Lecture 24 - Jan 27nd, 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 for Bernoulli, Gaussian, and Multiple Gaussians with unequal noise
- L24 (1/27): main rate distortion theorem, geometry, and computing $R(D)$
- 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 (today), 1/27/14 at 11:45pm. Only four problems, but these are good problems (and first three are on Gaussian channels).
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.
The rate-distortion function $R(D)$ for Bernoulli($p$) with $d(x, \hat{x}) = 1_{\{x \neq \hat{x}\}}$ (Hamming distortion) has the following form:

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

(24.1)
Theorem 24.2.1

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

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

(24.18)

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.
Theorem: 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.
- With constraint \( E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D \), we get:

**Theorem 24.2.1**

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
\[
R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i} = \sum_{i=1}^{m} R_i \tag{24.42}
\]

where
\[
D_i = \begin{cases} 
\lambda & \text{if } \lambda < \sigma_i^2 \ (\Rightarrow R_i > 0) \\
\sigma_i^2 & \text{if } \lambda \geq \sigma_i^2 \ (\Rightarrow R_i = 0)
\end{cases} = \min(\lambda, \sigma_i^2) \tag{24.43}
\]

and where \( \lambda \) is chosen so that \( \sum_i D_i = D \).
The next four slides are a bit more review of earlier slides.
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.

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{otherwise} \end{cases} \quad (24.35)$$

Then $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$
Definition 24.3.7

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

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

(24.36)

and a decoding function

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

(24.37)

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

The distortion of this code is

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

(24.38)
**Achievability and rate-distortion pairs**

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

$$D(R) = \inf \{ D : (R, D) \text{ is achievable} \} \quad (24.37)$$

- The next definition is very important.

**Definition 24.3.7**

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

$$R(I)(D) = \min_{p(\hat{x}|x) : \sum_x \sum_{\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \quad (24.38)$$

- Lets now spend a bit of time getting some intuition on this function.
Definition 24.3.7

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

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

(24.37)

- 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.,

\[
R(D) = \inf \{ R : (R, D) \text{ is achievable} \} \quad (24.38)
\]
Key Theorem

Theorem 24.3.1

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

$$R(D) = R^{(I)}(D) \quad (24.18)$$

- 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.
Converse of Theorem 24.3.1 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 24.3.1 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

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

(24.1)

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

Alternatively, for any achievable \((R, D)\) pair, we have that \( R \geq R^{(I)}(D) \).
Converse of Theorem 24.3.1 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 \((2^n R, n)\) 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 \( R > R^{(I)}(D) \)

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 24.3.1

$R(I)(D)$ 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 $D \uparrow$. 

![Graph showing the non-increasing and convex nature of $R(I)(D)$ as a function of $D$.]
**Lemma 24.3.1**

\( R^{(I)}(D) \) 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 24.3.1

\( R^{(I)}(D) \) 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
  \[
  D_\lambda = \lambda D_1 + (1 - \lambda)D_2 = \sum x, \hat{x} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x}).
  \]
Lemma 24.3.1

$R(I)(D)$ 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 24.3.1

$R^{(I)}(D)$ 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 $D_\lambda = \lambda D_1 + (1 - \lambda) D_2 = \sum_x \sum_{\hat{x}} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x})$.

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

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

Prof. Jeff Bilmes

EE515a/Winter 2014/Information Theory II – Lecture 24 - Jan 27nd, 2014
Proof.

Therefore,

\[ R(I)(D_\lambda) \leq I_{p\lambda}(X; \hat{X}) \leq \lambda I_{p1}(X; \hat{X}) + (1 - \lambda)I_{p2}(X; \hat{X}) \]

\[ = \lambda R(I)(D_1) + (1 - \lambda)R(I)(D_2) \]

(24.2)

(24.3)

(24.4)

Showing the convexity of \( R(I)(D) \)

Not surprisingly, shapes we’ve seen so far are of the form:
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:

\[
\hat{X}^n = \hat{X}^n(x^n) = g_n(f_n(x^n))
\]  

(24.5)
Proof of converse

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

\[
\begin{align*}
\text{proof of converse.} \\
\end{align*}
\]

(24.10)
Proof of converse proof of converse.

\[ nR \]

(24.10)
Proof of converse

proof of converse.

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

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

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n;X^n) \quad (24.6) \]
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} (24.6)

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

(24.10)
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} (24.6)

\[= H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \] \hspace{1cm} (24.7)

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

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n | X^n) = I(\hat{X}^n; X^n) \quad (24.6) \]

\[ = H(X^n) - H(X^n | \hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n | \hat{X}^n) \quad (24.7) \]

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | \hat{X}^n, X_{1:i-1}) \quad (24.8) \]

\[ \geq \sum_{i=1}^{n} R(\tilde{D}(X_i; \tilde{X}^n)) \quad (24.10) \]
Proof of converse

\begin{align*}
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)
\end{align*}

(24.6) (24.7) (24.8) (24.10)
Proof of converse

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n;X^n) \quad (24.6) \]

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \quad (24.7) \]

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \quad (24.8) \]

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \quad (24.10) \]
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \quad (24.6) \]

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \quad (24.7) \]

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{1:i-1}) \quad (24.8) \]

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R(I)(Ed(X_i, \hat{X}_i)) \quad (24.9) \]

\[ (24.10) \]
Proof of converse

Proof of converse.

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n | X^n) = I(\hat{X}^n; X^n) \quad (24.6) \]

\[ = H(X^n) - H(X^n | \hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n | \hat{X}^n) \quad (24.7) \]

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | \hat{X}^n, X_{1:i-1}) \quad (24.8) \]

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

\[ = n \sum_{i=1}^{n} \frac{1}{n} R(I)(Ed(X_i, \hat{X}_i)) \quad (24.10) \]
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} (24.6)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \] \hspace{1cm} (24.7)

\[ \geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) = \sum_{i=1}^{n} I(X_i; \hat{X}_i) \geq \sum_{i=1}^{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \] \hspace{1cm} (24.8)

\[ = n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \geq n R^{(I)}(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \] \hspace{1cm} (24.9)

\[ = n R^{(I)}(D) \] \hspace{1cm} (24.10)
Proof of converse

Proof of converse.

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

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

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

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

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

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

Therefore, \( R(D) = R^{(I)}(D) \).
Proof of converse

\[ nR \geq H(\hat{X}^n) \geq H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \] (24.6)

\[ = H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \] (24.7)

\[ = \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_1:i-1) \] (24.8)

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

\[ = n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \geq n R^{(I)}(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \] (24.9)

\[ = n R^{(I)}(Ed(X^n, \hat{X}^n)) = n R^{(I)}(D) \] (24.10)
Proof of converse

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

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

Theorem 24.3.2 (Achievability in 24.3.1)

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 $D$. 
Definition 24.3.3 (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:

\[
\left| - \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad x\text{-typical} \tag{24.11}
\]

\[
\left| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \hat{x}\text{-typical} \tag{24.12}
\]

\[
\left| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical} \tag{24.13}
\]

\[
|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| \leq \epsilon \quad \text{new, “distortion typical”} \tag{24.14}
\]
Typicality lives

**Definition 24.3.3 (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:

\[
| - \frac{1}{n} \log p(x^n) - H(X)| < \epsilon \quad x\text{-typical} \tag{24.11}
\]

\[
| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X})| < \epsilon \quad \hat{x}\text{-typical} \tag{24.12}
\]

\[
| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X})| < \epsilon \quad \text{jointly typical} \tag{24.13}
\]

\[
| d(x^n, \hat{x}^n) - E_d(X, \hat{X})| \leq \epsilon \quad \text{new, "distortion typical"} \tag{24.14}
\]
Typicality lives

Definition 24.3.3 (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:

1. $\left| - \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad x$-typical
2. $\left| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \hat{x}$-typical
3. $\left| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical}$
4. $|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| \leq \epsilon$
Typicality lives

**Definition 24.3.3 (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:

\[
\begin{align*}
| - \frac{1}{n} \log p(x^n) - H(X) | &< \epsilon & \text{x-typical} & (24.11) \\
| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) | &< \epsilon & \text{\hat{x}-typical} & (24.12) \\
| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) | &< \epsilon & \text{jointly typical} & (24.13) \\
| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) | &\leq \epsilon & \text{new, “distortion typical”} & (24.14)
\end{align*}
\]
Typicality lives

Definition 24.3.3 (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:

\[
\left| - \frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad \text{\textit{x-typical}} \tag{24.11}
\]

\[
\left| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \text{\textit{\hat{x}-typical}} \tag{24.12}
\]

\[
\left| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical} \tag{24.13}
\]

\[
\left| d(x^n, \hat{x}^n) - Ed(X, \hat{X}) \right| \leq \epsilon \quad \text{new, “distortion typical”} \tag{24.14}
\]

Any $x$ s.t. Equations (24.11)-(24.14) are true define the set $A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)}$. 
Lemma 24.3.4

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 24.3.4

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.
Lemma 24.3.4

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. \(\square\)

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 24.3.1.

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

...
proof of achievability in 24.3.1.

- 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 $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$.
proof of achievability in 24.3.1.

- 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 $\delta > 0$. 

...
Main Theorem: Achievability

Proof of achievability in 24.3.1.

- 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 24.3.1.

- 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, $\{\hat{x}_1:n(w)\}_{w=1,...,2^{nR}}$).
proof of achievability in 24.3.1.

- 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. $\sim \prod_{i=1}^n p(\hat{x}_i)$. ...
Main Theorem: Achievability

proof of achievability in 24.3.1.

- 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, $\{\hat{x}_1:n(w)\}_{w=1,...,2^{nR}}$). 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 24.3.1.

Encoding:

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

...
Main Theorem: Achievability

... proof of achievability in 24.3.1.

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

... proof of achievability in 24.3.1.

Encoding:
- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_d^{(n)}$.
- 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).
  $\Rightarrow$ rate $\approx R$. 

...
Main Theorem: Achievability

... proof of achievability in 24.3.1.

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).

$\Rightarrow$ rate $\approx R$.

Decoding:
...

Prof. Jeff Bilmes
EE515a/Winter 2014/Information Theory II – Lecture 24 - Jan 27nd, 2014
Main Theorem: Achievability

... proof of achievability in 24.3.1.

Encoding:
- We encode $x^n$ by $w$ if there exists a $w$ such that $(x^n, \hat{x}^n(w)) \in A_d^n$.
- 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).
  $\Rightarrow$ rate $\approx R$.

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

... proof of achievability in 24.3.1.

Distortion:

- Average distortion over both codebooks and codewords:

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

(24.15)
Main Theorem: Achievability

Proof of achievability in 24.3.1.

Distortion:

- Average distortion over both codebooks and codewords:

$$\bar{D} = E_{X^n,C}d(X^n, \hat{X}^n) = \sum_{C,x^n} \Pr(C)p(x^n)d(x^n, \hat{x}^n)$$ (24.15)

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

... proof of achievability in 24.3.1.

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

... proof of achievability in 24.3.1.

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

... proof of achievability in 24.3.1.

- 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 \( \Pr(A_{d,\epsilon}^{(n)}) \rightarrow 1 \).
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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.
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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 $\leq P_e d_{\text{max}}$.

\[ \sum P. D \leq D_{\text{max}} \]
Main Theorem: Achievability

...proof of achievability in 24.3.1.

- 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

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

(24.16)

...
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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. 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

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

(24.16)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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

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

    (24.16)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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

  $$\bar{D} = E d(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta$$  \hspace{1cm} (24.16)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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

$$\bar{D} = Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\text{max}} < D + \delta \quad (24.16)$$

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

...proof of achievability in 24.3.1.

- 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

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

(24.16)

for any $\delta > 0$ if $\epsilon$ is chosen small, and as long as $P_e \to 0$ as $n \to \infty$.
- Trick is to show that $P_e$ gets small fast with $n \to \infty$. ...
Main Theorem: Achievability

Proof of achievability in 24.3.1.

General idea first:

- What we will show is that

\[
P_e \leq \Pr \left( (X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)} \right) + e^{2n(R-I(X;\hat{X})-3\epsilon)}
\]  (24.17)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

General idea first:

What we will show is that

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

(24.17)

<\epsilon for n sufficiently large
Main Theorem: Achievability

...proof of achievability in 24.3.1.

General idea first:

- What we will show is that

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

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

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

(24.17)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

General idea first:

- What we will show is that

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

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

... proof of achievability in 24.3.1.

General idea first:

- What we will show is that

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

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

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 \( R^{(I)}(D) \).

- This gives

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

... proof of achievability in 24.3.1.

General idea first:

- This gives

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

(24.19)
Main Theorem: Achievability

...proof of achievability in 24.3.1.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \]  \hspace{1cm} (24.19)

- 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 24.3.1.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^n(R-I(X;\hat{X})-3\epsilon) \] (24.19)

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

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

... proof of achievability in 24.3.1.

General idea first:

- This gives

\[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X})-3\epsilon)} \]  \hspace{1cm} (24.19)

- 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 24.3.5

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

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

(24.20)

Proof.

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

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

(24.23)
Subsidiary Theorems

**Theorem 24.3.5**

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

\[ p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{n(I(X; \hat{X})+3\epsilon)} \] (24.20)

**Proof.**

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

\[ p(\hat{x}^n | x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} \] (24.23)
Subsidiary Theorems

**Theorem 24.3.5**

$$\forall (x^n, \hat{x}^n) \in A^{(n)}_{d,\epsilon}, \text{ we have}$$

$$p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \quad (24.20)$$

**Proof.**

$$\forall (x^n, \hat{x}^n) \in A^{(n)}_{d,\epsilon}, \text{ we have}$$

$$p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = p(\hat{x}^n) \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \quad (24.21)$$

$$\quad (24.23)$$
Subsidiary Theorems

**Theorem 24.3.5**

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

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

(24.20)

**Proof.**

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

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

(24.21)

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

(24.22)

(24.23)
Subsidiary Theorems

**Theorem 24.3.5**

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

\[ p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (24.20)

**Proof.**

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

\[ p(\hat{x}^n | x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \]  \hspace{1cm} (24.21)

\[ \leq p(\hat{x}^n) \frac{2^{-n(H(X;\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} \]  \hspace{1cm} (24.22)

\[ = p(\hat{x}^n) 2^{n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (24.23)
Subsidiary Theorems

**Theorem 24.3.6**

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

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

(24.24)

**Proof.**

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

Theorem 24.3.6

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \tag{24.24}$$

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 24.3.6

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (24.24)$$

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 24.3.6

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (24.24)$$

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 $0 \leq 1 - y \leq e^{-y}$, which is a variational lower bound.
... proof continued.

\[ \Rightarrow (1 - y)^n \leq e^{-yn} \] which already is the theorem for \( x = 1 \).
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 \( 1 \leq 1 + e^{-yn} \).
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.

- $\Rightarrow (1 - y)^n \leq e^{-yn}$ 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 $0 \leq x \leq 1$: 
... proof continued.

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

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

... proof continued.

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

\[
(1 - xy)^n = g_y(x)
\]
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 \( 0 \leq x \leq 1 \):

\[
(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \quad (24.25)
\]
... 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 \( 0 \leq x \leq 1 \):

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

(24.25)  

(24.26)
... proof continued.

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

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

(24.25)  (24.26)  (24.27)
... 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 \( 0 \leq x \leq 1 \):
  \[
  (1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \geq (1 - x)g_y(0) + xg_y(1) \\
  \leq 1 - x + xe^{-yn} \]
Subsidiary Theorems

... proof continued.

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

...proof of achievability in 24.3.1.

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 24.3.1.

- 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:

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

(24.30)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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:

$$J(C) = \left\{ x^n : \exists \hat{x}^n \in C \text{ s.t. } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)} \right\}$$ (24.30)

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

...proof of achievability in 24.3.1.

\[ P_e \]

\[ (24.34) \]
Main Theorem: Achievability

... proof of achievability in 24.3.1.

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

(24.31)

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

(24.34)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

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

(24.31)

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

(24.32)

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

(24.34)

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

\[ P_e = \sum_C \Pr(C) \sum_{x^n: x^n \notin J(C)} p(x^n) \]  \hspace{1cm} (24.31)

\[ = \sum_{x^n} p(x^n) \sum_{C: x^n \notin J(C)} \Pr(C) \]  \hspace{1cm} (24.32)

\[ = \sum_{x^n} p(x^n) \left\{ \text{total prob of all } 2^{nR} \text{ current } C \text{ code-words not being distortion typical with current } x^n \right\} \]  \hspace{1cm} (24.33)

\[ \text{with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \]  \hspace{1cm} (24.34)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

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

... proof of achievability in 24.3.1.

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

(24.31)

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

(24.32)

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

(24.33)

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

(24.34)

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

...proof of achievability in 24.3.1.

- 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 24.3.1.

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

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

(24.35)

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

(24.36)

(24.37)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

- 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

\[
q = \Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \Pr(K(x^n, \hat{X}^n) = 1) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \tag{24.35}
\]

\[
\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n) \tag{24.36}
\]

- This last line follows from Theorem 24.3.5.
Main Theorem: Achievability

...proof of achievability in 24.3.1.

then we have

\[ \Pr_e \]
Main Theorem: Achievability

...proof of achievability in 24.3.1.

then we have

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

(24.38)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

\[ P_e = \sum_{x^n} p(x^n)q^{2^{nR}} \]  \hspace{2cm} (24.38)

\[ \leq \sum_{x^n} p(x^n) \left( 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x)K(x,\hat{x}) \right)^{2^{nR}} \]  \hspace{2cm} (24.39)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

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

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

(24.38)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \]  \hspace{1cm} (24.38)

\[ \leq \sum_{x^n} p(x^n) \left( 1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^{2nR}} p(\hat{x}|x) K(x, \hat{x}) \right) \]  \hspace{1cm} (24.39)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

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

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

(24.38)

(24.39)
Main Theorem: Achievability

...proof of achievability in 24.3.1.

then we have

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

(24.38)

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

(24.39)
Main Theorem: Achievability

...proof of achievability in 24.3.1.

then we have

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

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

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

(24.38)

(24.39)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

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

(24.38)

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

(24.39)

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

(24.40)
Main Theorem: Achievability

... proof of achievability in 24.3.1.

then we have

\[ P_e = \sum_{x^n} p(x^n) q^{2nR} \] (24.38)

\[ \leq \sum_{x^n} p(x^n) \left( \frac{1}{1} - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}|x) K(x, \hat{x}) \right)^{2nR} \] (24.39)

\[ = \sum_{x^n} p(x^n)(1 - xy)^n \leq \sum_{x^n} p(x^n)(1 - x - e^{-yn}) \] (24.40)

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

... proof of achievability in 24.3.1.

Now

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

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

... proof of achievability in 24.3.1.

- Now

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

(24.42)

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

- Also

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

(24.43)

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

... proof of achievability in 24.3.1.

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} (24.42)

is just \(\Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)})\) and is \(< \epsilon\) 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} (24.43)

if \(R > I(X;\hat{X}) + 3\epsilon\). This is true when we chose \(p(\hat{x}|x)\) to be the distribution that achieves the minimum, and since \(R > I(X;\hat{X}) = R^{(I)}(D)\) in this case, we get \(R > I(X;\hat{X}) + 3\epsilon\) for all \(\epsilon\) as small as we want for large \(n\).
Reminder: Geometry of channel capacity

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

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

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

$$h(Y) \leq \frac{1}{2} \log[2\pi e (P + \sigma^2)] \quad (24.44)$$
$$h(Z) \leq \frac{1}{2} \log[2\pi e (\sigma^2)] \quad (24.45)$$
Reminder: Geometry of channel capacity

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

\[
  h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] \tag{24.44}
\]

\[
  h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)] \tag{24.45}
\]

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

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

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

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

  $$h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)]$$  \hspace{1cm} (24.44)

  $$h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)]$$  \hspace{1cm} (24.45)

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

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

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

  $$R = \frac{1}{2} \log(1 + P/\sigma^2)$$  \hspace{1cm} (24.46)
Reminder: Geometry of channel capacity

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

$$h(Y) \leq \frac{1}{2} \log[2\pi e(P + \sigma^2)] \quad (24.44)$$

$$h(Z) \leq \frac{1}{2} \log[2\pi e(\sigma^2)] \quad (24.45)$$

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

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

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

$$R = \frac{1}{2} \log(1 + P/\sigma^2) \quad (24.46)$$

- Assuming no overlap of volumes which is best we can do, so $R = C$. 

Prof. Jeff Bilmes
Reminder: Geometry of channel capacity

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

\[ \text{Goal is to pack as many small spheres in the big sphere as possible.} \]
Reminder: Geometry of channel capacity

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

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{n/2} \log[2\pi e \sigma^2]$$  \hspace{1cm} (24.47)
Reminder: Geometry of channel capacity

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

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{n/2} \log[2\pi e \sigma^2] \quad (24.47)$$

- This gives

$$r_{\sigma^2} = \Gamma^{1/2} \left(\frac{n}{2} + 1\right) (2e\sigma^2)^{1/2} \approx (2e\sigma^2 n)^{1/2} = \sqrt{2e\sigma^2 n} \quad (24.48)$$
Reminder: Geometry of channel capacity

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

- For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} r^n = 2^{n/2} \log[2\pi e \sigma^2]$$  \hspace{1cm} (24.47)

- This gives

$$r_{\sigma^2} = \Gamma^{1/2}\left(\frac{n}{2} + 1\right)(2e\sigma^2)^{1/2} \approx (2e\sigma^2 n)^{1/2} = \sqrt{2e\sigma^2 n}$$  \hspace{1cm} (24.48)

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

$$M \leq \frac{(r_{\sigma^2} + P)^n}{(r_{\sigma^2})^n} = \left(\frac{\sigma^2 + P}{\sigma^2}\right)^{n/2}$$  \hspace{1cm} (24.49)
Sphere-packing: typical set volume can be approximately identified with the volume of a sphere in $\mathcal{R}^n$.

For the noise

$$V(r, n) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} r^n = 2^{\frac{n}{2}} \log[2\pi e \sigma^2]$$  \hspace{1cm} (24.47)

This gives

$$r_{\sigma^2} = \Gamma^{1/2}(\frac{n}{2} + 1)(2e\sigma^2)^{1/2} \approx (2e\sigma^2 n)^{1/2} = \sqrt{2e\sigma^2 n}$$  \hspace{1cm} (24.48)

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

$$M \leq \frac{(r_{\sigma^2} + P)^n}{(r_{\sigma^2})^n} = \left(\frac{\sigma^2 + P}{\sigma^2}\right)^{n/2}$$  \hspace{1cm} (24.49)

Goal is to pack as many small spheres in the big sphere as possible.
Source $X \sim \mathcal{N}(0, \sigma^2)$. 
Source $X \sim \mathcal{N}(0, \sigma^2)$.

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

A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathcal{R}^n$ s.t. most $x^n$ are "near" a codeword.

Sources live in sphere of radius $\sqrt{nD}$.

Minimum number of such words is $\approx \left(\frac{n\sigma^2}{D}\right)^{1/2}$ for Gaussian sources.
Geometry of Rate Distortion

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

A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathbb{R}^n$ s.t. most $x^n$ are “near” a codeword.

Sources live in sphere of radius $(n\sigma^2)^{1/2}$.

Goal is to use the fewest number of codewords s.t. every source sequence $X^n$ is within $\sqrt{nD}$ of some codeword.

Minimum number of such words is

$$ \approx \frac{[(n\sigma^2)^{1/2}]^n}{[(nD)^{1/2}]^n} = \left(\frac{\sigma^2}{D}\right)^{n/2} = 2^{nR(D)} \quad (24.50) $$

for Gaussian sources.
Geometry of Rate Distortion

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

- Let $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ be a large sphere.

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

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

$h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi eD)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.

To make sure that no $x^n$ is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
Geometry of Rate Distortion

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

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

(24.51)
Geometry of Rate Distortion

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

\( h(\hat{X}|X) = h(Z) = \frac{1}{2} \log (2\pi e D) \) is a small sphere, and is a region corresponding to codeword \( \hat{x} \) in the form of a log volume.

To make sure that no \( x^n \) is too far away from codeword, need to spread out (or cover) the large volume as much as possible.

The minimum number of such codewords is then:

\[
\frac{2^n h(X)}{2^n h(Z)} \geq \frac{2^n \frac{1}{2} \log (2\pi e \sigma^2)}{2^n / 2 \log (2\pi e D)} = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^{nR(D)}
\]  

(24.51)
Geometry of Rate Distortion

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

\( h(\hat{X}|X) = h(Z) = \frac{1}{2} \log(2\pi e D) \) is a small sphere, and is a region corresponding to codeword \( \hat{x} \) in the form of a log volume.

To make sure that no \( x^n \) is too far away from codeword, need to spread out (or cover) the large volume as much as possible.

The minimum number of such codewords is then:

\[
\geq \frac{2^{nh(X)}}{2^{nh(Z)}} = \frac{2^{n/2} \log(2\pi e \sigma^2)}{2^{n/2} \log(2\pi e D)} = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^n R(D) \quad (24.51)
\]

Again, meaning \( R = \frac{1}{2} \log(\sigma^2/D) \) bits per source symbol to compress with distortion \( D \).
Suppose $X \sim \mathcal{N}(0, C)$ where $C$ is general positive definite matrix.
Suppose \( X \sim \mathcal{N}(0, C) \) where \( C \) is general positive definite matrix.

Then \( R(D) \) can be easily and analytically calculated by using the reverse water filling argument on the eigenvalues of \( C \).
Suppose $X \sim \mathcal{N}(0, C)$ where $C$ is a general positive definite matrix.

Then $R(D)$ can be easily and analytically calculated by using the reverse water filling argument on the eigenvalues of $C$.

This is analogous to the water filling argument for computing the channel capacity of a general Gaussian.
We’ve seen that for certain special cases (e.g., Bernoulli sources, Gaussian sources, and now also for Gaussian vector w. full covariance matrix), we can compute $R(D) = R(I)(D)$ exactly.
We’ve seen that for certain special cases (e.g., Bernoulli sources, Gaussian sources, and now also for Gaussian vector w. full covariance matrix), we can compute $R(D) = R^I(D)$ exactly.

What if we do not have such simple properties of $X$?
We’ve seen that for certain special cases (e.g., Bernoulli sources, Gaussian sources, and now also for Gaussian vector w. full covariance matrix), we can compute $R(D) = R(I)(D)$ exactly.

What if we do not have such simple properties of $X$?

Let $X \sim p(x)$ which is over multi-alphabet $\mathcal{X}$.
We’ve seen that for certain special cases (e.g., Bernoulli sources, Gaussian sources, and now also for Gaussian vector w. full covariance matrix), we can compute $R(D) = R^{(I)}(D)$ exactly.

What if we do not have such simple properties of $X$?

Let $X \sim p(x)$ which is over multi-alphabet $\mathcal{X}$.

Goal: compute in general

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

(24.52)
Can restate this problem as:

\[
R(D) = \min_{q(\hat{x}|x)} I(X; \hat{X}) \tag{24.53}
\]

subject to:

\[
q(\hat{x}|x) \geq 0 \quad \forall \hat{x}, x \tag{24.54}
\]

\[
\sum_{\hat{x}} q(\hat{x}|x) = 1 \quad \forall x \tag{24.55}
\]

\[
\sum_{\hat{x}, x} q(\hat{x}|x)p(x)d(x, \hat{x}) = D \tag{24.56}
\]

where

\[
I(X; \hat{X}) = \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} \tag{24.57}
\]

and

\[
q(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x) \tag{24.58}
\]
Marginalization vs. Projection

We’re going to see that the marginalization 
\[ q(\hat{x}) = \sum_x p(x)q(\hat{x}|x) \]
can be viewed as a form of projection.
Marginalization vs. Projection

- We’re going to see that the marginalization $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$ can be viewed as a form of projection.
- Generic projection. We have distance measure $d(\cdot, \cdot)$ and a constraint set $\mathcal{P}$, and a vector $\hat{x} \notin \mathcal{P}$. 
Marginalization vs. Projection

- We’re going to see that the marginalization \( q(\hat{x}) = \sum_x p(x)q(\hat{x}|x) \) can be viewed as a form of projection.
- Generic projection. We have distance measure \( d(\cdot, \cdot) \) and a constraint set \( \mathcal{P} \), and a vector \( \hat{x} \notin \mathcal{P} \).
- We want to find the member of \( \mathcal{P} \) that is closest to \( \hat{x} \) where closeness is measured via \( d(\cdot, \cdot) \), i.e.,

\[
x^* \in \arg\min_{x \in \mathcal{P}} d(\hat{x}, x)
\]  

(24.59)
Computing $R(D)$

- Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55) & (24.56)).
Computing $R(D)$

- Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55)&(24.56)).
- We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$. 

Q: Why ok to equal $D$ and not $\leq D$ in the above?

A: we know can only make $R(D)$ smaller by making $D$ larger.

For the moment, lets ignore the inequality constraint $q(\hat{x}|x) \geq 0$ and hope that we find everywhere positive solutions.

We get objective (Lagrangian) in the form:

$$J(Q) = \sum x, \hat{x} p(x) q(\hat{x}|x) \log q(\hat{x}|x) q(\hat{x}) + \lambda \left( \sum x, \hat{x} p(x) q(\hat{x}|x) d(x, \hat{x}) - D \right) + \sum x \nu(x) \left( \sum \hat{x} q(\hat{x}|x) - 1 \right) \text{(24.60)}$$
Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55)&(24.56)).

We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$.

Q: Why ok to equal $D$ and not $\leq D$ in the above?
Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55)&(24.56)).
We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$.
Q: Why ok to equal $D$ and not $\leq D$ in the above? A: we know can only make $R(D)$ smaller by making $D$ larger.
Computing $R(D)$

- Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55)&(24.56)).
- We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$.
- Q: Why ok to equal $D$ and not $\leq D$ in the above? A: we know can only make $R(D)$ smaller by making $D$ larger.
- For the moment, lets ignore the inequality constraint $q(\hat{x}|x) \geq 0$ and hope that we find everywhere positive solutions.
Computing $R(D)$

- Both inequality (Eq.(24.54)) and equality constraints (Eqs.(24.55)&(24.56)).
- We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$.
- Q: Why ok to equal $D$ and not $\leq D$ in the above? A: we know can only make $R(D)$ smaller by making $D$ larger.
- For the moment, lets ignore the inequality constraint $q(\hat{x}|x) \geq 0$ and hope that we find everywhere positive solutions.
- We get objective (Lagrangian) in the form:

$$J(Q) = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})}$$

$$+ \lambda(\sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) - D)$$

$$+ \sum_x \nu(x) \left( \sum_{\hat{x}} q(\hat{x}|x) - 1 \right) \quad (24.60)$$
Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.
Computing $R(D)$

- Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.
- Thus, happily, the KKT conditions hold,
Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.

Thus, happily, the KKT conditions hold,

so we need

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}| |\hat{\mathcal{X}}| \text{ equations}
\] (24.61)
Computing $R(D)$

- Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.
- Thus, happily, the KKT conditions hold, so we need

$$\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}| |\hat{\mathcal{X}}| \text{ equations} \quad (24.61)$$

- The equality constraints give us an additional $|\mathcal{X}| + 1$ equations.
Computing $R(D)$

- Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.
- Thus, happily, the KKT conditions hold, so we need

$$\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}||\hat{\mathcal{X}}| \text{ equations} \quad (24.61)$$

- The equality constraints give us an additional $|\mathcal{X}| + 1$ equations.
- How many unknowns?
Computing $R(D)$

- Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.
- Thus, happily, the KKT conditions hold,
- so we need

$$\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}| |\hat{\mathcal{X}}| \text{ equations} \quad (24.61)$$

- The equality constraints give us an additional $|\mathcal{X}| + 1$ equations.
- How many unknowns? $|\mathcal{X}| |\hat{\mathcal{X}}|$, so this should be solvable (if feasible, and assume is).
Eq (24.60) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.

Thus, happily, the KKT conditions hold, so we need

$$\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}| |\hat{\mathcal{X}}| \text{ equations} \quad (24.61)$$

The equality constraints give us an additional $|\mathcal{X}| + 1$ equations.

How many unknowns? $|\mathcal{X}| |\hat{\mathcal{X}}|$, so this should be solvable (if feasible, and assume is).

If solution is such that $q(\hat{x}|x) > 0$ for all $x, \hat{x}$ then we have solved the problem since this solution is a minimum without requiring the inequality constraints (and we still have equality constraint $\sum_{\hat{x}} q(\hat{x}|x) = 1$ so solution is feasible).
Computing $R(D)$

- When might this occur?
When might this occur? Small $D$ requires a high rate, meaning we need to use all of the symbols in $\hat{\mathcal{X}}$. 
Computing $R(D)$

- When might this occur? Small $D$ requires a high rate, meaning we need to use all of the symbols in $\hat{X}$.
- It might even be that for all $0 \leq D \leq D_{\text{max}}$, we use all symbols.
Computing $R(D)$

- When might this occur? Small $D$ requires a high rate, meaning we need to use all of the symbols in $\hat{X}$.
- It might even be that for all $0 \leq D \leq D_{\text{max}}$, we use all symbols.
- If, on the other hand, there are one or more $q(\hat{x}|x)$ that are negative, then the min of $I(X;\hat{X})$ falls outside of the probability simplex (defined as $\triangle_{q(\hat{x},x)}$).
When might this occur? Small $D$ requires a high rate, meaning we need to use all of the symbols in $\hat{X}$.

It might even be that for all $0 \leq D \leq D_{\text{max}}$, we use all symbols.

If, on the other hand, there are one or more $q(\hat{x} | x)$ that are negative, then the min of $I(X; \hat{X})$ falls outside of the probability simplex (defined as $\Delta_q(\hat{x}, x)$).

$\Delta_q(\hat{x}, x)$ is convex, so in this case, the minimization with the constraint must lie on a boundary of this convex set (essentially KKT conditions).
So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\triangle q(\hat{x},x)$ at boundary will set it to zero.
So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\triangle q(\hat{x}, x)$ at boundary will set it to zero.

That is, we get that $q(\hat{x}'|x) = 0$
So, if solution of problem without inequality constraint is negative (say \( q(\hat{x}'|x) < 0 \) for some \( \hat{x}' \)), then projection back onto convex \( \triangle q(\hat{x},x) \) at boundary will set it to zero.

That is, we get that \( q(\hat{x}'|x) = 0 \)

This means we can essentially “remove” an alphabet symbol (or several such symbols) and form a compressed alphabet \( \hat{X}' \) where \( |\hat{X}'| < |\hat{X}| \)
So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\Delta q(\hat{x},x)$ at boundary will set it to zero.

That is, we get that $q(\hat{x}'|x) = 0$

This means we can essentially “remove” an alphabet symbol (or several such symbols) and form a compressed alphabet $\hat{X}'$ where $|\hat{X}'| < |\hat{X}|$

This happens when $D$ is big, and by minimizing the rate, we don’t need to “use up” all the symbols in $\hat{X}$. 
So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\Delta q(\hat{x},x)$ at boundary will set it to zero.

That is, we get that $q(\hat{x}'|x) = 0$

This means we can essentially “remove” an alphabet symbol (or several such symbols) and form a compressed alphabet $\hat{X}'$ where $|\hat{X}'| < |\hat{X}|$

This happens when $D$ is big, and by minimizing the rate, we don’t need to “use up” all the symbols in $\hat{X}$.

General result: convex minimization over a convex set, an update to minimize objective without constraint need only be projected back to nearest point in convex set.
So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\Delta q(\hat{x},x)$ at boundary will set it to zero.

That is, we get that $q(\hat{x}'|x) = 0$

This means we can essentially “remove” an alphabet symbol (or several such symbols) and form a compressed alphabet $\hat{X}'$ where $|\hat{X}'| < |\hat{X}|$

This happens when $D$ is big, and by minimizing the rate, we don’t need to “use up” all the symbols in $\hat{X}$.

General result: convex minimization over a convex set, an update to minimize objective without constraint need only be projected back to nearest point in convex set.

Thus, lets first solve the partially unconstrained problem (i.e., the one without the inequality constraints, Eq.(24.54).
First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \frac{\nu(x)}{p(x)}$. 

The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$J(q) = \sum_{x, \hat{x}} p(x) q(\hat{x} | x) \left[ \log q(\hat{x} | x) q(\hat{x}) + \lambda d(x, \hat{x}) + \nu(x) p(x) \right]$$

where

$$\beta(x, \hat{x}) = \left[ \log q(\hat{x} | x) q(\hat{x}) \mu(x) + \lambda d(x, \hat{x}) \right]$$
Computing $R(D)$

1. First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \nu(x)/p(x)$.
2. The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$J(q) = \sum_{x, \hat{x}} p(x) q(\hat{x} | x) \left[ \log q(\hat{x} | x) q(\hat{x}) + \lambda d(x, \hat{x}) + \nu(x)/p(x) \right]$$

(24.62)

$$\implies \sum_{x, \hat{x}} p(x) q(\hat{x} | x) \beta(x, \hat{x})$$

(24.64)
Computing $R(D)$

- First, define $\mu(x) \triangleq e^{-\nu(x)/p(x)}$ so that $-\log \mu(x) = \nu(x)/p(x)$.
- The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$J(q)$$

(24.64)
Computing $R(D)$

- First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \frac{\nu(x)}{p(x)}$.
- The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$J(q) = \sum_{x, \hat{x}} p(x) q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right]$$  \hspace{1cm} (24.62)

where

$$\beta(x, \hat{x}) = \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right]$$  \hspace{1cm} (24.65)
Computing $R(D)$

- First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \nu(x)/p(x)$.
- The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$J(q) = \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right]$$  \hspace{1cm} (24.62)

$$= \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right]$$  \hspace{1cm} (24.63)

$$= \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right]$$  \hspace{1cm} (24.64)
First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \nu(x)/p(x)$.

The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

$$ J(q) = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right] $$

(24.62)

$$ = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] $$

(24.63)

$$ = \sum_{x,\hat{x}} p(x)q(\hat{x}|x)\beta(x, \hat{x}) $$

(24.64)
Computing $R(D)$

- First, define $\mu(x) \triangleq e^{-\frac{\nu(x)}{p(x)}}$ so that $-\log \mu(x) = \nu(x)/p(x)$.
- The objective can be stated, ignoring constants that cancel out when taking derivatives, as:

\[
J(q) = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right] \tag{24.62}
\]

\[
= \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] \tag{24.63}
\]

\[
= \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \beta(x, \hat{x}) \tag{24.64}
\]

where

\[
\beta(x, \hat{x}) = \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] \tag{24.65}
\]
After some algebra, we get:

$$\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \quad (24.66)$$
After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \tag{24.66}
\]

Implied that

\[
q(\hat{x}|x) = q(\hat{x}) e^{-\lambda d(x, \hat{x})} \mu(x) \tag{24.67}
\]
After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0
\]

(24.66)

Implying that

\[
q(\hat{x}|x) = q(\hat{x}) e^{-\lambda d(x, \hat{x})\mu(x)}
\]

(24.67)
Computing \( R(D) \)

- After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \tag{24.66}
\]

- Implying that

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} \tag{24.67}
\]
After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \tag{24.66}
\]

Implying that

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}} \tag{24.67}
\]
Computing $R(D)$

After some algebra, we get:

$$\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \quad (24.66)$$

Implying that

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}} \quad (24.67)$$

This expresses $q(\hat{x}|x)$ in terms of $q(\hat{x})$, so if we can solve for $q(\hat{x})$, then we can get $q(\hat{x}|x)$. We first do a little intuition.
The update is:

\[
q(\hat{x}|x) = q(\hat{x})e^{-\lambda d(x, \hat{x})} \mu(x) = q(\hat{x})e^{-\lambda d(x, \hat{x})} \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}
\]

(24.68)

Note that \(\mu(x) = \sum_{\hat{x}} q(\hat{x})e^{-\lambda d(x, \hat{x})}\) since \(\sum_{\hat{x}} q(\hat{x}|x) = 1\). If \(d(x, \hat{x})\) is large, then \(q(\hat{x}|x)\) will be small. Makes sense that we don’t in general want to use \(\hat{x}\) for \(x\) if distortion is large. This, however, is balanced by overall \(q(\hat{x})\) which will force us to start using \(\hat{x}\) for \(x\) if \(q(\hat{x})\) is large.
The update is:

\[ q(\hat{x}|x) = q(\hat{x}) e^{-\lambda d(x,\hat{x})} \mu(x) \]

(24.68)
The update is:

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} \]  

(24.68)
The update is:

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  

(24.68)
The update is:

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}
\]  \hspace{1cm} (24.68)

Note that \(\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}\) since \(\sum_{\hat{x}} q(\hat{x}|x) = 1\).
The update is:

\[
q(\hat{x} | x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}} \tag{24.68}
\]

- Note that \(\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}\) since \(\sum_{\hat{x}} q(\hat{x} | x) = 1\).
- If \(d(x, \hat{x})\) is large, then \(q(\hat{x} | x)\) will be small.
The update is:

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}
\]  \hspace{1cm} (24.68)

Note that \(\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}\) since \(\sum_{\hat{x}} q(\hat{x}|x) = 1\).

If \(d(x, \hat{x})\) is large, then \(q(\hat{x}|x)\) will be small. Makes sense that we don’t in general want to use \(\hat{x}\) for \(x\) if distortion is large.
The update is:

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  (24.68)

- Note that \( \mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})} \) since \( \sum_{\hat{x}} q(\hat{x}|x) = 1 \).
- If \( d(x, \hat{x}) \) is large, then \( q(\hat{x}|x) \) will be small. Makes sense that we don’t in general want to use \( \hat{x} \) for \( x \) if distortion is large.
- This, however, is balanced by overall \( q(\hat{x}) \) which will force us to start using \( \hat{x} \) for \( x \) if \( q(\hat{x}) \) is large.
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$$

(24.70)
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x)e^{-\lambda d(x,\hat{x})}/\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}$$

(24.70)
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right)$$

(24.69)

(24.70)
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right)$$  \hspace{1cm} (24.69)$$

$$= q(\hat{x}) \frac{\sum_x p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (24.70)$$
To solve for $q(\hat{x})$, we find

$$q(\hat{x}) = \sum_x p(x) q(\hat{x}|x),$$

yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y}) e^{-\lambda d(x,\hat{y})}} \right)$$

(24.69)

$$= \frac{q(\hat{x}) \sum_x p(x) e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y}) e^{-\lambda d(x,\hat{y})}}$$

(24.70)
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right)$$  \hspace{1cm} (24.69)$$

$$= \frac{\sum_x p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (24.70)$$

So, for all $\hat{x}$ such that $q(\hat{x}) > 0$ we have

$$C'(\hat{x}) = \sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} = 1$$  \hspace{1cm} (24.71)$$
Computing $R(D)$

- To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right)$$

$$= q(\hat{x}) \frac{\sum_x p(x)e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$ (24.69) (24.70)

- So, for all $\hat{x}$ such that $q(\hat{x}) > 0$ we have

$$C(\hat{x}) = \sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} = 1$$ (24.71)

- Thus, if $q(\hat{x}) > 0$ for all $\hat{x}$, then this defines $|\hat{X}|$ simultaneous equations ($\{C(\hat{x}) = 1\}_{\forall \hat{x}}$) which, along with the distortion constraint equation, can be used to solve the $|\hat{X}|$ unknown quantities ($\{q(\hat{x})\}_{\forall \hat{x}}$), for the current $\lambda$. 

According to Eq. (24.68), as long as $q(\hat{x}) > 0$ then $q(\hat{x}|x) > 0$ as well.
According to Eq. (24.68), as long as $q(\hat{x}) > 0$ then $q(\hat{x}|x) > 0$ as well.

So, we can then choose a $\lambda$ and use it to compute particular point on the $R(D)$ curve.
Computing $R(D)$

- For a given set of values $(\lambda, \{q(\hat{x})\})$, we have

$$D = \sum_{x, \hat{x}} p(x) \frac{q(\hat{x})}{\mu(x)} e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\})$$

(24.72)
Computing $R(D)$

- For a given set of values $(\lambda, \{q(\hat{x})\})$, we have

$$D = \sum_{x, \hat{x}} \frac{p(x)}{\mu(x)} q(\hat{x}) e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\}) \quad (24.72)$$

- It can also be shown that

$$R = - (\lambda D + \sum_x p(x) \log \mu(x)) = R(\lambda, \{q(\hat{x})\}) \quad (24.73)$$

$$= sD + \sum_x p(x) \log 1/\mu(x) \quad \text{where } s = -\lambda \quad (24.74)$$
For a given set of values \((\lambda, \{q(\hat{x})\})\), we have
\[
D = \sum_{x, \hat{x}} \frac{p(x)}{\mu(x)} q(\hat{x}) e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\}) \tag{24.72}
\]

It can also be shown that
\[
R = - (\lambda D + \sum_{x} p(x) \log \mu(x)) = R(\lambda, \{q(\hat{x})\}) \tag{24.73}
\]
\[
= sD + \sum_{x} p(x) \log 1/\mu(x) \quad \text{where } s = -\lambda \tag{24.74}
\]

Since \(s = -\lambda\) determines \(D\), if \(s\) yields a large enough \(D\) we will ultimately get some cases where \(q(\hat{x}) \leq 0\).
For a given set of values \((\lambda, \{q(\hat{x})\})\), we have

\[
D = \sum_{x, \hat{x}} \frac{p(x)}{\mu(x)} q(\hat{x}) e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\})
\]  

(24.72)

It can also be shown that

\[
R = - (\lambda D + \sum_{x} p(x) \log \mu(x)) = R(\lambda, \{q(\hat{x})\})
\]  

(24.73)

\[
= sD + \sum_{x} p(x) \log 1/\mu(x)
\]

where \(s = -\lambda\)  

(24.74)

Since \(s = -\lambda\) determines \(D\), if \(s\) yields a large enough \(D\) we will ultimately get some cases where \(q(\hat{x}) \leq 0\).

In fact, this is sufficient to eliminate all instances of \(\hat{x}\) as we now show.
Theorem 24.5.1

∀s > −∞, for optimal q(ˆx), if q(ˆx|x) = 0 for any one x then q(ˆx|x) = 0 for all x. Thus, that particular ˆx may be deleted from the alphabet.

Proof.

- Lets bring the inequality constraints back in for q(ˆx|x) ≥ 0 for a particular (ˆx, x) pair:

\[ L(q) = J(q) + \gamma q(\hat{x}|x) \]  
(24.75)
Computing $R(D)$

**Theorem 24.5.1**

$\forall s > -\infty$, for optimal $q(\hat{x})$, if $q(\hat{x}|x) = 0$ for any one $x$ then $q(\hat{x}|x) = 0$ for all $x$. Thus, that particular $\hat{x}$ may be deleted from the alphabet.

**Proof.**

- Lets bring the inequality constraints back in for $q(\hat{x}|x) \geq 0$ for a particular $(\hat{x}, x)$ pair:

\[
L(q) = J(q) + \gamma q(\hat{x}|x)
\]  

(24.75)

- Then setting $\frac{\partial L}{\partial p(\hat{y}|x)} = 0$ gives

\[
q(\hat{y}|x) = \begin{cases} 
\frac{q(\hat{y})}{\mu(x)} e^{sd(x,\hat{y})} & \text{if } \hat{y} \neq \hat{x} \\
\frac{q(\hat{x})}{\mu(x)} e^{(sd(x,\hat{x}) + \gamma/p(x))} & \text{if } \hat{y} = \hat{x}
\end{cases}
\]  

(24.76)

...
Proof.

Since $1 = \sum \hat{y} q(\hat{y}|x)$, we have that

$$\mu(x) = \sum \hat{y} q(\hat{y}) e^{f(\hat{y},x)} > 0$$

(24.77)
Proof.

- Since \( 1 = \sum \hat{y} q(\hat{y} | x) \), we have that

\[
\mu(x) = \sum_{\hat{y}} q(\hat{y}) e^f(\hat{y}, x) > 0 \quad (24.77)
\]

- Note that the above inequality is strict since \( e^f(\hat{y}, x) > 0 \) and \( q(\hat{y}) \geq 0 \).
Computing $R(D)$

**Proof.**

- Since $1 = \sum_{\hat{y}} q(\hat{y}|x)$, we have that

\[ \mu(x) = \sum_{\hat{y}} q(\hat{y})e^f(\hat{y},x) > 0 \]  

(24.77)

- Note that the above inequality is strict since $e^f(\hat{y},x) > 0$ and $q(\hat{y}) \geq 0$.
- Thus, $\mu(x) > 0$ strictly for all $x$. 
Proof.

Since $1 = \sum_{\hat{y}} q(\hat{y} | x)$, we have that

$$\mu(x) = \sum_{\hat{y}} q(\hat{y}) e^{f(\hat{y}, x)} > 0 \quad (24.77)$$

Note that the above inequality is strict since $e^{f(\hat{y}, x)} > 0$ and $q(\hat{y}) \geq 0$.

Thus, $\mu(x) > 0$ strictly for all $x$.

If $q(\hat{x} | x) = 0$, it must be due to $q(\hat{x}) = 0$ and not due to either $\mu(x)$ or to the $e^{()}$ factor, as long as $s > -\infty$. 
Proof.

Since $1 = \sum \hat{y} q(\hat{y}|x)$, we have that

$$\mu(x) = \sum_{\hat{y}} q(\hat{y}) e^f(\hat{y},x) > 0$$  \hspace{1cm} (24.77)

Note that the above inequality is strict since $e^f(\hat{y},x) > 0$ and $q(\hat{y}) \geq 0$.

Thus, $\mu(x) > 0$ strictly for all $x$.

If $q(\hat{x}|x) = 0$, it must be due to $q(\hat{x}) = 0$ and not due to either $\mu(x)$ or to the $e^{()}$ factor, as long as $s > -\infty$.

But if $q(\hat{x}) = 0$ then we have $q(\hat{x}|x) = 0$ for all $x$. 


More intuition: From previous definition, we have

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  

(24.78)
More intuition: From previous definition, we have

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}
\]  

(24.78)

If \(q(\hat{x}|x) = 0\) for some \(\hat{x}\), then this must be due to \(q(\hat{x}) = 0\) since nothing else in the definition can be 0.
More intuition: From previous definition, we have

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}
\]  

(24.78)

If \( q(\hat{x}|x) = 0 \) for some \( \hat{x} \), then this must be due to \( q(\hat{x}) = 0 \) since nothing else in the definition can be 0.

We also have a nice meaning for \( s = -\lambda \).
Computing $R(D)$

**Theorem 24.5.2**

The parameter $s = -\lambda$ represents the slope of the rate-distortion function at the point $(D_s, R_s)$ that one generates parametrically from the parametric form above. I.e.

$$R' = \left. \frac{dR}{dD} \right|_{D_s} = s \quad (24.79)$$

**Proof.**

Take derivatives and use the chain rule . . .
Computing $R(D)$

**Theorem 24.5.2**

The parameter $s = -\lambda$ represents the slope of the rate-distortion function at the point $(D_s, R_s)$ that one generates parametrically from the parametric form above. I.e.

$$R' = \frac{dR}{dD} \bigg|_{D_s} = s$$  \hspace{1cm} (24.79)

**Proof.**

Take derivatives and use the chain rule . . .

- Pictorially,
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$. 
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$.

To get the resulting distribution, we need to find the $q(\hat{x})$ values, and if $< 0$ remove symbols, and repeat.
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$.

To get the resulting distribution, we need to find the $q(\hat{x})$ values, and if $< 0$ remove symbols, and repeat.

We continue this process until all are positive.
Computing $R(D)$

- Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$.
- To get the resulting distribution, we need to find the $q(\hat{x})$ values, and if $< 0$ remove symbols, and repeat.
- We continue this process until all are positive.
- If we have only one left, then we have a $R = 0$ case.
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$.

To get the resulting distribution, we need to find the $q(\hat{x})$ values, and if $< 0$ remove symbols, and repeat.

We continue this process until all are positive.

If we have only one left, then we have a $R = 0$ case.

Also, solution to the set of equations might be hard (or an analytical solution might not exist).
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$.

To get the resulting distribution, we need to find the $q(\hat{x})$ values, and if $< 0$ remove symbols, and repeat.

We continue this process until all are positive.

If we have only one left, then we have a $R = 0$ case.

Also, solution to the set of equations might be hard (or an analytical solution might not exist).

Fortunately, there is a better way to do this.