EE515a – Information Theory II
Winter Quarter 2014

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Winter Quarter, 2014
http://j.ee.washington.edu/~bilmes/classes/ee515a_winter_2014/

Lecture 23 - Jan 22nd, 2014
Class Road Map - IT-I

- **L19 (1/6)**: Overview, Communications, Gaussian Channel
- **L20 (1/8)**: Gaussian Channel, band limitation, parallel channels, optimization and duality
- **L21 (1/13)**: parallel channels, colored noise, feedback, matrix inequalities
- **L22 (1/15)**: matrix inequalities, rate distortion.
- **– (1/20)**: Monday holiday
- **L23 (1/22)**: rate distortion.

- **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)**:

Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
Read Ch. 9 in our book (Cover & Thomas, “Information Theory”)
Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book
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/)).
Additional Reading on Rate-Distortion Theory


- “Information Geometry and Alternating Minimization Procedures”, Csiszár & Tusnády, 1983

Homework

Homework 1 posted on canvas, due Monday, 1/27/14 at 11:45pm. Only four problems, but these are good problems (and first three are on Gaussian channels so you can start today).
Announcements

- Office hours on Mondays, 3:30-4:30.
- As always, email me if you want to skype/google hangout rather than come to office hours, also at different times.
Combined Bound for Feedback

Corollary 23.2.7

\[ C_{n,FB} \leq \min \{2C_n, C_n + 1/2\} \]  \hspace{1cm} (23.32)

So unfortunately, feedback in this model is not as useful as we might think it would be.
We know that the source compresses down to the entropy $H$, but no further.

We also know that the signal may be sent through the channel at a rate no more than $C$. 
Coding/Compression with distortion

- What if we want to compress $R < H$ or transmit $R > C$? ⇒ Error.
- Similarly, what if we allow for errors, but rather than measure error or no error, measure average distortion.
- But are all errors created equally? Are all errors as bad as others?
- We can measure errors with a distortion function, and we have generalization of the previously stated results.
- Rate-distortion curves with achievable region

![Diagram showing rate-distortion curves with achievable and unachievable regions.](image-url)
Vector Quantization

- We have symbols $X \in \mathcal{X}$ which could be a continuous or a (say big) discrete domain.
- We quantize this region to $\hat{\mathcal{X}}$ where $\hat{\mathcal{X}}$ is discrete and not too big (if $\mathcal{X}$ is discrete, then $|\hat{\mathcal{X}}| \ll |\mathcal{X}|$).

In above, $\hat{\mathcal{X}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_7\}$, $|\hat{\mathcal{X}}| = 7 = M$

There are a set of regions $\mathcal{R}_i$, $i = 1, \ldots, M$, disjoint so that $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for $i \neq j$, and with $\hat{x}_i \in \mathcal{R}_i$ for all $i$. 

Vector Quantization: What determines quality?

- The “quality” of the codebook depends on:
  1. The specific values \( \{\hat{x}_i\}^{M}_{i=1} \), the representatives.
  2. How big \( M \) is, or rather \( R \), with \( M = 2^nR \).
  3. How often each of the values within \( \{\hat{x}_i\}^{M}_{i=1} \) are used, or more accurately, a probability distribution \( p(x) \) over \( \mathcal{X} \).
  4. A measure of how bad it is to represent \( x \in \mathcal{X} \) by \( g(f(x)) \), or a distortion \( d(x, g(f(x))) \).

- Expected distortion \( D = E_{p(x)} d(X, g(f(X))) \).
Rate distortion: set up

- A source produces $x_1, x_2, \cdots \sim p(x)$ based on source distribution $p(x)$ with $x_i \in \mathcal{X}$ for all $i$.
- An encoder $f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$ takes a sequence of source symbols $x_1:n$ and maps them to an integer:
- A decoder $g_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{\mathcal{X}}^n$ takes an integer and maps to quantized vector (i.e., a codeword).
- A distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+$ measures how bad the mapping is. I.e., $d(x, \hat{x})$ measures the “cost” of representing $x \in \mathcal{X}$ by $\hat{x} \in \hat{\mathcal{X}}$.
- Distortion is bounded (sometimes needed) if $\exists d_{\text{max}}$ such that $d_{\text{max}} \triangleq \max_{x, \hat{x}} d(x, \hat{x}) < \infty$.
- Ex: Hamming (probability of error) distortion.

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{otherwise} \end{cases} \quad (23.35)$$

Then $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$
Definition 23.2.7

A \((2^nR, n)\) rate distortion code consists of an encoding function

\[
f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}
\]

(23.36)

and a decoding function

\[
g_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{\mathcal{X}}^n
\]

(23.37)

(Note, \(H(\hat{\mathcal{X}}^n) \leq nR\) since only \(2^{nR}\) different codewords.)

The distortion of this code is

\[
D = Ed(X_{1:n}, g_n(f_n(X_{1:n}))) = \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n})d(x_{1:n}, g_n(f_n(x_{1:n})))
\]

(23.38)
### Achievability and rate-distortion pairs

**Definition 23.2.7**

A rate-distortion pair \((R, D)\) is said to be **achievable** if \(\exists\) a sequence of \((2^{nR}, n)\) codes \((f_n, g_n)\) with

\[
\lim_{n \to \infty} Ed(X_{1:n}, g_n(f_n(X_{1:n}))) \leq D
\]

(23.37)

- So \(D\) is the max allowable distortion for code at this rate \(R\).
- We can make errors, but not too many (bounded average distortion).
- The type of errors we can make is entirely dependent on the distortion function.
- Def: A **rate distortion region** for a source is the closure of achievable rate distortion pairs \((R, D)\)
- Def: A **rate distortion function** \(R(D)\) is the infimum of rates \(R\) such that \((R, D)\) is in rate distortion region. I.e.,

\[
R(D) = \inf \{ R : (R, D) \text{ is achievable} \}
\]

(23.38)
Def: A distortion rate function $D(R)$ is the infimum of distortions $D$ such that $(R, D)$ is in rate distortion region. I.e.,

$$D(R) = \inf \{ D : (R, D) \text{ is achievable} \}$$

(23.37)

The next definition is very important

**Definition 23.2.7**

The “information” rate distortion function $R^{(I)}(D)$ for source $X$ and distortion $d(x, \hat{x})$ is defined as

$$R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum x, \hat{x} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X})$$

(23.38)

Let’s now spend a bit of time getting some intuition on this function.
Intuition: Information Rate Distortion Function

\[ R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \]  

(23.38)

- Related to lossless entropy compression.
- Suppose \( D = Ed(X, \hat{X}) = 0 \) (no distortion) and recall \( d(x, \hat{x}) \geq 0 \) by definition.
- Thus, we must at least have that \( \forall x, \hat{x} : p(x, \hat{x}) > 0, d(x, \hat{x}) = 0 \).
- Consider distortions of the form: \( d(x, \hat{x}) = 0 \Rightarrow \{ x = \hat{x} \} \). For example, \( d(x, \hat{x}) = 1_{\{ x \neq \hat{x} \}} \), or alternatively \( d(x, \hat{x}) = (x - \hat{x})^2 \).
- Thus, \( \forall x, \hat{x} \), if \( p(x, \hat{x}) > 0 \) then \( x = \hat{x} \). Or, the random variables are such that \( X = \hat{X} \).
- Hence,

\[ I(X; \hat{X}) = I(X; X) = H(X) \]  

(23.39)

- And if \( p(x) \) is uniform, then

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

(23.40)
Binary Source $R(D)$, $X \sim \text{Bernoulli}(p)$ r.v.

**Theorem 23.3.1**

The rate-distortion function $R(D)$ for $\text{Bernoulli}(p)$ with 
(Hamming distortion) has the following form:

\[
R(D) = \begin{cases} 
H(p) - H(D) & \text{if } 0 \leq D \leq \min \{p, 1-p\} \\
0 & \text{if } D > \min \{p, 1-p\}
\end{cases}
\]  

(23.1)

$R(D) = \inf \{R : (R, D) \text{ is achievable}\}$

Rate Distortion Region
Binary Source $R(D)$, $X \sim \text{Bernoulli}(p)$ r.v.

- When $D = 0$, minimum rate is the entropy, and can’t compress below the entropy with zero distortion.
Binary Source $R(D)$, $X \sim \text{Bernoulli}(p)$ r.v.

- When $D = 0$, minimum rate is the entropy, and can’t compress below the entropy with zero distortion.
- As $D \uparrow$, we can “compress” more, below the entropy, but we suffer some distortion, and the cyan curve (as we will soon see) gives the limits of achievability.
Binary Source $R(D)$, $X \sim \text{Bernoulli}(p)$ r.v.

- When $D = 0$, minimum rate is the entropy, and can’t compress below the entropy with zero distortion.
- As $D \uparrow$, we can “compress” more, below the entropy, but we suffer some distortion, and the cyan curve (as we will soon see) gives the limits of achievability.
- If we have $D > p$, then random noise will have that distortion, so we can just decode noise and achieve a rate of zero.
Distortion vs. Error

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

- Is it, in general, 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.
**Distortion vs. Error**

- Is it, in general, 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 Rate R vs. Distortion]

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

- Is it, in general, 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]

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

**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.
Distortion vs. Error

- Is it, in general, 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 $\{D = 0\} \Leftrightarrow P_e = 0$.
- Key point (again): distortion not necessarily the same as error.
Distortion vs. Error

- Is it, in general, 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?

Achievable region
unachievable region

Rate $R$
Distortion vs. Error

- Is it, in general, 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 $\{D = 0\} \Leftrightarrow P_e = 0$.
- 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, in general, 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.
Proof of Theorem 23.3.1.

- \( X \sim \mathcal{B}(p) \) and assume, \( D < p \) (for now) and w.l.o.g., that \( 0 \leq D < p \leq 1/2 \) (so \( \min\{p, 1-p\} = p \)).
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- $X \sim \mathcal{B}(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so $\min\{p, 1 - p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$. 
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- \( X \sim \mathcal{B}(p) \) and assume, \( D < p \) (for now) and w.l.o.g., that \( 0 \leq D < p \leq 1/2 \) (so \( \min\{p, 1 - p\} = p \)).

- \( \oplus \) is the xor operator, so \( \{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\} \).

- Approach (like before), find a lower bound on \( I(X; \hat{X}) \) which does not depend on \( p(\hat{x}|x) \), but then find a procedure that “achieves” this lower bound.
Proof of Theorem 23.3.1.

- $X \sim \mathcal{B}(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so $\min\{p, 1-p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. We get:
Proof of Theorem 23.3.1.

- $X \sim \mathcal{B}(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so $\min\{p, 1-p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. We get:

\[ I(X; \hat{X}) \]
Proof of Theorem 23.3.1.

- $X \sim B(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so $\min\{p, 1 - p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. we get:

$$I(X; \hat{X}) = H(X) - H(X|\hat{X})$$  \hspace{1cm} (23.2)
Proof of Theorem 23.3.1.

- $X \sim \mathcal{B}(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so $\min\{p, 1 - p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. we get:

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \quad (23.2)$$
$$\quad = H(X) - H(X \oplus \hat{X}|\hat{X}) \quad (23.3)$$
Proof of Theorem 23.3.1.

- $X \sim B(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so min $\{p, 1 - p\} = p$).
- $\oplus$ is the xor operator, so $\{x \oplus \hat{x} = 1\} \equiv \{x \neq \hat{x}\}$.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. We get:

\[
I(X; \hat{X}) = H(X) - H(X|\hat{X})
\]

(23.2)

\[
= H(X) - H(X \oplus \hat{X}|\hat{X})
\]

(23.3)

\[
\geq H(p) - H(X \oplus \hat{X})
\]

(23.4)
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- $X \sim B(p)$ and assume, $D < p$ (for now) and w.l.o.g., that $0 \leq D < p \leq 1/2$ (so min \{p, 1 − p\} = p).
- $\oplus$ is the xor operator, so \{x $\oplus$ $\hat{x}$ = 1\} ≡ \{x $\neq$ $\hat{x}$\}.
- Approach (like before), find a lower bound on $I(X; \hat{X})$ which does not depend on $p(\hat{x}|x)$, but then find a procedure that “achieves” this lower bound. we get:

\[
I(X; \hat{X}) = H(X) - H(X|\hat{X}) \tag{23.2}
\]

\[
= H(X) - H(X \oplus \hat{X}|\hat{X}) \tag{23.3}
\]

\[
\geq H(p) - H(X \oplus \hat{X}) \tag{23.4}
\]

\[
= H(p) - H(\Pr\{X \neq \bar{X}\}) \tag{23.5}
\]

...
Proof of Theorem 23.3.1.

- \( X \sim \mathcal{B}(p) \) and assume, \( D < p \) (for now) and w.l.o.g., that \( 0 \leq D < p \leq 1/2 \) (so \( \min \{ p, 1 - p \} = p \)).

- \( \oplus \) is the xor operator, so \( \{ x \oplus \hat{x} = 1 \} \equiv \{ x \neq \hat{x} \} \).

- Approach (like before), find a lower bound on \( I(X; \hat{X}) \) which does not depend on \( p(\hat{x}|x) \), but then find a procedure that “achieves” this lower bound. we get:

\[
I(X; \hat{X}) = H(X) - H(X|\hat{X}) \\
= H(X) - H(X \oplus \hat{X}|\hat{X}) \\
\geq H(p) - H(X \oplus \hat{X}) \\
\geq H(p) - H(Pr\{ X \neq \bar{X} \}) \\
\geq H(p) - H(D)
\]
Proof of Theorem 23.3.1.

This last step follows since 1) $D < p \leq 1/2$; 2) that $H(D)$ is a non-negative monotone non-decreasing function of $D$ from $0 \leq D \leq 1/2$; and 3) by the constraint (assumed to be true):

$$Ed(X, \hat{X}) = \Pr(\{X \neq \hat{X}\}) \leq D \quad (23.7)$$

Hence, we have that $H(\Pr(\{X \neq \hat{X}\})) \leq H(D)$. 

...
Proof of Theorem 23.3.1.

- This last step follows since 1) \( D < p \leq 1/2 \); 2) that \( H(D) \) is a non-negative monotone non-decreasing function of \( D \) from \( 0 \leq D \leq 1/2 \); and 3) by the constraint (assumed to be true):

\[
E_d(X, \hat{X}) = \Pr(\{X \neq \bar{X}\}) \leq D
\]  

(23.7)

Hence, we have that \( H(\Pr(\{X \neq \bar{X}\})) \leq H(D) \).

- Now we need to show a distribution \( p(x, \hat{x}) \) that 1) achieves this lower bound and that 2) has tight rate \( R(D) \) with this lower bound \( H(p) - H(D) \).
Proof of Theorem 23.3.1.

- This last step follows since 1) $D < p \leq 1/2$; 2) that $H(D)$ is a non-negative monotone non-decreasing function of $D$ from $0 \leq D \leq 1/2$; and 3) by the constraint (assumed to be true):

$$Ed(X, \hat{X}) = \Pr(\{X \neq \hat{X}\}) \leq D \quad (23.7)$$

Hence, we have that $H(\Pr(\{X \neq \hat{X}\})) \leq H(D)$.

- Now we need to show a distribution $p(x, \hat{x})$ that 1) achieves this lower bound and that 2) has tight rate $R(D)$ with this lower bound $H(p) - H(D)$.

- For case $D = 0$, Hamming requires $P_e = 0$ and $R(0) = H(p)$. 

...
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- This last step follows since 1) $D < p \leq 1/2$; 2) that $H(D)$ is a non-negative monotone non-decreasing function of $D$ from $0 \leq D \leq 1/2$; and 3) by the constraint (assumed to be true):

\[ Ed(X, \hat{X}) = \Pr(\{X \neq \bar{X}\}) \leq D \quad (23.7) \]

Hence, we have that $H(\Pr(\{X \neq \bar{X}\})) \leq H(D)$.

- Now we need to show a distribution $p(x, \hat{x})$ that 1) achieves this lower bound and that 2) has tight rate $R(D)$ with this lower bound $H(p) - H(D)$.

- For case $D = 0$, Hamming requires $P_e = 0$ and $R(0) = H(p)$.

- For case $0 \leq D < p$: we just fix $\Pr(X = 1) = p$ (so $H(X) = H(p)$) and then choose a joint distribution $p(\hat{x}, x)$ that achieves rate $R(D) = H(p) - H(D)$.
$D \leq p < 1/2$ rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- Let $p(x|\hat{x})$ be like a BSC with crossover probability $D$, i.e.,

\[
p(x|\hat{x}) = \begin{cases} 
1 - D & \text{if } x = \hat{x} \\
D & \text{if } x \neq \hat{x}
\end{cases}
\]  

(23.8)
Proof of Theorem 23.3.1.

- Let $p(x|\hat{x})$ be like a BSC with crossover probability $D$, i.e.,

\[
p(x|\hat{x}) = \begin{cases} 
1 - D & \text{if } x = \hat{x} \\
D & \text{if } x \neq \hat{x}
\end{cases}
\] (23.8)

- Then for any $\Pr(\hat{X})$, we have

\[
Ed(X, \hat{X}) = \Pr(X \neq \hat{X}) = D
\]
Proof of Theorem 23.3.1.

Let $p(x|\hat{x})$ be like a BSC with crossover probability $D$, i.e.,

$$p(x|\hat{x}) = \begin{cases} 1 - D & \text{if } x = \hat{x} \\ D & \text{if } x \neq \hat{x} \end{cases} \quad (23.8)$$

Then for any $\Pr(\hat{X})$, we have $Ed(X, \hat{X}) = \Pr(X \neq \hat{X}) = D$, and

$$p \leq D \leq p < \frac{1}{2} \quad (23.10)$$
Proof of Theorem 23.3.1.

- Let \( p(x|\hat{x}) \) be like a BSC with crossover probability \( D \), i.e.,

\[
p(x|\hat{x}) = \begin{cases} 
1 - D & \text{if } x = \hat{x} \\
D & \text{if } x \neq \hat{x}
\end{cases}
\]  

(23.8)

- Then for any \( \Pr(\hat{X}) \), we have \( E_d(X, \hat{X}) = \Pr(X \neq \hat{X}) = D \), and

\[
p = \Pr(X = 1)
\]  

(23.9)

(23.10)
Proof of Theorem 23.3.1.

- Let $p(x | \hat{x})$ be like a BSC with crossover probability $D$, i.e.,

$$p(x | \hat{x}) = \begin{cases} 1 - D & \text{if } x = \hat{x} \\ D & \text{if } x \neq \hat{x} \end{cases} \quad (23.8)$$

- Then for any $\Pr(\hat{X})$, we have $Ed(X, \hat{X}) = \Pr(X \neq \hat{X}) = D$, and

$$p = \Pr(X = 1) \quad (23.9)$$

$$= \Pr(X = 1 | \hat{X} = 0) \Pr(\hat{X} = 0) + \Pr(X = 1 | \hat{X} = 1) \Pr(\hat{X} = 1) \quad (23.10)$$

...
Proof of Theorem 23.3.1.

- Let \( p(x|\hat{x}) \) be like a BSC with crossover probability \( D \), i.e.,

\[
p(x|\hat{x}) = \begin{cases} 
1 - D & \text{if } x = \hat{x} \\
D & \text{if } x \neq \hat{x}
\end{cases}
\]  

(23.8)

- Then for any \( \Pr(\hat{X}) \), we have \( Ed(X, \hat{X}) = Pr(X \neq \hat{X}) = D \), and

\[
p = \Pr(X = 1) = \Pr(X = 1|\hat{X} = 0) \Pr(\hat{X} = 0) + \Pr(X = 1|\hat{X} = 1) \Pr(\hat{X} = 1)
\]

(23.9)

\[
= D(1 - \Pr(\hat{X} = 1)) + (1 - D) \Pr(\hat{X} = 1)
\]  

(23.10)
Proof of Theorem 23.3.1.

Let \( p(x|\hat{x}) \) be like a BSC with crossover probability \( D \), i.e.,

\[
p(x|\hat{x}) = \begin{cases} 
1 - D & \text{if } x = \hat{x} \\
D & \text{if } x \neq \hat{x}
\end{cases}
\]  
(23.8)

Then for any \( \Pr(\hat{X}) \), we have \( Ed(X, \hat{X}) = \Pr(X \neq \hat{X}) = D \), and

\[
p = \Pr(X = 1) = \Pr(X = 1|\hat{X} = 0) \Pr(\hat{X} = 0) + \Pr(X = 1|\hat{X} = 1) \Pr(\hat{X} = 1)
\]

\[
= D(1 - \Pr(\hat{X} = 1)) + (1 - D) \Pr(\hat{X} = 1)
\]  
(23.10)

Solving for \( \Pr(\hat{X} = 1) \) we get

\[
\Pr(\hat{X} = 1) = \frac{p - D}{1 - 2D}
\]  
(23.11)

...
Proof of Theorem 23.3.1.

- Let $p(x|\hat{x})$ be like a BSC with crossover probability $D$, i.e.,
  \[ p(x|\hat{x}) = \begin{cases} 
  1 - D & \text{if } x = \hat{x} \\
  D & \text{if } x \neq \hat{x}
  \end{cases} \] (23.8)

- Then for any $\Pr(\hat{X})$, we have $Ed(X, \hat{X}) = \Pr(X \neq \hat{X}) = D$, and
  \[ p = \Pr(X = 1) \]
  \[ = \Pr(X = 1|\hat{X} = 0) \Pr(\hat{X} = 0) + \Pr(X = 1|\hat{X} = 1) \Pr(\hat{X} = 1) \]
  \[ = D(1 - \Pr(\hat{X} = 1)) + (1 - D) \Pr(\hat{X} = 1) \] (23.10)

- Solving for $\Pr(\hat{X} = 1)$ we get
  \[ \Pr(\hat{X} = 1) = \frac{p - D}{1 - 2D} \leq p \]
  if $0 \leq D < p \leq 1/2$ (23.11)

...
Aside: rate-distortion for Bernoulli r.v.

To show this last inequality is true, assume $D < p \leq 1/2$ and that 

$$\text{Pr}(\hat{X} = 1) = \frac{p - D}{1 - 2D}.$$
Aside: rate-distortion for Bernoulli r.v.

To show this last inequality is true, assume $D < p \leq 1/2$ and that $\Pr(\hat{X} = 1) = \frac{p-D}{1-2D}$.

Then $p \geq \frac{(p - D)}{(1 - 2D)}$ since

$$\frac{p-D}{1-2D} - p = \frac{p - D - p + 2Dp}{1 - 2D} = \frac{D(2p - 1)}{1 - 2D} \leq 0 \quad (23.12)$$
Proof of Theorem 23.3.1.

- So we have right distortion $E_d(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$. 

...
Proof of Theorem 23.3.1.

- So we have right distortion $Ed(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.
- Starting with the lower bound, we get:

\begin{align}
H(p) - H(D) &\leq I(X; \hat{X}) \\
&= H(X) - H(X | \hat{X}) \\
&= H(p) - H(D) \tag{23.14}
\end{align}

...
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- So we have right distortion $E_d(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.
- Starting with the lower bound, we get:

$$H(p) - H(D)$$

(23.14)
Proof of Theorem 23.3.1.

- So we have right distortion $Ed(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.
- Starting with the lower bound, we get:

  $$H(p) - H(D) \leq I(X; \hat{X})$$

  (23.14)
Proof of Theorem 23.3.1.

- So we have right distortion $Ed(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.

- Starting with the lower bound, we get:

\[
H(p) - H(D) \leq I(X; \hat{X}) = H(X) - H(X|\hat{X})
\]

(23.13)

(23.14)

...
Proof of Theorem 23.3.1.

- So we have right distortion $Ed(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.
- Starting with the lower bound, we get:

$$H(p) - H(D) \leq I(X; \hat{X}) = H(X) - H(X|\hat{X})$$

(23.13)

$$= H(p) - H(D)$$

(23.14)
Proof of Theorem 23.3.1.

- So we have right distortion $Ed(X, \hat{X}) = D$, we just need to show that $I(X; \hat{X}) = H(p) - H(D)$.

- Starting with the lower bound, we get:

$$H(p) - H(D) \leq I(X; \hat{X}) = H(X) - H(X|\hat{X})$$

$$= H(p) - H(D)$$  \hfill (23.13)

$$= H(p) - H(D)$$  \hfill (23.14)

- And thus $I(X; \hat{X}) = H(p) - H(D)$. ...
Proof of Theorem 23.3.1.

- If $D \geq p$ then we must show that we can achieve this distortion with a rate of $R = 0$.
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

- If $D \geq p$ then we must show that we can achieve this distortion with a rate of $R = 0$.
- To do this, hard-code $\hat{X} = 0$ (i.e., $P(\hat{X} = 0) = 1$) which can be done with a rate of $R = 0$. 

\[ R(D) = R(I)(D) \]
Proof of Theorem 23.3.1.

- If $D \geq p$ then we must show that we can achieve this distortion with a rate of $R = 0$.
- To do this, hard-code $\hat{X} = 0$ (i.e., $P(\hat{X} = 0) = 1$) which can be done with a rate of $R = 0$.
- Then $Ed(X, \hat{X}) = Pr(X \neq \hat{X}) = Pr(\text{error})$ and with $\hat{X} = 0$, we see that

\[
Pr(\text{error}) = Pr(X = 0)Pr(\hat{X} = 1|X = 0) + Pr(X = 1)Pr(\hat{X} = 0|X = 1)
\]

\[
= Pr(X = 0) \times 0 + Pr(X = 1) \times 1
\]

\[
= p
\]
Theorem 23.4.1

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

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

(23.19)

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 $R > R^{(I)}(D)$, and (2) if $R > R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$. 

Like in channel capacity and entropy compression case, what happens at $R = R^{(I)}(D)$ depends on the very specific case that one is analyzing.

Before proving this key theorem, let's look at Gaussian sources.
Key Theorem

**Theorem 23.4.1**

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

$$R(D) = R^{(I)}(D)$$  \hspace{0.5cm} (23.19)

- 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 $R > R^{(I)}(D)$, and (2) if $R > R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$. 

Like in channel capacity and entropy compression case, what happens at $R = R^{(I)}(D)$ depends on the very specific case that one is analyzing.

Before proving this key theorem, let's look at Gaussian sources.
Theorem 23.4.1

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

$$R(D) = R^{(I)}(D) \tag{23.19}$$

- 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 $R > R^{(I)}(D)$, and (2) if $R > R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$.
- Like in channel capacity and entropy compression case, what happens at $R = R^{(I)}(D)$ depends on the very specific case that one is analyzing.
Key Theorem

Theorem 23.4.1

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

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

(23.19)

- 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 $R > R^{(I)}(D)$, and (2) if $R > R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$.
- Like in channel capacity and entropy compression case, what happens at $R = R^{(I)}(D)$ depends on the very specific case that one is analyzing.
- Before proving this key theorem, let’s look at Gaussian sources.
Gaussian Channels

Theorem 23.5.1

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

$$R^{(I)}(D) = \begin{cases} 
\frac{1}{2} \log \frac{\sigma^2}{D} & \text{if } 0 \leq D \leq \sigma^2 \\
0 & \text{otherwise.}
\end{cases} \tag{23.20}$$

Thus, $R^{(I)}(D)$ has the same plot profile that we have seen.
Theorem 23.5.1

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

$$R^{(I)}(D) = \begin{cases} 
\frac{1}{2} \log \frac{\sigma^2}{D} & \text{if } 0 \leq D \leq \sigma^2 \\
0 & \text{otherwise}
\end{cases} \quad (23.20)$$

- Thus, $R^{(I)}(D)$ has the same plot profile that we have seen.
- What happens when $D$ gets very close to zero and why?
**Theorem 23.5.1**

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

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

(23.20)

- Thus, $R^{(I)}(D)$ has the same plot profile that we have seen.
- What happens when $D$ gets very close to zero and why?
- **A:** basically, at zero distortion we are needing to code an infinite resolution Gaussian which will require an infinite rate (infinite precision), similar to what happened with the Gaussian channel without a source power constraint.
Proof of Theorem 23.5.1

Proof.

We have that

\[
R(I)(D) = \min_{f(\hat{x}|x): E(\hat{X} - X)^2 \leq D} I(X; \hat{X})
\]  
(23.21)

So we lower bound \( I(X; \hat{X}) \) under \( E(\hat{X} - X)^2 \leq D \):

\[
I(X; \hat{X}) = h(X) - h(X|\hat{X}) = \frac{1}{2} \log((2\pi e)\sigma^2) - h(X - \hat{X}|\hat{X})
\]

\[
\geq \frac{1}{2} \log((2\pi e)\sigma^2) - h(X - \hat{X})
\]  
(23.22)

\[
\geq \frac{1}{2} \log((2\pi e)\sigma^2) - h(\mathcal{N}(0, E(X - \hat{X})^2))
\]  
(23.23)

\[
\geq \frac{1}{2} \log((2\pi e)\sigma^2) - \frac{1}{2} \log((2\pi e)D)
\]  
(23.24)

\[
= \frac{1}{2} \log(\sigma^2 / D)
\]  
(23.25)

\[ \text{...} \]
Proof of Theorem 23.5.1

Proof.

Thus, $R(D) \geq \frac{1}{2} \log(\sigma^2/D)$
Proof of Theorem 23.5.1

Proof.

Thus, \( R(D) \geq \frac{1}{2} \log(\sigma^2 / D) \)

Like before, we construct \( f(\hat{x}) \) and \( f(x|\hat{x}) \) to give a joint that achieves equality/tightness in the lower bound & distortion \( D \).
Proof of Theorem 23.5.1

Proof.

Thus, \( R(D) \geq \frac{1}{2} \log(\sigma^2/D) \)

Like before, we construct \( f(\hat{x}) \) and \( f(x|\hat{x}) \) to give a joint that achieves equality/tightness in the lower bound & distortion \( D \).

We define it as follows, where \( \hat{X} \perp Z \):

\[
\begin{align*}
X &= \hat{X} + Z, & \hat{X} &\sim \mathcal{N}(0, \sigma^2 - D), & Z &\sim \mathcal{N}(0, D), \\
Z &\sim \mathcal{N}(0, D) \\
\hat{X} &\sim \mathcal{N}(0, \sigma^2 - D) & + & X &\sim \mathcal{N}(0, \sigma^2)
\end{align*}
\]

Note also, \( E(X - \hat{X})^2 = EZ^2 \leq D \) so we have achieved the distortion constraint.
Proof of Theorem 23.5.1

Proof.

Thus, $R(D) \geq \frac{1}{2} \log(\sigma^2/D)$

Like before, we construct $f(\hat{x})$ and $f(x|\hat{x})$ to give a joint that achieves equality/tightness in the lower bound & distortion $D$.

We define it as follows, where $\hat{X} \perp \perp Z$:

$$X = \hat{X} + Z, \quad \hat{X} \sim \mathcal{N}(0, \sigma^2 - D), \quad Z \sim \mathcal{N}(0, D),$$

Note also, $E(X - \hat{X})^2 = EZ^2 \leq D$ so we have achieved the distortion constraint. We have:

$$I(X; \hat{X}) = h(X) - h(X|\hat{X}) = \frac{1}{2} \log(2\pi e)\sigma^2 - h(Z)$$ \hspace{1cm} (23.26)

$$= \frac{1}{2} \log(\sigma^2/D)$$ \hspace{1cm} (23.27)
Proof of Theorem 23.5.1

Proof.

- If $D > \sigma^2$ we can choose $\hat{x} = 0$ w.p.1 for a zero rate code, achieving a distortion of $\sigma^2$.

- To summarize, we then get:

$$R(D) = \max \left\{ \frac{1}{2} \log \frac{\sigma^2}{D}, 0 \right\}$$  \hspace{1cm} (23.28)

- As always, this rate is achieved by longer block lengths, so short block lengths would not get this rate.
Proof of Theorem 23.5.1

Proof.

- If $D > \sigma^2$ we can choose $\hat{x} = 0$ w.p.1 for a zero rate code, achieving a distortion of $\sigma^2$.
- To summarize, we then get:

$$R(D) = \max \left\{ \frac{1}{2} \log \frac{\sigma^2}{D}, 0 \right\}$$  \hspace{1cm} (23.28)

As always, this rate is achieved by longer block lengths, so short block lengths would not get this rate.

- Note again to keep in mind: if $D > \sigma^2$ then we can use a rate of $R = 0$. If $D < \sigma^2$ then we need to allocate some bits.
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.
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
  \[
  E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D \text{ where } D \text{ is overall distortion constraint.}
  \]
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:

$$R(D) = \min_{f(\hat{x}_1:m|x_1:m): E[d(X_1:m, \hat{X}_1:m)] \leq D} I(X_1:m; \hat{X}_1:m)$$ (23.29)
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:

$$R(D) = \min_{f(\hat{x}_{1:m}|x_{1:m}): E[d(X_{1:m}, \hat{X}_{1:m})] \leq D} I(X_{1:m}; \hat{X}_{1:m})$$ (23.29)

- 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

- We expand MI as follows:

\[ I(X_{1:m}; \hat{X}_{1:m}) = h(X_{1:m}) - h(X_{1:m} | \hat{X}_{1:m}) \]

\[ = \sum_i h(X_i) - \sum_i h(X_i | \hat{X}_{1:m}, X_{1:i-1}) \]  

\[ \geq \sum_i h(X_i) - \sum_i h(X_i | \hat{X}_i) \]  

\[ = \sum_i I(X_i; \hat{X}_i) \]  

\[ \geq \sum_i R(D_i) = \sum_i \max \left\{ \frac{1}{2} \log \frac{\sigma^2}{D_i}, 0 \right\} \]  

where \( D_i = E(X_i - \hat{X}_i)^2 \).
Example: Multiple Gaussians Unequal Noise

To achieve equality, we set

\[
f(\hat{x}_{1:m} | x_{1:m}) = \prod_{i} f(\hat{x}_i | x_i)
\]  \hspace{1cm} (23.35)
Example: Multiple Gaussians Unequal Noise

- To achieve equality, we set

\[ f(\hat{x}_{1:m}|x_{1:m}) = \prod_i f(\hat{x}_i|x_i) \]  

(23.35)

- And also

\[ \hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i) = \mathcal{N}(0, \hat{\sigma}_i^2) \]  

(23.36)
Example: Multiple Gaussians Unequal Noise

To achieve equality, we set

\[ f(\hat{x}_{1:m} | x_{1:m}) = \prod_i f(\hat{x}_i | x_i) \]  (23.35)

And also

\[ \hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i) = \mathcal{N}(0, \hat{\sigma}_i^2) \]  (23.36)

Thus, the problem becomes:
Example: Multiple Gaussians Unequal Noise

The problem becomes:

\[ R(D) = \min_{\{D_i\}_i: \sum_i D_i = D} \sum_{i=1}^{m} \max \left\{ \frac{1}{2} \log \frac{\sigma_i^2}{D_i}, 0 \right\} \]

(23.37)
Example: Multiple Gaussians Unequal Noise

- The problem becomes:

\[
R(D) = \min_{\{D_i\}_i : \sum_i D_i = D} \sum_{i=1}^m \max \left\{ \frac{1}{2} \log \frac{\sigma_i^2}{D_i}, 0 \right\}
\]

(23.37)

- This is a convex minimization problem, and strong duality and KKT necessary optimality conditions hold, and can be written as:

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^m R_i \\
\text{over} & \quad \{R_i\}_i, \{D_i\}_i \\
\text{subject to} & \quad \sum_i D_i = D \\
& \quad R_i \geq \frac{1}{2} \log \frac{\sigma_i^2}{D_i} \quad \forall i \\
& \quad R_i \geq 0 \quad \forall i 
\end{align*}
\]

(23.38) (23.39) (23.40) (23.41)
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i \implies R_i > 0$, then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

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

\[
J(D) = \frac{1}{2} \sum_{i=1}^{m} \left( \ln \sigma_i^2 - D_i + \lambda D_i \right)
\]

\[
\partial J \partial D_i = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0
\]

\[
D_i = \frac{1}{2\lambda} \Rightarrow \lambda' = \frac{D}{m} < \sigma_i^2 \text{ for all } i,
\]

so this says that in this case we use the same distortion amount for all source symbols, which is feasible when $\lambda' = \frac{D}{m} < \sigma_i^2$ for all $i$. That is, when $\lambda' = \frac{D}{m} < \sigma_i^2$ min, then $\lambda < \sigma_i^2$ for all $i$, and have $R_i > 0$ (i.e., each source gets to transmit) $\forall i$. As total distortion $D$ increases, $\lambda$ will also increase eventually hitting one or more of the $\sigma_i^2$ values (i.e., $\lambda = \min_i \sigma_i^2$). This will shut off some sources as we'll get $R_i = 0$ for those.
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ ($\Rightarrow R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

\[
J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right)
\]  

(23.42)

\[
\Rightarrow \partial J \partial D_i = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0
\]  

(23.43)

\[
D_i = \frac{1}{2\lambda} = \lambda' \forall i
\]  

(23.44)

So this says that in this case we use the same distortion amount for all source symbols, which is feasible when $D_i < \sigma_i^2$ for all $i$. That is, when $\lambda' = D/m < \sigma_{\min}^2$ where $\sigma_{\min}^2 = \min_i \sigma_i^2$, then $\lambda < \sigma_i^2$ for all $i$, and have $R_i > 0$ (i.e., each source gets to transmit) $\forall i$. As total distortion $D$ increases, $\lambda$ will also increase eventually hitting one or more of the $\sigma_i^2$ values (i.e., $\lambda = \min_i \sigma_i^2$). This will shut off some sources as we’ll get $R_i = 0$ for those.
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ ($\Rightarrow R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

$$J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right)$$

(23.42)

$$\Rightarrow \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0$$

(23.43)

$$\Rightarrow D_i = \frac{1}{2\lambda}$$

(23.44)

So this says that in this case we use the same distortion amount for all source symbols, which is feasible when $D_i < \sigma_i^2$ for all $i$. That is, when $\lambda' = \frac{D}{m} < \sigma_{\min}^2$ where $\sigma_{\min}^2 = \min_i \sigma_i^2$, then $\lambda < \sigma_i^2$ for all $i$, and have $R_i > 0$ (i.e., each source gets to transmit) $\forall i$. As total distortion $D$ increases, $\lambda$ will also increase eventually hitting one or more of the $\sigma_i^2$ values (i.e., $\lambda = \min_i \sigma_i^2$). This will shut off some sources as we’ll get $R_i = 0$ for those.
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ ($\Rightarrow R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

$$J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right)$$

(23.42)

$$\Rightarrow \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0$$

(23.43)

$$\Rightarrow D_i = \frac{1}{(2\lambda)} = \lambda' \quad \forall i$$

(23.44)
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ (⇒ $R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

\[
J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right) \tag{23.42}
\]

⇒ \[
\frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0 \tag{23.43}
\]

⇒ \[
D_i = 1/(2\lambda) = \lambda' \quad \forall i \tag{23.44}
\]

- So this says that in this case we use the same distortion amount for all source symbols, which is feasible when $D_i < \sigma_i^2$ for all $i$. 

Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ ($\Rightarrow R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

\[
J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma^2}{D_i} + \lambda D_i \right)
\]  
(23.42)

\[
\Rightarrow \quad \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0
\]  
(23.43)

\[
\Rightarrow \quad D_i = \frac{1}{(2\lambda)} = \lambda' \quad \forall i
\]  
(23.44)

- So this says that in this case we use the same distortion amount for all source symbols, which is feasible when $D_i < \sigma_i^2$ for all $i$.

- That is, when $\lambda' = D/m < \sigma_{\text{min}}^2$, where $\sigma_{\text{min}}^2 = \min_i \sigma_i^2$, then $\lambda < \sigma_i^2$ for all $i$, and can have $R_i > 0$ (i.e., each source gets to transmit) $\forall i$. 

Prof. Jeff Bilmes
Example: Multiple Gaussians Unequal Noise

- If $D_i < \sigma_i^2$ for all $i$ ($\Rightarrow R_i > 0$), then this simplifies and we can avoid the constraints on $R_i$ (and $R_i$ altogether, as all rates are guaranteed positive), yielding Lagrangian

$$J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right)$$  \hspace{1cm} (23.42)

$$\Rightarrow \quad \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0$$  \hspace{1cm} (23.43)

$$\Rightarrow \quad D_i = \frac{1}{(2\lambda)} = \lambda' \quad \forall i$$  \hspace{1cm} (23.44)

- So this says that in this case we use the same distortion amount for all source symbols, which is feasible when $D_i < \sigma_i^2$ for all $i$.

- That is, when $\lambda' = D/m < \sigma_{\min}^2$ where $\sigma_{\min}^2 = \min_i \sigma_i^2$, then $\lambda < \sigma_i^2$ for all $i$, and can have $R_i > 0$ (i.e., each source gets to transmit) $\forall i$.

- As total distortion $D$ increases, $\lambda$ will also increase eventually hitting one or more of the $\sigma_i^2$ values (i.e., $\lambda = \min_i \sigma_i^2$).
Example: Multiple Gaussians Unequal Noise

- If \( D_i < \sigma_i^2 \) for all \( i \) (\( \Rightarrow R_i > 0 \)), then this simplifies and we can avoid the constraints on \( R_i \) (and \( R_i \) altogether, as all rates are guaranteed positive), yielding Lagrangian

\[
J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right) \quad (23.42)
\]

\[
\Rightarrow \quad \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0 \quad (23.43)
\]

\[
\Rightarrow \quad D_i = 1/(2\lambda) = \lambda' \quad \forall i \quad (23.44)
\]

- So this says that in this case we use the same distortion amount for all source symbols, which is feasible when \( D_i < \sigma_i^2 \) for all \( i \).
- That is, when \( \lambda' = D/m < \sigma_{\min}^2 \) where \( \sigma_{\min}^2 = \min_i \sigma_i^2 \), then \( \lambda < \sigma_i^2 \) for all \( i \), and can have \( R_i > 0 \) (i.e., each source gets to transmit) \( \forall i \).
- As total distortion \( D \) increases, \( \lambda' \) will also increase eventually hitting one or more of the \( \sigma_i^2 \) values (i.e., \( \lambda' = \min_i \sigma_i^2 \)). This will shut off some sources as we’ll get \( R_i = 0 \) for those.
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

\[ R(D) = R^{(I)}(D) \]
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 23.6.1**

*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*

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i} = \sum_{i=1}^{m} R_i \quad (23.45)$$

*where*

$$D_i = \begin{cases} 
\lambda & \text{if } \lambda < \sigma_i^2 (\Rightarrow R_i > 0) \\
\sigma_i^2 & \text{if } \lambda \geq \sigma_i^2 (\Rightarrow R_i = 0) 
\end{cases} = \min(\lambda, \sigma_i^2) \quad (23.46)$$

*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 $\sigma^2_i$ is too small (so that $\lambda > \sigma^2_i$), we allocate no bits to that source symbol.
- If $\sigma^2_i$ is sufficiently large, we allocate $R_i = \frac{1}{2} \log \frac{\sigma^2_i}{\lambda}$ bits.
- This is the well known reverse water filling argument (or reverse gravity water filling of tanks hanging from a ceiling).
- Let $\hat{\sigma}^2_i = \sigma^2_i - D_i$. Water fills tanks hanging from ceiling in reverse gravity, current water line defines $\lambda$ which descends and pushes down any $D_i$ with it. This happens until $\sum_i D_i = D$. 

\[ \begin{align*} 
R_i &= \frac{1}{2} \log \frac{\sigma^2_i}{\lambda} \\
\sum_i D_i &= D 
\end{align*} \]
Rate-Distortion Theorem: Converse

- Converse of Theorem 23.4.1 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 \( R > R^{(I)}(D) \)
Converse of Theorem 23.4.1 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

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

(23.47)

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

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

- Converse of Theorem 23.4.1 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

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

(23.47)

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

- 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 23.7.1

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

Proof.

First, as $D \uparrow$, we are taking the minimum over a larger set so necessarily $R^{(I)}(D) \downarrow$ as $D \uparrow$. 

\begin{align*}
R(D) &= R^{(I)}(D) \\
&= R(I)(D)
\end{align*}
Lemma 23.7.1

\(R(I)(D)\) is: (1) non-increasing in \(D\), and (2) convex in \(D\).

Proof.

- First, as \(D \uparrow\), we are taking the minimum over a larger set so necessarily \(R(I)(D) \downarrow\) as \(D \uparrow\).

- Now, consider \((R_1, D_1)\) and \((R_2, D_2)\) on \(R-D\) curve of \(R(I)(D)\) with, respectively, \(p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x)\) and \(p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x)\) being two distributions that achieve those pairs respectively.
Lemma 23.7.1

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

Proof.

- First, as \( D \uparrow \), we are taking the minimum over a larger set so necessarily \( R^{(I)}(D) \downarrow \) as \( D \uparrow \).

- Now, consider \((R_1, D_1)\) and \((R_2, D_2)\) on \( R-D \) curve of \( R^{(I)}(D) \) with, respectively, \( p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x) \) and \( p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x) \) being two distributions that achieve those pairs respectively.

- Mix them, \( p_\lambda = \lambda p_1 + (1 - \lambda)p_2 \) which achieves distortion \( D_\lambda = \lambda D_1 + (1 - \lambda)D_2 = \sum_{x,\hat{x}} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x}) \).

Recall mutual information is convex in conditional distribution for fixed \( p(x) \). Hence, \( I_{p_\lambda}(X;\hat{X}) \leq \lambda I_{p_1}(X;\hat{X}) + (1 - \lambda)I_{p_2}(X;\hat{X}) \).
Lemma 23.7.1

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

Proof.

- First, as \( D \uparrow \), we are taking the minimum over a larger set so necessarily \( R^{(I)}(D) \downarrow \) as \( D \uparrow \).

- Now, consider \((R_1, D_1)\) and \((R_2, D_2)\) on \( R-D \) curve of \( R^{(I)}(D) \) with, respectively, \( p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x) \) and \( p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x) \) being two distributions that achieve those pairs respectively.

- Mix them, \( p_\lambda = \lambda p_1 + (1-\lambda)p_2 \) which achieves distortion \( D_\lambda = \lambda D_1 + (1-\lambda)D_2 = \sum_{x,\hat{x}} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x}) \).

- Recall mutual information is convex in conditional distribution for fixed \( p(x) \).
Lemma 23.7.1

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

Proof.

- First, as \( D \uparrow \), we are taking the minimum over a larger set so necessarily \( R^{(I)}(D) \downarrow \) as \( D \uparrow \).

- Now, consider \((R_1, D_1)\) and \((R_2, D_2)\) on \( R-D \) curve of \( R^{(I)}(D) \) with, respectively, \( p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x) \) and \( p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x) \) being two distributions that achieve those pairs respectively.

- Mix them, \( p_\lambda = \lambda p_1 + (1 - \lambda) p_2 \) which achieves distortion \( D_\lambda = \lambda D_1 + (1 - \lambda) D_2 = \sum_{x, \hat{x}} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x}) \).

- Recall mutual information is convex in conditional distribution for fixed \( p(x) \).

- Hence, \( I_{p_\lambda}(X; \hat{X}) \leq \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda) I_{p_2}(X; \hat{X}) \).
Proof.

- Therefore,

\[ R^{(I)}(D_{\lambda}) \leq I_{p_{\lambda}}(X; \hat{X}) \leq \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda)I_{p_2}(X; \hat{X}) = \lambda R^{(I)}(D_1) + (1 - \lambda)R^{(I)}(D_2) \]

- Showing the convexity of \( R^{(I)}(D) \)

- Not surprisingly, shapes we’ve seen so far are of the form:
The next four slides are a bit of review.
Rate distortion: set up

- A source produces $x_1, x_2, \cdots \sim p(x)$ based on source distribution $p(x)$ with $x_i \in \mathcal{X}$ for all $i$.
- An encoder $f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$ takes a sequence of source symbols $x_{1:n}$ and maps them to an integer.
- A decoder $g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$ takes an integer and maps to quantized vector (i.e., a codeword).
- A distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ measures how bad the mapping is. I.e., $d(x, \hat{x})$ measures the “cost” of representing $x \in \mathcal{X}$ by $\hat{x} \in \hat{\mathcal{X}}$.
- Distortion is bounded (sometimes needed) if $\exists d_{\text{max}}$ such that $d_{\text{max}} \triangleq \max_{x, \hat{x}} d(x, \hat{x}) < \infty$.
- Ex: Hamming (probability of error) distortion.

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{otherwise} \end{cases} \quad (23.35)$$

Then $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$
Definition 23.7.7

A \((2^nR, n)\) rate distortion code consists of an encoding function

\[ f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^nR\} \quad (23.36) \]

and a decoding function

\[ g_n : \{1, 2, \ldots, 2^nR\} \rightarrow \hat{\mathcal{X}}^n \quad (23.37) \]

(Note, \(H(\hat{\mathcal{X}}^n) \leq nR\) since only \(2^nR\) different codewords.)

The distortion of this code is

\[ D = Ed(X_1:n, g_n(f_n(X_1:n))) = \sum_{x_1:n \in \mathcal{X}^n} p(x_1:n)d(x_1:n, g_n(f_n(x_1:n))) \quad (23.38) \]
A rate-distortion pair \((R, D)\) is said to be achievable if \(\exists\) a sequence of \((2^{nR}, n)\) codes \((f_n, g_n)\) with

\[
\lim_{n \to \infty} Ed(X_1:n, g_n(f_n(X_1:n))) \leq D
\]  \hspace{1cm} (23.37)

- So \(D\) is the max allowable distortion for code at this rate \(R\).
- We can make errors, but not too many (bounded average distortion).
- The type of errors we can make is entirely dependent on the distortion function.
- Def: A rate distortion region for a source is the closure of achievable rate distortion pairs \((R, D)\)
- Def: A rate distortion function \(R(D)\) is the infimum of rates \(R\) such that \((R, D)\) is in rate distortion region. I.e.,

\[
R(D) = \inf \{ R : (R, D) \text{ is achievable} \} \hspace{1cm} (23.38)
\]
Theorem 23.7.1

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

\[
R(D) = R^{(I)}(D)
\]  
(23.19)

- 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 \( R > R^{(I)}(D) \), and (2) if \( R > R^{(I)}(D) \), then there exists a sequence of codes that can achieve rate-distortion pair \( (R, D) \).
- Like in channel capacity and entropy compression case, what happens at \( R = R^{(I)}(D) \) depends on the very specific case that one is analyzing.
- Before proving this key theorem, let's look at Gaussian sources.
Proof of converse

Converse: any \( (2^{nR}, n) \) code w. distortion at most \( D \Rightarrow R \geq R^{(I)}(D) \).

**proof of converse.**

- Reminder: given a \( (2^{nR}, n) \) code defined by functions \( f_n \) and \( g_n \), the reproduction of sequence \( X^n \) is given by:

\[
\hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n)) \tag{23.51}
\]
Proof of converse

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

(23.56)
Proof of converse

Proof of converse.

\[ nR \]

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

(23.56)
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \]
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) \]

(23.56)
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n;X^n) \quad (23.52) \]

Therefore, \[ R \geq R(I(D)) \].
Proof of converse

proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) \]

\[ \geq nR \geq R(I)(D) \hspace{1cm} (23.56) \]
Proof of converse

proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \hspace{1cm} (23.53)

\[ \geq n \sum_{i=1}^{n} R(\hat{X}_i; X_{i-1}) \]  \hspace{1cm} (23.56)
Proof of converse

proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \hspace{1cm} (23.53)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \]  \hspace{1cm} (23.54)

\[ \geq n \sum_{i=1}^{n} R(I(\hat{X}^n)) \geq \sum_{i=1}^{n} R(I(\hat{X}^n, X_{1:i-1})) \]  \hspace{1cm} (23.56)
Proof of converse

proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  
\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  
\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \]  
\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) \]  

(23.56)
Proof of converse

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \hspace{1cm} (23.53)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \]  \hspace{1cm} (23.54)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \]  \hspace{1cm} (23.56)
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \hspace{1cm} (23.53)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \]  \hspace{1cm} (23.54)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(I)(Ed(X_i, \hat{X}_i)) \]  \hspace{1cm} (23.56)
Proof of converse

Proof of converse.

\( nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \)  \( (23.52) \)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \( (23.53) \)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \]  \( (23.54) \)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \]

\[ = n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \]

\( (23.56) \)
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  \hspace{1cm} (23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  \hspace{1cm} (23.53)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \]  \hspace{1cm} (23.54)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(I)(Ed(X_i, \hat{X}_i)) \]

\[ = n \sum_{i=1}^{n} \frac{1}{n} R(I)(Ed(X_i, \hat{X}_i)) \geq nR(I)(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \]  \hspace{1cm} (23.55)

\[ \geq nR(I)(D) \]  \hspace{1cm} (23.56)
Proof of converse

Theorem of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n;X^n) \quad (23.52) \]

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \quad (23.53) \]

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \quad (23.54) \]

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i;\hat{X}_i) \geq \sum_{i=1}^{n} R(I)(Ed(X_i, \hat{X}_i)) \]

\[ = n \sum_{i=1}^{n} \frac{1}{n} R(I)(Ed(X_i, \hat{X}_i)) \geq nR(I) \left( \frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i) \right) \quad (23.55) \]

\[ = nR(I)(Ed(X^n, \hat{X}^n)) \quad (23.56) \]
Proof of converse

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  
(23.52)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  
(23.53)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \]  
(23.54)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(I)(Ed(X_i, \hat{X}_i)) \]

\[ = n \sum_{i=1}^{n} \frac{1}{n} R(I)(Ed(X_i, \hat{X}_i)) \geq nR(I)(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \]  
(23.55)

\[ = nR(I)(Ed(X^n, \hat{X}^n)) = nR(I)(D) \]  
(23.56)

Therefore,
\[ R(D) = R(I)(D) \]
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n | X^n) = I(\hat{X}^n; X^n) \tag{23.52} \]

\[ = H(X^n) - H(X^n | \hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n | \hat{X}^n) \tag{23.53} \]

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | \hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(I)(\text{Ed}(X_i, \hat{X}_i)) \]

\[ = n \sum_{i=1}^{n} \frac{1}{n} R(I)(\text{Ed}(X_i, \hat{X}_i)) \geq nR(I)(\frac{1}{n} \sum_{i=1}^{n} \text{Ed}(X_i, \hat{X}_i)) \tag{23.55} \]

\[ = nR(I)(\text{Ed}(X^n, \hat{X}^n)) = nR(I)(D) \tag{23.56} \]

Therefore, \( R \geq R(I)(D) \).
Main Theorem: Achievability

Theorem 23.7.2 (Achievability in 23.4.1)

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 $D$. 

$RD \text{ Bernoulli}$  

$RD \text{ & Gaussians}$  

$RD \& > 1 \text{ Gaussians}$ 

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

Definition 23.7.3 (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:

\[
| - \frac{1}{n} \log p(x^n) - H(X) | < \epsilon \tag{23.57}
\]

$x$-typical

\[
(23.60)
\]
Definition 23.7.3 (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:

\[
\left| - \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad \text{x-typical} \quad (23.57)
\]
\[
\left| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \hat{x}\text{-typical} \quad (23.58)
\]
\[
\left| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) \right| \leq \epsilon \quad \text{new, "distortion typical"} \quad (23.60)
\]
Typicality lives

**Definition 23.7.3 (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:

\[
\left| - \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad \text{$x$-typical} \tag{23.57}
\]

\[
\left| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \text{$\hat{x}$-typical} \tag{23.58}
\]

\[
\left| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical} \tag{23.59}
\]

\[
\left| d(x^n, \hat{x}^n) - E_d(X, \hat{X}) \right| \leq \epsilon \quad \text{new, “distortion typical”} \tag{23.60}
\]
Typicality lives

Definition 23.7.3 (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:

\begin{align*}
| - \frac{1}{n} \log p(x^n) - H(X) | < \epsilon \\
| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) | < \epsilon \\
| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) | < \epsilon \\
| d(x^n, \hat{x}^n) - E_d(X, \hat{X}) | \leq \epsilon 
\end{align*}

$x$-typical \quad (23.57) \\
\hat{x}$-typical \quad (23.58) \\
jointly typical \quad (23.59) \\
new, “distortion typical” \quad (23.60)
Typicality lives

**Definition 23.7.3 (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:

\[
\frac{1}{n} \log p(x^n) - H(X) < \epsilon \quad \text{($x$-typical)} \tag{23.57}
\]

\[
\frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) < \epsilon \quad \text{($\hat{x}$-typical)} \tag{23.58}
\]

\[
\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) < \epsilon \quad \text{jointly typical} \tag{23.59}
\]

\[
|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| \leq \epsilon \quad \text{new, “distortion typical”} \tag{23.60}
\]

Any $x$ s.t. Equations (23.57)-(23.60) are true define the set $A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)}$. 
Probability of typicality

**Lemma 23.7.4**

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

Lemma 23.7.4

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 weak law of large numbers, just like before.
Probability of typicality

Lemma 23.7.4

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

Proof.

Simple application of the weak 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.
proof of achievability in 23.4.1.

- 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 23.4.1.

- 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 23.4.1.

- 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 \( p(\hat{x}) = \sum_x p(x)p(\hat{x}|x) \).
- Choose \( \epsilon > 0 \) and \( \delta > 0 \).
proof of achievability in 23.4.1.

- 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 23.4.1.

- 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 a random codebook.
- Generate a random codebook $\mathcal{C}$ (a set of $2^{nR}$ codewords, $\{\hat{x}_1:n(w)\}_{w=1,...,2^{nR}}$).
proof of achievability in 23.4.1.

- 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 \( \mathcal{C} \) (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. \( \sim \prod_{i=1}^{n} p(\hat{x}_i) \).

...
Main Theorem: Achievability

proof of achievability in 23.4.1.

- 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 23.4.1.

Encoding:

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

⇒ rate $\approx R$. 

Decoding:

Just produce $\hat{x}^n(w)$. ...
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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 $w$. 

...
Main Theorem: Achievability

... proof of achievability in 23.4.1.

Encoding:

- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_{d,\varepsilon}^{(n)}$.
- 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).
  \[ \Rightarrow \text{rate} \approx R. \]
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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).
  \[ \Rightarrow \text{rate} \approx R. \]

Decoding:

...
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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).
  \[ \Rightarrow \text{rate} \approx R. \]

Decoding:

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

... proof of achievability in 23.4.1.

Distortion:

- Average distortion over both codebooks and codewords:

\[
\bar{D} = E_{X^n,C}d(X^n, \hat{X}^n) = \sum_{C,x^n} \Pr(C)p(x^n)d(x^n, \hat{x}^n)
\]  \hspace{1cm} (23.61)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

Distortion:

- Average distortion over both codebooks and codewords:

  \[ \bar{D} = E_{X^n,C}d(X^n, \hat{X}^n) = \sum_{C,x^n} \Pr(C)p(x^n)d(x^n, \hat{x}^n) \]  
  \[ (23.61) \]

- 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 \( p(x^n) \). ...
... proof of achievability in 23.4.1.

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

$$R(D) = R^{(1)}(D)$$
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $d(x^n, \hat{x}^n(w)) < D + \epsilon$.
Main Theorem: Achievability

...proof of achievability in 23.4.1.

1. 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 $(x^n, \hat{x}^n(w)) \in A^{(n)}_{d,\epsilon}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A^{(n)}_{d,\epsilon}) \to 1$. 

2. 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 $\bar{D} = \mathbb{E}d(x^n, \hat{x}^n(x^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$ (23.62) for any $\delta > 0$ if $\epsilon$ is chosen small, and as long as $P_e \to 0$ as $n \to \infty$.

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

...
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $\Pr(A_d^{(n)}) \to 1$.

  - **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 23.4.1.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:
  - Category A: $x^n : \exists \hat{x}^n(w) \text{ with } (x^n, \hat{x}^n(w)) \in A_d^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) \to 1$.
  - 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 $\bar{D} = \mathbb{E}d(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$ (23.62) 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 $n \to \infty$. 

---

Prof. Jeff Bilmes
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $\Pr(A_{d,\epsilon}^{(n)}) \to 1$.
  - 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

$$\bar{D} \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$$  \hspace{1cm} (23.62)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $(x^n, \hat{x}^n(w)) \in A^{(n)}_{d,\epsilon}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $Pr(A^{(n)}_{d,\epsilon}) \to 1$.

- 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

$$\bar{D} = Ed(X^n, \hat{X}^n(X^n))$$  (23.62)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $\Pr(A_{d,\epsilon}^{(n)}) \to 1$.

  - 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

  \[
  \bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} \tag{23.62}
  \]
Main Theorem: Achievability

...proof of achievability in 23.4.1.

- 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 $(x^n, \hat{x}^n(w)) \in A^{(n)}_{d,\epsilon}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A^{(n)}_{d,\epsilon}) \to 1$.

  - 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

  $$\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$$  \hspace{1cm} (23.62)
### Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $(x^n, \hat{x}^n(w)) \in A_d^{(n)}_{d,\epsilon}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_d^{(n)}_{d,\epsilon}) \rightarrow 1$.
  - **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

\[
\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta \quad (23.62)
\]

for any $\delta > 0$ if $\epsilon$ is chosen small, and as long as $P_e \rightarrow 0$ as $n \rightarrow \infty$
Main Theorem: Achievability

... proof of achievability in 23.4.1.

- 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 $\Pr(A_{d,\epsilon}^{(n)}) \rightarrow 1$.

  - 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

  \[
  \bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta \quad (23.62)
  \]

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

- Trick is to show that $P_e$ gets small fast with $n \rightarrow \infty$. …
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

What we will show is that

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

(23.63)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

What we will show is that

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

\(<\epsilon \text{ for } n \text{ sufficiently large} \quad (23.63)\)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- What we will show is that

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

exponentially fast to zero if \( R > I + 3\epsilon \)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- What we will show is that

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

that is, if we can choose $p(\hat{x}|x)$ to get value $R^{(I)}(D)$ in the limit. In such case $I(X;\hat{X})$ above becomes $R^{(I)}(D)$.
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- What we will show is that

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

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

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

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

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \] (23.64)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- This gives

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

(23.65)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- This gives

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

(23.65)

- 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 23.4.1.

General idea first:

- This gives
  \[ P_e \leq \epsilon + (e^2)^{-n(R - \hat{I}(X;\hat{X}) - 3\epsilon)} \] (23.65)

- 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 \( \mathcal{C}^* \) with this rate, block-length, and distortion.
Main Theorem: Achievability

... proof of achievability in 23.4.1.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \]  \hspace{1cm} (23.65)

- 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 \( R > R^{(I)}(D) \).
Theorem 23.7.5

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

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

(23.66)

Proof.

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

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

(23.69)
Subsidiary Theorems

Theorem 23.7.5

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

\[ p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (23.66)

Proof.

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

\[ p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} \]  \hspace{1cm} (23.69)
Subsidiary Theorems

**Theorem 23.7.5**

∀(x^n, ŷ^n) ∈ A_{d,ε}^{(n)}, we have

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

(23.66)

**Proof.**

∀(x^n, ŷ^n) ∈ A_{d,ε}^{(n)}, we have

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

(23.67)

(23.69)
Subsidiary Theorems

Theorem 23.7.5

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

\[ p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \quad (23.66) \]

Proof.

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

\[ p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = p(x^n) \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \quad (23.67) \]

\[ \leq p(\hat{x}^n) \frac{2^{-n(H(X;\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} \quad (23.68) \]

\[ \leq p(\hat{x}^n)2^{-n(2H(X)+\epsilon)} \quad (23.69) \]
Subsidiary Theorems

Theorem 23.7.5

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

\[ p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \]  
(23.66)

Proof.

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

\[ p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = \frac{p(x^n)}{p(x^n)p(\hat{x}^n)} \]  
(23.67)

\[ \leq \frac{2^{-n(H(X;\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} \]  
(23.68)

\[ = p(\hat{x}^n)2^{n(I(X;\hat{X})+3\epsilon)} \]  
(23.69)
Subsidiary Theorems

Theorem 23.7.6

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

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

(23.70)

Proof.

- \(f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0\).
Subsidiary Theorems

Theorem 23.7.6

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (23.70)$$

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 23.7.6

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (23.70)$$

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 23.7.6

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

$$\left(1 - xy\right)^n \leq 1 - x + e^{-yn}$$  (23.70)

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 \( x = 1 \).
Subsidiary Theorems

... 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 \(1 \leq 1 + e^{-yn}\).
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 \( \frac{\partial^2 g_y}{\partial x^2} \geq 0 \).
... 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 \( 0 \leq x \leq 1 \):

\[
(23.75)
\]
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 \(0 \leq x \leq 1\):

\[
(1 - xy)^n
\] 

(23.75)
\[ R(D) = R^{(I)}(D) \]

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 \( 0 \leq x \leq 1 \):

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

(23.75)
... 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 \( 0 \leq x \leq 1 \):

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

(23.71)
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 \( 0 \leq x \leq 1 \):

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

(23.71) [23.72] [23.75]
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 \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \\
\leq (1 - x)g_y(0) + xg_y(1) \\
= (1 - x) \cdot 1 + x \cdot (1 - y)^n
\] (23.71)

(23.72)

(23.73)

(23.74)

(23.75)
... 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 $0 \leq x \leq 1$:

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

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

...proof of achievability in 23.4.1.

- 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 23.4.1.

- 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:

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

(23.76)
Main Theorem: Achievability

...proof of achievability in 23.4.1.

- 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} (23.76)

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

... proof of achievability in 23.4.1.

$$P_e$$

(23.80)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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

(23.77)

\[
(23.80)
\]
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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

(23.77)

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

(23.78)

(23.80)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

\[ P_e = \sum\limits_C \Pr(C) \sum\limits_{x^n : x^n \notin J(C)} p(x^n) \]  
\[ = \sum\limits_{x^n} p(x^n) \sum\limits_{C : x^n \notin J(C)} \Pr(C) \]  
\[ = \sum\limits_{x^n} p(x^n) \left\{ \text{total prob of all } 2^{nR} \text{ current } C \text{ code-words not being distortion typical with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \right\} \]
Main Theorem: Achievability

... proof of achievability in 23.4.1.

\[ P_e = \sum_{C} \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n) \]  \hspace{1cm} (23.77)

\[ = \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C) \]  \hspace{1cm} (23.78)

\[ = \sum_{x^n} p(x^n) \begin{cases} \text{total prob of all } 2^nR \text{ current } C \text{ code-} \\ \text{words not being distortion typical} \\ \text{with current } x^n \text{ (i.e., prob. of choos-} \\ \text{ing codebook not good for current} \\ x^n) \end{cases} \]  \hspace{1cm} (23.79)

\[ = \sum_{x^n} p(x^n) q^{2^nR} \]  \hspace{1cm} (23.80)
Main Theorem: Achievability

...proof of achievability in 23.4.1.

\[ P_e = \sum_C \Pr(C) \sum_{x^n: x^n \notin J(C)} p(x^n) \]  \hspace{1cm} (23.77)

\[ = \sum_{x^n} p(x^n) \sum_{C: x^n \notin J(C)} \Pr(C) \]  \hspace{1cm} (23.78)

\[ = \sum_{x^n} p(x^n) \left\{ \sum_{C: x^n \notin J(C)} \Pr(C) \right\} \]  \hspace{1cm} \text{total prob of all } 2^{nR} \text{ current } C \text{ code-words not being distortion typical with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \]  \hspace{1cm} (23.79)

\[ = \sum_{x^n} p(x^n) q^{2^{nR}} \]  \hspace{1cm} (23.80)

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

... proof of achievability in 23.4.1.

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

... proof of achievability in 23.4.1.

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

$$q = \Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \Pr(K(x^n, \hat{X}^n) = 1) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n)$$ (23.81)

$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n)$$ (23.83)

This last line follows from Theorem 23.7.5.
Main Theorem: Achievability

... proof of achievability in 23.4.1.

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

Then

\[
q = \Pr((x^n, \hat{X}^n) \notin A_{d,e}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \Pr(K(x^n, \hat{X}^n) = 1) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \\
\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n)
\]

This last line follows from Theorem 23.7.5.

...
Main Theorem: Achievability

... proof of achievability in 23.4.1.

then we have

\[ P_e \]
Main Theorem: Achievability

...proof of achievability in 23.4.1.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \]  \hspace{1cm} (23.84)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

Then we have

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

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

... proof of achievability in 23.4.1.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \]  \hspace{1cm} (23.84)

\[ \leq \sum_{x^n} p(x^n) \left( 1 - \frac{2^{-n(I(X;\hat{X})+3\epsilon)}}{1} \right) \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \]  \hspace{1cm} (23.85)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

then we have

\[
P_e = \sum_{x^n} p(x^n) q^{2nR}
\]  \hspace{1cm} (23.84)

\[
\leq \sum_{x^n} p(x^n) \left( \frac{1}{1} - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x,\hat{x}) \right)^{2nR}
\]  \hspace{1cm} (23.85)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

then we have

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

(23.84)

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

(23.85)
Main Theorem: Achievability

... proof of achievability in 23.4.1.

then we have

\[
P_e = \sum_{x^n} p(x^n) q^{2^n R} \tag{23.84}
\]

\[
\leq \sum_{x^n} p(x^n) \left( \frac{1}{1} - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right) \tag{23.85}
\]
Main Theorem: Achievability

...proof of achievability in 23.4.1.

then we have

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

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

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \]
Main Theorem: Achievability

...proof of achievability in 23.4.1.

then we have

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

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

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \leq \sum_{x^n} p(x^n)(1 - x - e^{-yn}) \]
Main Theorem: Achievability

... proof of achievability in 23.4.1.

then we have

\[ P_e = \sum_{x^n} p(x^n)q^{2^nR} \] (23.84)

\[ \leq \sum_{x^n} p(x^n) \left( \frac{1}{1 - 2^{-n(I(X;\hat{X})+3\epsilon)}} \sum_{\hat{x}^n} p(\hat{x}|x)K(x, \hat{x}) \right)^n \] (23.85)

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \leq \sum_{x^n} p(x^n)(1 - x - e^{-yn}) \] (23.86)

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

... proof of achievability in 23.4.1.

Now

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

(23.88)

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

... proof of achievability in 23.4.1.

- Now

\[
1 - \sum_{x^n, \hat{x}^n} p(x^n) p(\hat{x}^n | x^n) K(x^n, \hat{x}^n)
\]  (23.88)

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

\[
\exp(-2^n(R-I(X; \hat{X})-3\epsilon)) \to 0
\]  (23.89)

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

...proof of achievability in 23.4.1.

Now

\[
1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n)
\]  \hspace{1cm} (23.88)

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

\[
\exp(-2^n(R - I(X;\hat{X}) - 3\epsilon)) \to 0
\]  \hspace{1cm} (23.89)

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 \( R > R(I(D)) \) implying that \( R > I(X;\hat{X}) + 3\epsilon \) for all \( \epsilon \) as small as we want.