion $\epsilon$-typical)

Let $p(x, \hat{x})$ be a joint distortion, $d(x, \hat{x})$ a distortion. For any $\epsilon > 0$, $(x^n, \hat{x}^n)$ (a pair of sequences) is distortion $\epsilon$-typical if all four of the below are true:

1. $| - \frac{1}{n} p(x^n) - H(X) | < \epsilon \quad x$-typical
2. $| - \frac{1}{n} p(\hat{x}^n) - H(\hat{X}) | < \epsilon \quad \hat{x}$-typical
3. $| - \frac{1}{n} p(x^n, \hat{x}^n) - H(X, \hat{X}) | < \epsilon \quad \text{jointly typical}$
4. $| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) | \leq \epsilon \quad \text{new, "distortion typical"}$

Any $x$ s.t. Equations (8)-(11) are true define the set $A_d, \epsilon \subseteq A_{\epsilon}$. 
Typicality lives

**Definition 3.2 (distortion $\epsilon$-typical)**

Let $p(x, \hat{x})$ be a joint distortion, $d(x, \hat{x})$ a distortion. For any $\epsilon > 0$, $(x^n, \hat{x}^n)$ (a pair of sequences) is distortion $\epsilon$-typical if all four of the below are true:

1. $\left| - \frac{1}{n} p(x^n) - H(X) \right| < \epsilon$ \hspace{1cm} $x$-typical \hspace{1cm} (8)
2. $\left| - \frac{1}{n} p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon$ \hspace{1cm} $\hat{x}$-typical \hspace{1cm} (9)
3. $\left| - \frac{1}{n} p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon$ \hspace{1cm} jointly typical \hspace{1cm} (10)
4. $\left| d(x^n, \hat{x}^n) - E_d(X, \hat{X}) \right| \leq \epsilon$ \hspace{1cm} new, “distortion typical” \hspace{1cm} (11)

Any $x$ s.t. Equations (8)-(11) are true define the set $A_d,\epsilon \subseteq A_{\epsilon}$. 

---

Prof. Jeff Bilmes  
page 14
Typicality lives

Definition 3.2 (distortion $\epsilon$-typical)

Let $p(x, \hat{x})$ be a joint distortion, $d(x, \hat{x})$ a distortion. For any $\epsilon > 0$, $(x^n, \hat{x}^n)$ (a pair of sequences) is distortion $\epsilon$-typical if all four of the below are true:

\[
| - \frac{1}{n} p(x^n) - H(X) | < \epsilon \quad \text{x-typical} \tag{8}
\]

\[
| - \frac{1}{n} p(\hat{x}^n) - H(\hat{X}) | < \epsilon \quad \text{\hat{x}-typical} \tag{9}
\]

\[
| - \frac{1}{n} p(x^n, \hat{x}^n) - H(X, \hat{X}) | < \epsilon \quad \text{jointly typical} \tag{10}
\]

\[
| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) | \leq \epsilon \quad \text{new, “distortion typical”} \tag{11}
\]
Typicality lives

Definition 3.2 (distortion \( \epsilon \)-typical)

Let \( p(x, \hat{x}) \) be a joint distortion, \( d(x, \hat{x}) \) a distortion. For any \( \epsilon > 0 \), \((x^n, \hat{x}^n)\) (a pair of sequences) is distortion \( \epsilon \)-typical if all four of the below are true:

\[
\left| - \frac{1}{n} p(x^n) - H(X) \right| < \epsilon \quad \text{x-typical} \quad (8)
\]

\[
\left| - \frac{1}{n} p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \text{\( \hat{x} \)-typical} \quad (9)
\]

\[
\left| - \frac{1}{n} p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical} \quad (10)
\]

\[
|d(x^n, \hat{x}^n) - E_d(X, \hat{X})| \leq \epsilon \quad \text{new, “distortion typical”} \quad (11)
\]

Any \( x \) s.t. Equations (8)-(11) are true define the set \( A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)} \).
Lemma 3.3

Let \((x_i, \hat{x}_i) \sim p(x, \hat{x})\). Then \(Pr(A_{d,\epsilon}^{(n)}) \to 1\) as \(n \to \infty\).
Probability of typicality

Lemma 3.3

Let \((x_i, \hat{x}_i) \sim p(x, \hat{x})\). Then \(Pr(A_{d,\epsilon}^{(n)}) \to 1 \text{ as } n \to \infty\).

Proof.

Simple application of the law of large numbers, just like before.
Let \((x_i, \hat{x}_i) \sim p(x, \hat{x})\). Then \(\Pr(A_{d, \epsilon}^{(n)}) \to 1\) as \(n \to \infty\).

**Proof.**

Simple application of the law of large numbers, just like before.

Note, this is the same as earlier, except for the distortion but since
\[d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i),\]
we see that \(d(x^n, \hat{x}^n) \to Ed(X, \hat{X})\) by the w.l.l.n. as well.
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as $n \to \infty$. 

...
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as $n \to \infty$.
- Fix $p(\hat{x}|x)$ and then calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$. 

...
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as $n \to \infty$.
- Fix $p(\hat{x}|x)$ and then calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$.
- Chose $\epsilon > 0$ and $\delta > 0$. 

...
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as \( n \to \infty \).
- Fix \( p(\hat{x}|x) \) and then calculate \( p(\hat{x}) = \sum_x p(x)p(\hat{x}|x) \).
- Chose \( \epsilon > 0 \) and \( \delta > 0 \).
- We will show that for any \( R > R^{(I)}(D) \), there exists a code with distortion \( \leq D + \delta \) by generating random codebook.
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as \( n \to \infty \).
- Fix \( p(\hat{x}|x) \) and then calculate \( p(\hat{x}) = \sum_x p(x)p(\hat{x}|x) \).
- Chose \( \epsilon > 0 \) and \( \delta > 0 \).
- We will show that for any \( R > R^{(I)}(D) \), there exists a code with distortion \( \leq D + \delta \) by generating random codebook.

- Generate a random codebook \( C \) (a set of \( 2^{nR} \) codewords, \( \{\hat{x}_1:n(w)\}_{w=1,...,2^{nR}} \)).
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as $n \rightarrow \infty$.
- Fix $p(\hat{x}|x)$ and then calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$.
- Chose $\epsilon > 0$ and $\delta > 0$.
- We will show that for any $R > R(I)(D)$, there exists a code with distortion $\leq D + \delta$ by generating random codebook.
- Generate a random codebook $C$ (a set of $2^{nR}$ codewords, $\{\hat{x}_1:n(w)\}_{w=1,...,2^{nR}}$). So we need $2^{nR}$ length-$n$ sequences, $\hat{x}^n$ drawn i.i.d. $\sim \prod_{i=1}^n p(\hat{x}_i)$.

...
Main Theorem: Achievability

proof of achievability in 2.2.

- We show that we can construct a random code, and use joint typicality to bound the probability of error as $n \to \infty$.
- Fix $p(\hat{x}|x)$ and then calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$.
- Chose $\epsilon > 0$ and $\delta > 0$.
- We will show that for any $R > R(I)(D)$, there exists a code with distortion $\leq D + \delta$ by generating random codebook.
- Generate a random codebook $C$ (a set of $2^{nR}$ codewords, $\{\hat{x}_{1:n}(w)\}_{w=1,\ldots,2^{nR}}$. So we need $2^{nR}$ length-$n$ sequences, $\hat{x}^n$ drawn i.i.d. $\sim \prod_{i=1}^{n} p(\hat{x}_i)$.
- Use $w \in \{1, \ldots, 2^{nR}\}$ to index this codebook, and both the encoder and decoder knows the codebook.
Main Theorem: Achievability

... proof of achievability in 2.2.

Encoding:

- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$.
Main Theorem: Achievability

... proof of achievability in 2.2.

Encoding:

- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_{d, \epsilon}^{(n)}$.
- If such a $w$ does not exist, set $w = 1$. If more than one exists, use least $w$. 

We need $nR$ bits to describe the codewords (since $2^{nR}$ codewords).
Main Theorem: Achievability

... proof of achievability in 2.2.

Encoding:

- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A^{(n)}_{d,\epsilon}$.
- If such a $w$ does not exist, set $w = 1$. If more than one exists, use least $w$.
- We need $nR$ bits to describe the codewords (since $2^{nR}$ codewords).
Main Theorem: Achievability

... proof of achievability in 2.2.

Encoding:

- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A^{(n)}_{d,\epsilon}$.
- If such a $w$ does not exist, set $w = 1$. If more than one exists, use least $w$.
- We need $nR$ bits to describe the codewords (since $2^{nR}$ codewords).

Decoding:

...
Main Theorem: Achievability

... proof of achievability in 2.2.

Encoding:
- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$.
- If such a $w$ does not exist, set $w = 1$. If more than one exists, use least $w$.
- We need $nR$ bits to describe the codewords (since $2^{nR}$ codewords).

Decoding:
- Just produce $\hat{x}^n(w)$. 

...
Main Theorem: Achievability

... proof of achievability in 2.2.

Distortion:

- Average distortion over both codebooks and codewords:

\[
\tilde{D} = E_{X^n,C} d(X^n, \hat{X}^n) = \sum_{C, x^n} \Pr(C) p(x^n) d(x^n, \hat{x}^n)
\]  

(12)
Main Theorem: Achievability

... proof of achievability in 2.2.

Distortion:

- Average distortion over both codebooks and codewords:

\[
\tilde{D} = E_{X^n,C}d(X^n, \hat{X}^n) = \sum_{C,x^n} \Pr(C)p(x^n)d(x^n, \hat{x}^n)
\] (12)

- In the above, we take expectation over both random choice of codebooks \( C = \{ \hat{x}^n(1), \hat{x}^n(2), \ldots, \hat{x}^n(2^{nR}) \} \) based on probability model \( \Pr(C) \), and also random source strings based on \( p(x^n) \).
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

- **Category A:** $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$.  

...
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

- Category A: $x^n$ : $\exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $Pr(A_{d,\epsilon}^{(n)}) \rightarrow 1$. 
Main Theorem: Achievability

...proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:
  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_d^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_d^{(n)}) = 1$.
  - Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences.
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

- Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.

- Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\max}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\max}$.
Main Theorem: Achievability

...proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_d^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.

  - Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\text{max}}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\text{max}}$.

- Total distortion is then

$$\overline{D} \leq D + \epsilon + P_e d_{\text{max}}$$  \hspace{1cm} (13)
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:
  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.
  - Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\text{max}}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\text{max}}$.
- Total distortion is then

$$\bar{D} = Ed(X^n, \hat{X}^n(X^n))$$

(13)

...
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

  - **Category A:** $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.

  - **Category B:** $x^n$ such that there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\text{max}}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\text{max}}$.

- Total distortion is then

$$\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} \quad (13)$$
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.

  - Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\text{max}}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\text{max}}$.

  - Total distortion is then

    $$\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$$ (13)
Main Theorem: Achievability

...proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_d^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_d^{(n)}) = 1$.

  - Category B: $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{\text{max}}$ is the max distortion, then total distortion for this set is $\leq P_e d_{\text{max}}$.

- Total distortion is then

  \[ \bar{D} = Ed(x^n, \hat{x}^n(x^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta \]  

  for any $\delta > 0$ if $\epsilon$ is chosen small, and as long as $P_e \to 0$ as $n \to \infty$.


...
Main Theorem: Achievability

... proof of achievability in 2.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:

  **Category A:** $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) = 1$.

  **Category B:** $x^n$ s.t. there exists no $w$ with $\hat{x}^n(w)$ jointly distortion typical. Let $P_e$ be the probability of these sequences. If $d_{max}$ is the max distortion, then total distortion for this set is $\leq P_e d_{max}$.

- Total distortion is then

\[
\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{max} < D + \delta
\]  

for any $\delta > 0$ if $\epsilon$ is chosen small, and as long as $P_e \to 0$ as $n \to \infty$.

- Trick is to show that $P_e$ gets small fast with $n \to \infty$. 

...
Main Theorem: Achievability

...proof of achievability in 2.2.

General idea first:

- What we will show is that

\[ P_e \leq \Pr((X^n, \hat{X}^n) \notin A^{(n)}_{d,\epsilon}) + e^{-2^n(R-I(X;\hat{X})-3\epsilon)} \]  \hspace{1cm} (14)
Main Theorem: Achievability

... proof of achievability in 2.2.

General idea first:

- What we will show is that

\[ P_e \leq \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) + e^{-2n(R-I(X;\hat{X})-3\epsilon)} \]

where \( <\epsilon \) for \( n \) sufficiently large

(14)
Main Theorem: Achievability

...proof of achievability in 2.2.

General idea first:

What we will show is that

\[ P_e \leq \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) + e^{-2n(R-I(X;\hat{X})-3\epsilon)} \]

\[ < \epsilon \text{ for } n \text{ sufficiently large} \]

exponentially fast to zero if \( R > I + 3\epsilon \)

(14)
Main Theorem: Achievability

...proof of achievability in 2.2.

General idea first:

- What we will show is that

\[
P_e \leq \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) + e^{-2n(R-I(X;\hat{X})-3\epsilon)}
\]

that is, if we can choose \( p(\hat{x}|x) \) to get the \( R(D) \) limit. In such case

\[
I(X;\hat{X}) = R^{(I)}(D).
\]
Main Theorem: Achievability

... proof of achievability in 2.2.

General idea first:

- What we will show is that

\[
P_e \leq \text{Pr}((X^n, \hat{X}^n) \notin A_d^{(n)}) + e^{-2n(R - I(X; \hat{X}) - 3\epsilon)}
\]

that is, if we can chose \( p(\hat{x}|x) \) to get the \( R(D) \) limit. In such case \( I(X; \hat{X}) \to R(D) \).

- This gives

\[
P_e \leq \epsilon + (e^2)^{-n(R - I(X; \hat{X}) - 3\epsilon)}
\]
Main Theorem: Achievability

...proof of achievability in 2.2.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \]  

(16)
Main Theorem: Achievability

... proof of achievability in 2.2.

General idea first:

- This gives

  \[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \]  

- So for any \( \delta > 0 \) \( \exists \epsilon, n \) s.t. over all randomly chosen rate \( R \) codes of block length \( n \), the expected distortion < \( D + \delta \).
Main Theorem: Achievability

...proof of achievability in 2.2.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{n(R-I(X;\hat{X})-3\epsilon)} \]  \hspace{1cm} (16)

- So for any \( \delta > 0 \exists \epsilon, n \) s.t. over all randomly chosen rate \( R \) codes of block length \( n \), the expected distortion \( < D + \delta \).

- This means there must be at least one code \( C^* \) with this rate, block-length, and distortion.
Main Theorem: Achievability

... proof of achievability in 2.2.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \] (16)

- So for any \( \delta > 0 \) \( \exists \epsilon, n \) s.t. over all randomly chosen rate \( R \) codes of block length \( n \), the expected distortion \( < D + \delta \).

- This means there must be at least one code \( C^* \) with this rate, block-length, and distortion.

- \( \delta \) is arbitrary \( \Rightarrow (R, D) \) is achievable if \( R > R^{(I)}(D) \).
Subsidiary Theorems

Theorem 3.4

\[ \forall (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}, \text{ we have} \]

\[ p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \]  
(17)

Proof.

\[ \forall (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}, \text{ we have} \]

\[ p(x^n|\hat{x}^n) \]  
(20)
Subsidiary Theorems

Theorem 3.4

∀(x^n, \hat{x}^n) ∈ A_{d,\epsilon}^{(n)}, we have

\[ p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (17)

Proof.

∀(x^n, \hat{x}^n) ∈ A_{d,\epsilon}^{(n)}, we have

\[ p(x^n | \hat{x}^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} \]  \hspace{1cm} (20)
Subsidiary Theorems

Theorem 3.4

∀(x^n, ̂x^n) ∈ A_d,ε, we have

\[ p( ̂x^n) \geq p( ̂x^n | x^n) 2^{-n(I(X; ̂X)+3\epsilon)} \] (17)

Proof.

∀(x^n, ̂x^n) ∈ A_d,ε, we have

\[ p(x^n | ̂x^n) = \frac{p( ̂x^n, x^n)}{p(x^n)} = p(x^n) \frac{p( ̂x^n, x^n)}{p(x^n)p( ̂x^n)} \] (18)

(20)
Subsidiary Theorems

Theorem 3.4

∀(x^n, ˆx^n) ∈ A^{(n)}_{d,ϵ}, we have

\[ p(ˆx^n) \geq p(ˆx^n|x^n)2^{-n(I(X; ˆX)+3ϵ)} \tag{17} \]

Proof.

∀(x^n, ˆx^n) ∈ A^{(n)}_{d,ϵ}, we have

\[ p(x^n| ˆx^n) = \frac{p(ˆx^n, x^n)}{p(x^n)} = \frac{p(ˆx^n, x^n)}{p(ˆx^n)p(x^n)} \tag{18} \]

\[ \leq p(ˆx^n) \frac{2^{-n(H(X; ˆX)-ϵ)}}{2^{-n(H(X)+ϵ)}2^{-n(H(ˆX)+ϵ)}} \tag{19} \]

\[ \leq p(ˆx^n) \frac{2^{-n(H(X; ˆX)-ϵ)}}{2^{-n(H(X)+ϵ)}2^{-n(H(ˆX)+ϵ)}} \tag{20} \]
Subsidiary Theorems

Theorem 3.4

∀(x^n, ˆx^n) ∈ A_d,ϵ (n), we have

\[ p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{-n(I(X; \hat{X}) + 3\epsilon)} \]  

(17)

Proof.

∀(x^n, ˆx^n) ∈ A_d,ϵ (n), we have

\[ p(x^n | \hat{x}^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = p(x^n) \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \]  

(18)

\[ \leq p(\hat{x}^n) \frac{2^{-n(H(X; \hat{X}) - \epsilon)}}{2^{-n(H(X) + \epsilon)}2^{-n(H(\hat{X}) + \epsilon)}} \]  

(19)

\[ = p(\hat{x}^n) 2^n(I(X; \hat{X}) + 3\epsilon) \]  

(20)
Subsidiary Theorems

Theorem 3.5

For $0 \leq x, y \leq 1$ and $n > 0$, we have

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (21)$$

Proof.

- $f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0.$
Theorem 3.5

For $0 \leq x, y \leq 1$ and $n > 0$, we have

$$(1 - xy)^n \leq 1 - x + e^{-yn}$$

(21)

Proof.

- $f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0$.
- and $f'(y) = -e^{-y} + 1 > 0$ for all $y > 0$. 

...
Subsidiary Theorems

**Theorem 3.5**

For $0 \leq x, y \leq 1$ and $n > 0$, we have

$$ (1 - xy)^n \leq 1 - x + e^{-yn} \quad (21) $$

**Proof.**

- $f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0$.
- and $f'(y) = -e^{-y} + 1 > 0$ for all $y > 0$.
- Thus, $f(y) > 0$ for all $y > 0$. 

...
Subsidiary Theorems

Theorem 3.5

For $0 \leq x, y \leq 1$ and $n > 0$, we have

$$(1 - xy)^n \leq 1 - x + e^{-yn}$$  \hspace{1cm} (21)$$

Proof.

- $f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0$.
- and $f'(y) = -e^{-y} + 1 > 0$ for all $y > 0$.
- Thus, $f(y) > 0$ for all $y > 0$.
- $\Rightarrow$ for $0 \leq y \leq 1$, we have $1 - y \leq e^{-y}$, which is a variational lower bound.
Subsidiary Theorems

... proof continued.

- \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
... proof continued.

- \[ (1 - y)^n \leq e^{-yn} \] which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
Subsidiary Theorems

... proof continued.

- \( \Rightarrow (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
- Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
... proof continued.

- ⇒ \((1 - y)^n \leq e^{-yn}\) which already is the theorem for \(x = 1\).
- Also, theorem is clearly true for \(x = 0\) since \(1 \leq 1 + e^{-yn}\).
- Now, \(g_y(x) = (1 - xy)^n\) is convex in \(x\) since \(\frac{\partial^2 g_y}{\partial x^2} \geq 0\).
- Thus, for all \(0 \leq x \leq 1\):

\[
\Rightarrow (1 - y)^n \leq e^{-yn}
\]
Subsidiary Theorems

... proof continued.

- \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
- Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
- Thus, for all \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n
\]
Subsidiary Theorems

... proof continued.

- $\Rightarrow (1 - y)^n \leq e^{-yn}$ which already is the theorem for $x = 1$.
- Also, theorem is clearly true for $x = 0$ since $1 \leq 1 + e^{-yn}$.
- Now, $g_y(x) = (1 - xy)^n$ is convex in $x$ since $\frac{\partial^2 g_y}{\partial x^2} \geq 0$.
- Thus, for all $0 \leq x \leq 1$:

\[(1 - xy)^n = g_y(x)\]

(26)
Subsidiary Theorems

... proof continued.

- \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
- Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
- Thus, for all \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \tag{22}
\]

(26)
... proof continued.

- \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
- Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
- Thus, for all \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \tag{22}
\]
\[
\leq (1 - x)g_y(0) + xg_y(1) \tag{23}
\]

\[
\leq 1 - x + xe^{-yn} \tag{25}
\]

\[
\leq 1 - x + e^{-yn} \tag{26}
\]
... proof continued.

1. \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
2. Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
3. Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
4. Thus, for all \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \geq (1 - x)g_y(0) + xg_y(1) \leq (1 - x) \cdot 1 + x \cdot (1 - y)^n
\] (22)

(23)

(24)

(25)

(26)
... proof continued.

- \((1 - y)^n \leq e^{-yn}\) which already is the theorem for \(x = 1\).
- Also, theorem is clearly true for \(x = 0\) since \(1 \leq 1 + e^{-yn}\).
- Now, \(g_y(x) = (1 - xy)^n\) is convex in \(x\) since \(\frac{\partial^2 g_y}{\partial x^2} \geq 0\).
- Thus, for all \(0 \leq x \leq 1\):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \\
\leq (1 - x)g_y(0) + xg_y(1) \\
= (1 - x) \cdot 1 + x \cdot (1 - y)^n \\
\leq 1 - x + xe^{-y}
\]
... proof continued.

- \( (1 - y)^n \leq e^{-yn} \) which already is the theorem for \( x = 1 \).
- Also, theorem is clearly true for \( x = 0 \) since \( 1 \leq 1 + e^{-yn} \).
- Now, \( g_y(x) = (1 - xy)^n \) is convex in \( x \) since \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
- Thus, for all \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \\
\leq (1 - x)g_y(0) + xg_y(1) \\
= (1 - x) \cdot 1 + x \cdot (1 - y)^n \\
\leq 1 - x + xe^{-y} \\
\leq 1 - x + e^{-yn}
\]
Main Theorem: Achievability

...proof of achievability in 2.2.

- Next, we calculate $P_e$ for a randomly chosen source sequence and randomly chosen codebook where there exists no codeword that is distortion typical with the source sequence.
Main Theorem: Achievability

... proof of achievability in 2.2.

- Next, we calculate $P_e$ for a randomly chosen source sequence and randomly chosen codebook where there exists no codeword that is distortion typical with the source sequence.

- The set of source sequences s.t. there is at least one codeword in $C$ that is distortion typical with it, is defined as:

$$J(C) = \left\{ x^n : \exists \hat{x}^n \in C \text{ s.t. } (x^n, \hat{x}^n) \in A^{(n)}_{d, \epsilon} \right\}$$ (27)
Main Theorem: Achievability

... proof of achievability in 2.2.

- Next, we calculate $P_e$ for a randomly chosen source sequence and randomly chosen codebook where there exists no codeword that is distortion typical with the source sequence.

- The set of source sequences s.t. there is at least one codeword in $C$ that is distortion typical with it, is defined as:

$$J(C) = \left\{ x^n : \exists \hat{x}^n \in C \text{ s.t. } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)} \right\}$$  \hspace{1cm} (27)

- Then, an expression for $P_e$ follows next ...
Main Theorem: Achievability

\[ R(D) = R(I)(D) \]

... proof of achievability in 2.2.

\[ P_e \]

(31)
Main Theorem: Achievability

\[ P_e = \sum_{C} \Pr(C) \sum_{x^n: x^n \notin J(C)} p(x^n) \]  

(28)

...proof of achievability in 2.2.

\[ Pe = \sum_{C} \Pr(C) \sum_{x^n: x^n \notin J(C)} p(x^n) \]  

(31)
Main Theorem: Achievability

... proof of achievability in 2.2.

\[ P_e = \sum_C \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n) \]  

\[ = \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C) \]  

\[ \text{(31)} \]
Main Theorem: Achievability

... proof of achievability in 2.2.

\[ P_e = \sum_C \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n) \]  \hspace{1cm} (28)

\[ = \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C) \]  \hspace{1cm} (29)

\[ = \sum_{x^n} p(x^n) \begin{cases} \text{total prob of all } 2^{nR} \text{ current } C \text{ code-} \\
\text{words not being distortion typical with current } x^n \text{ (i.e., prob. of choos-} \\
\text{ing codebook not good for current } x^n) \end{cases} \]  \hspace{1cm} (30)

\[ \text{where } q \text{ is the probability that a single } \]  \hspace{1cm} (31)
Main Theorem: Achievability

... proof of achievability in 2.2.

\[ P_e = \sum_C \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n) \]

\[ = \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C) \]  \hspace{1cm} (28)

\[ = \sum_{x^n} p(x^n) \left\{ \begin{array}{l} \text{total prob of all } 2^{nR} \text{ current } C \text{ code-} \\
\text{words not being distortion typical with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \end{array} \right\} \]  \hspace{1cm} (29)

\[ = \sum_{x^n} p(x^n) q^{2^{nR}} \]  \hspace{1cm} (30)

where \( q \) is the probability that a single random codeword is not jointly typical with the current \( x^n \).
Main Theorem: Achievability

... proof of achievability in 2.2.

\[
P_e = \sum_C \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n)
\]

(28)

\[
= \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C)
\]

(29)

\[
= \sum_{x^n} p(x^n) \left\{ \begin{array}{l}
\text{total prob of all } 2^{nR} \text{ current } C \text{ code-}
\text{words not being distortion typical with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \\
\end{array} \right\}
\]

(30)

\[
= \sum_{x^n} p(x^n) q^{2^{nR}}
\]

(31)

where \( q \) is the probability that a single random codeword is not jointly typical with the current } x^n.
Main Theorem: Achievability

... proof of achievability in 2.2.

- Define $K(x^n, \hat{x}^n) = \begin{cases} 1 & \text{if } (x^n, \hat{x}^n) \in A_d^{(n)} \\ 0 & \text{else} \end{cases}$
Main Theorem: Achievability

...proof of achievability in 2.2.

- Define $K(x^n, \hat{x}^n) = \begin{cases} 1 & \text{if } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)} \\ 0 & \text{else} \end{cases}$

- Then

$$q = \Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \Pr(K(x^n, \hat{X}^n) = 1) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)$$

$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X}) + 3\epsilon)} K(x^n, \hat{x}^n)$$

This last line follows from Theorem 3.4.
Main Theorem: Achievability

... proof of achievability in 2.2.

- Define $K(x^n, \hat{x}^n) = \begin{cases} 1 & \text{if } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)} \\ 0 & \text{else} \end{cases}$

- Then

$$q = \Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0)$$

$$= 1 - \Pr(K(x^n, \hat{X}^n) = 1) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n)K(x^n, \hat{x}^n)$$

$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)}K(x^n, \hat{x}^n)$$

- This last line follows from Theorem 3.4.
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e \]

(38)
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2^n R} \]  

(35)
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n)q^{2nR} \]  \hspace{1cm} (35)

\[ \leq \sum_{x^n} \left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x)K(x,\hat{x})\right)^{2nR} \]  \hspace{1cm} (36)

(38)

...
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \]  

\[ \leq \sum_{x^n} \left( 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right)^{2nR} \]

...
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \] (35)

\[ \leq \sum_{x^n} \left( \begin{array}{c}
1 \\
1
\end{array} \right) \left[ 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \right] \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right)^{2nR} \] (36)

(38)

\[ \ldots \]
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2^{nR}} \]  \hspace{1cm} (35)

\[ \leq \sum_{x^n} \left( \begin{array}{c} 1 \\ 1 \end{array} - 2^{-n(I(X;\hat{X})+3\epsilon)} \right) \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \]  \hspace{1cm} (36)

\[ \leq 1 - \sum_{x^n} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \]  \hspace{1cm} (38)
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \]  \hspace{1cm} (35)

\[ \leq \sum_{x^n} \left( \begin{array}{c} 1 \\ 1 \end{array} \right) - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right) \]  \hspace{1cm} (36)

\[ = 1 - \sum_{x^n, \hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) + \exp(-2^{n(R-I(X;\hat{X})-3\epsilon)}) \]  \hspace{1cm} (38)

...
Main Theorem: Achievability

\[ P_e = \sum_{x^n} p(x^n) q^{2^nR} \]  
\[ \leq \sum_{x^n} \left( 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right)^n \]  
\[ = \sum_{x^n} p(x^n) (1 - xy)^n \]
Main Theorem: Achievability

... proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n)q^{2nR} \]  \hspace{1cm} (35)

\[ \leq \sum_{x^n} \left( \frac{1}{1 - 2^{-n(I(X;\hat{X})+3\epsilon)}} \sum_{\hat{x}^n} p(\hat{x}|x)K(x, \hat{x}) \right)^n 2^{nR} \]  \hspace{1cm} (36)

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \leq \sum_{x^n} p(x^n)(1 - x - e^{-yn}) \]  \hspace{1cm} (37)

(38)
Main Theorem: Achievability

...proof of achievability in 2.2.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \] (35)

\[ \leq \sum_{x^n} \left( \frac{1}{1} - \frac{2^{-n(I(X;\hat{X})+3\epsilon)}}{y} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right)^n \] (36)

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \leq \sum_{x^n} p(x^n)(1 - x - e^{-yn}) \] (37)

\[ = 1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) + \exp\left(-2^n(R-I(X;\hat{X})-3\epsilon)\right) \] (38)

...
Main Theorem: Achievability

... proof of achievability in 2.2.

Now

\[ 1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) \]  

is just \( \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) < \epsilon \) and can be made as small as we want by making \( n \) large.
Main Theorem: Achievability

... proof of achievability in 2.2.

- Now

\[
1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n| x^n)K(x^n, \hat{x}^n)
\]

(39)

is just \(\Pr((X^n, \hat{X}^n) \notin A_{d, \epsilon}^{(n)}) < \epsilon\) and can be made as small as we want by making \(n\) large.

- Also

\[
\exp(-2^n(R - I(X; \hat{X}) - 3\epsilon)) \to 0
\]

(40)

if \(R > I(X; \hat{X}) + 3\epsilon\).
Main Theorem: Achievability

... proof of achievability in 2.2.

- Now

\[ 1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) \]  

(39)

is just \( \Pr((X^n, \hat{X}^n) \not\in A_d^{(n)}) < \epsilon \) and can be made as small as we want by making \( n \) large.

- Also

\[ \exp(-2^n(R-I(X;\hat{X})-3\epsilon)) \to 0 \]  

(40)

if \( R > I(X;\hat{X}) + 3\epsilon \). This is true if we chose \( p(\hat{x}|x) \) to be the distribution that achieves the minimum, so that \( R > R(D) \) implying that \( R > I(X;\hat{X}) + 3\epsilon \) for all \( \epsilon \) as small as we want.
Reminder: Geometric of channel capacity

- \( Y = X + Z \) where \( Z \sim \mathcal{N}(0, \sigma^2) \), \( X \sim \mathcal{N}(0, P) \) and \( X \perp Z \).
Reminder: Geometric of channel capacity

- \( Y = X + Z \) where \( Z \sim \mathcal{N}(0, \sigma^2) \), \( X \sim \mathcal{N}(0, P) \) and \( X \perp Z \).
- Typical \( X \)-set \( A^{(n)}_{\epsilon} \) with volume \( \leq 2^{n(h(X)+\epsilon)} \), \( Y \)-given-\( X \) conditional typical set with volume \( \leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)} \), and unconditional typical \( Y \)-set has volume \( \leq 2^{n(h(Y)+\epsilon)} \), and

\[
\begin{align*}
    h(Y) &\leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] & (41) \\
    h(Z) &\leq \frac{1}{2} \log[2\pi e(\sigma^2)] & (42)
\end{align*}
\]
Reminder: Geometric of channel capacity

- $Y = X + Z$ where $Z \sim \mathcal{N}(0, \sigma^2)$, $X \sim \mathcal{N}(0, P)$ and $X \perp Z$.
- Typical $X$-set $A_{\epsilon}^{(n)}$ with volume $\leq 2^{n(h(X)+\epsilon)}$, $Y$-given-$X$ conditional typical set with volume $\leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)}$, and unconditional typical $Y$-set has volume $\leq 2^{n(h(Y)+\epsilon)}$, and

$$h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] \quad (41)$$

$$h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)] \quad (42)$$

- Number of $X$-conditional volumes packable into available $Y$-volume:

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

- \( Y = X + Z \) where \( Z \sim \mathcal{N}(0, \sigma^2) \), \( X \sim \mathcal{N}(0, P) \) and \( X \perp Z \).

- Typical \( X \)-set \( A^{(n)}_\epsilon \) with volume \( \leq 2^n(h(X)+\epsilon) \), \( Y \)-given-\( X \) conditional typical set with volume \( \leq 2^n(h(Y|X)+\epsilon) = 2^n(h(Z)+\epsilon) \), and unconditional typical \( Y \)-set has volume \( \leq 2^n(h(Y)+\epsilon) \), and

\[
    h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] \tag{41}
\]

\[
    h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)] \tag{42}
\]

- Number of \( X \)-conditional volumes packable into available \( Y \)-volume:

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

- 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

\[
    R = \frac{1}{2} \log(1 + P/\sigma^2) \tag{43}
\]
Reminder: Geometric of channel capacity

- \( Y = X + Z \) where \( Z \sim \mathcal{N}(0, \sigma^2) \), \( X \sim \mathcal{N}(0, P) \) and \( X \perp Z \).

- Typical \( X \)-set \( A_{\epsilon}^{(n)} \) with volume \( \leq 2^{n(h(X)+\epsilon)} \), \( Y \)-given-\( X \) conditional typical set with volume \( \leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)} \), and unconditional typical \( Y \)-set has volume \( \leq 2^{n(h(Y)+\epsilon)} \), and

\[
h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] \tag{41}
\]

\[
h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)] \tag{42}
\]

- Number of \( X \)-conditional volumes packable into available \( Y \)-volume:

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

- 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

\[
R = \frac{1}{2} \log(1 + P/\sigma^2) \tag{43}
\]

- Assuming no overlap of volumes which is best we can do, so \( R = C \).
Reminder: Geometric of channel capacity

- Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$. 

For the noise $V(r, n) = \pi^{n/2} \Gamma(n^2 + 1) r^n = 2n^2 \log[e^2 \sigma_r^2] (44)$

This gives $r_\sigma = \Gamma(1/2)(n^2 + 1)(2e^2 \sigma_r^2)^{1/2} \approx (2e^2 n^2)^{1/2} = \sqrt{2}^n (45)$

So, the number of messages $M$ is of the form:

$$M \leq \left( r_\sigma^2 + P \right)^n = (\sigma_r^2 + P \sigma_r^2)^n/2 (46)$$

Goal is to pack as many small spheres in the bit sphere as possible.
Reminder: Geometric of channel capacity

- Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$.
- For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{\frac{n}{2}} \log[2\pi e \sigma^2]$$  \hfill (44)
Reminder: Geometric of channel capacity

- Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$.
- For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{n/2} \log[2\pi e \sigma^2]$$  \hfill (44)

- This gives

$$r_\sigma^2 = \Gamma^{1/2} \left(\frac{n}{2} + 1\right) (2e\sigma^2)^{1/2} \approx (2e\sigma^2 n)^{1/2} = \sqrt{2e\sigma^2 n}$$  \hfill (45)
Reminder: Geometric of channel capacity

- Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$.

- For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{\frac{n}{2}} \log[2\pi e \sigma^2]$$  \hspace{1cm} (44)

- This gives

$$r_{\sigma^2} = \Gamma^{1/2} \left(\frac{n}{2} + 1\right) (2e \sigma^2)^{1/2} \approx (2e \sigma^2 n)^{1/2} = \sqrt{2e \sigma^2 n}$$  \hspace{1cm} (45)

- So, the number of messages $M$ is of the form:

$$M \leq \frac{(r_{\sigma^2} + P)^n}{(r_{\sigma^2})^n} = \left(\frac{\sigma^2 + P}{\sigma^2}\right)^{n/2}$$  \hspace{1cm} (46)
Reminder: Geometric of channel capacity

- Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$.

- For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{\frac{n}{2}} \log[2\pi e \sigma^2]$$ (44)

- This gives

$$r^{\sigma^2} = \Gamma^{1/2}\left(\frac{n}{2} + 1\right)(2e\sigma^2)^{1/2} \approx (2e\sigma^2 n)^{1/2} = \sqrt{2e\sigma^2 n}$$ (45)

- So, the number of messages $M$ is of the form:

$$M \leq \frac{(r^{\sigma^2} + P)^n}{(r^{\sigma^2})^n} = \left(\frac{\sigma^2 + P}{\sigma^2}\right)^{n/2}$$ (46)

- Goal is to pack as many small spheres in the bit sphere as possible.
Geometric of Rate Distortion

- Source $X \sim \mathcal{N}(0, \sigma^2)$. 
Geometric of Rate Distortion

- Source $X \sim \mathcal{N}(0, \sigma^2)$.
- A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathcal{R}^n$ s.t. most $x^n$ are “near” a codeword.
Geometric of Rate Distortion

- Source $X \sim \mathcal{N}(0, \sigma^2)$.
- A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathbb{R}^n$ s.t. most $x^n$ are “near” a codeword.
- Sources live in sphere of radius $(n\sigma^2)^{1/2}$. 
Geometric of Rate Distortion

- Source $X \sim \mathcal{N}(0, \sigma^2)$.
- A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathcal{R}^n$ s.t. most $x^n$ are “near” a codeword.
- Sources live in sphere of radius $(n\sigma^2)^{1/2}$.
- Goal is to use the fewest number of codewords s.t. every source sequence $X^n$ is within $\sqrt{nD}$ of some codeword.
Geometric of Rate Distortion

- Source \( X \sim \mathcal{N}(0, \sigma^2) \).
- A \((2^{nR}, n)\) code with distribution \( < D \) is a set of \( 2^{nR} \) sequences in \( \mathbb{R}^n \) s.t. most \( x^n \) are “near” a codeword.
- Sources live in sphere of radius \((n\sigma^2)^{1/2}\).
- Goal is to use the fewest number of codewords s.t. every source sequence \( X^n \) is within \( \sqrt{nD} \) of some codeword.
- Minimum number of such words is
  \[
  \approx \frac{[(n\sigma^2)^{1/2}]^n}{[nD^{1/2}]^n} = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^{nR(D)} \tag{47}
  \]
  for Gaussian sources.
Geometric of Rate Distortion

Let \( h(X) = \frac{1}{2} \log(2\pi e\sigma^2) \) be a large sphere.
Geometric of Rate Distortion

- Let $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ be a large sphere.
- $h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi e D)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.
Geometric of Rate Distortion

- Let $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ be a large sphere.
- $h(\hat{X} | X) = h(Z) = \frac{1}{2} \log(2\pi e D)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.
- To make sure that no $x^n$ is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
Geometric of Rate Distortion

- Let \( h(X) = \frac{1}{2} \log(2\pi e\sigma^2) \) be a large sphere.
- \( h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi eD) \) is a small sphere, and is a region corresponding to codeword \( \hat{x} \) in the form of a log volume.
- To make sure that no \( x^n \) is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
- The minimum number of such codewords is then:

\[
\geq \frac{2^{nh(X)}}{2^{nh(Z)}} = \frac{2^{\frac{n}{2} \log(2\pi e\sigma^2)}}{2^{n/2 \log(2\pi eD)}} = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^{nR(D)}
\]  

(48)
Geometric of Rate Distortion

- Let $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ be a large sphere.
- $h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi e D)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.
- To make sure that no $x^n$ is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
- The minimum number of such codewords is then:

$$\geq \frac{2^n h(X)}{2^n h(Z)} = \frac{2^n \frac{1}{2} \log(2\pi e \sigma^2)}{2^{n/2} \log(2\pi e D)} = \left(\frac{\sigma^2}{D}\right)^{n/2} = 2^n R(D)$$  \hspace{1cm} (48)
Geometric of Rate Distortion

- Let $h(X) = \frac{1}{2} \log(2\pi e\sigma^2)$ be a large sphere.
- $h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi eD)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.
- To make sure that no $x^n$ is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
- The minimum number of such codewords is then:

$$\geq \frac{2nh(X)}{2nh(Z)} = \frac{2^{n/2} \log(2\pi e\sigma^2)}{2^{n/2} \log(2\pi eD)} = \left(\frac{\sigma^2}{D}\right)^{n/2} = 2^{nR(D)} \quad (48)$$

- Again, meaning $R = \frac{1}{2} \log(\sigma^2/D)$ bits per source symbol to compress with distortion $D$. 
<table>
<thead>
<tr>
<th>Logistics</th>
<th>Review</th>
<th>$R(D) = R^{(T)}(D)$</th>
<th>Geometry</th>
<th>Scratch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scratch Paper
$R(D) = R^{(I)}(D)$
Logistics

Review

Geometry

Scratch

Scratch Paper