Cumulative Outstanding Reading

- Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
- Read Ch. 9 in our book (Cover & Thomas, “Information Theory”)
- Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page [http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/]).

Additional Reading on Rate-Distortion Theory

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

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

Announcements

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

**Corollary 23.2.7**

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

So unfortunately, feedback in this model is not as useful as we might think it would be.

---

**Coding/Compression and Transmission**

- We know that the source compresses down to the entropy \( H \), but no further.

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

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

Vector Quantization

- We have symbols \( X \in \mathcal{X} \) which could be a continuous or a (say big) discrete domain.
- We quantize this region to \( \hat{\mathcal{X}} \) where \( \hat{\mathcal{X}} \) is discrete and not too big (if \( \mathcal{X} \) is discrete, then \( |\hat{\mathcal{X}}| \ll |\mathcal{X}| \)).
- In above, \( \hat{\mathcal{X}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_7\}, |\hat{\mathcal{X}}| = 7 = M \)
- There are a set of regions \( \mathcal{R}_i, i = 1, \ldots, M \), disjoint so that \( \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \) for \( i \neq j \), and with \( \hat{x}_i \in \mathcal{R}_i \) for all \( i \).
Vector Quantization: What determines quality?

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

- Expected distortion $D = E_{p(x)} d(X, g(f(X)))$.

Rate distortion: set up

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

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

Then $Ed(X, \hat{X}) = \Pr(X \neq \hat{X})$. 
A \((2^nR, n)\) rate distortion code consists of an encoding function
\[ f_n : X^n \rightarrow \{1, 2, \ldots, 2^nR \} \] (23.36)
and a decoding function
\[ g_n : \{1, 2, \ldots, 2^nR\} \rightarrow \hat{X}^n \] (23.37)

(Note, \(H(\hat{X}^n) \leq nR\) since only \(2^nR\) different codewords.)
The distortion of this code is
\[ D = Ed(X_{1:n}, g_n(f_n(X_{1:n}))) = \sum_{x_{1:n} \in X^n} p(x_{1:n})d(x_{1:n}, g_n(f_n(x_{1:n}))) \] (23.38)

Achievability and rate-distortion pairs

Definition 23.2.7
A rate-distortion pair \((R, D)\) is said to be achievable if \(\exists\) a sequence of \((2^nR, n)\) codes \((f_n, g_n)\) with
\[ \lim_{n \rightarrow \infty} Ed(X_{1:n}, g_n(f_n(X_{1:n}))) \leq D \] (23.37)

- \(D\) is the max allowable distortion for code at this rate \(R\).
- We can make errors, but not too many (bounded average distortion).
- The type of errors we can make is entirely dependent on the distortion function.
- Def: A rate distortion region for a source is the closure of achievable rate distortion pairs \((R, D)\)
- Def: A rate distortion function \(R(D)\) is the infimum of rates \(R\) such that \((R, D)\) is in rate distortion region. I.e.,
\[ R(D) = \inf \{ R : (R, D) \text{ is achievable} \} \] (23.38)
Def: A distortion rate function $D(R)$ is the infimum of distortions $D$ such that $(R, D)$ is in rate distortion region. I.e.,

$$D(R) = \inf \{ D : (R, D) \text{ is achievable} \} \quad (23.37)$$

The next definition is very important

**Definition 23.2.7**

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

$$R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum_x x \hat{x} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \quad (23.38)$$

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

**Intuition: Information Rate Distortion Function**

$$R^{(I)}(D) = \min_{p(\hat{x}|x) : \sum_x x \hat{x} p(x)p(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \quad (23.38)$$

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

$$I(X; \hat{X}) = I(X; X) = H(X) \quad (23.39)$$

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

$$R^{(I)}(D) = H(X) \quad (23.40)$$
Theorem 23.3.1

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:

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

(23.1)

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

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

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

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

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \tag{23.2}$$
$$= H(X) - H(X \oplus \hat{X}|\hat{X}) \tag{23.3}$$
$$\geq H(p) - H(X \oplus \hat{X}) \tag{23.4}$$
$$= H(p) - H(\text{Pr}\{X \neq \hat{X}\}) \tag{23.5}$$
$$\geq H(p) - H(D) \tag{23.6}$$
rate-distortion for Bernoulli r.v.

Proof of Theorem 23.3.1.

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

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

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

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

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

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

\[
 D \leq p < 1/2 \text{ rate-distortion for Bernoulli r.v.}
\]

Proof of Theorem 23.3.1.

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

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

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

\[
p = \Pr(X = 1) = \Pr(X = 1|\hat{X} = 0)\Pr(\hat{X} = 0) + \Pr(X = 1|\hat{X} = 1)\Pr(\hat{X} = 1)
= D(1 - \Pr(\hat{X} = 1)) + (1 - D)\Pr(\hat{X} = 1) \tag{23.10}
\]

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

\[
 \Pr(\hat{X} = 1) = \frac{p - D}{1 - 2D} \leq p \quad \text{if } 0 \leq D < p \leq 1/2 \tag{23.11}
\]
Aside: rate-distortion for Bernoulli r.v.

- Assume $D < p \leq 1/2$ and that $\Pr(\hat{X} = 1) = (p - D)/(1 - 2D)$.
- Then $p \geq (p - D)/(1 - 2D)$ since

$$
\frac{p - D}{1 - 2D} - p = \frac{D(2p - 1)}{1 - 2D} \leq 0 \quad (23.12)
$$

Proof of Theorem 23.3.1.

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

$$
H(p) - H(D) \leq I(X; \hat{X}) = H(X) - H(X|\hat{X}) \quad (23.13)
$$
$$
= H(p) - H(X|\hat{X}) \quad (23.14)
$$
$$
= H(\hat{X}) - H(\hat{X}|X) \quad (23.15)
$$
$$
= H(\hat{X}) - H(D) \quad (23.16)
$$
$$
\leq H(p) - H(D) \quad (23.17)
$$

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

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

**Proof of Theorem 23.3.1.**

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

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

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

\[
= p
\]

**Key Theorem**

**Theorem 23.3.2**

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

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

- This means that the minimum coding rate for achieving distortion \( D \) is, perhaps now unsurprisingly, \( R^{(I)}(D) \).
- Two things to prove: (1) that if \( (R, D) \) is achievable, than \( R > R^{(I)}(D) \), and (2) if \( R > R^{(I)}(D) \), then there exists a sequence of codes that can achieve rate-distortion pair \( (R, D) \).
- Like in channel capacity and entropy compression case, what happens at \( R = R^{(I)}(D) \) depends on the very specific case that one is analyzing.
- For now, let’s look at Gaussian sources.
Gaussian Channels

Theorem 23.4.1

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

$$R^{(I)}(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & \text{if } 0 \leq D \leq \sigma^2 \\ 0 & \text{otherwise.} \end{cases} \quad (23.23)$$

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

Proof of Theorem 23.4.1

Proof.

- We have that

$$R^{(I)}(D) = \min_{f(\hat{x}|x):E(\hat{X}-X)^2 \leq D} I(X;\hat{X}) \quad (23.24)$$

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

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

$$\geq \frac{1}{2} \log((2\pi e)\sigma^2) - h(X - \hat{X}) \quad (23.25)$$

$$\geq \frac{1}{2} \log((2\pi e)\sigma^2) - h(\mathcal{N}(0, E(X - \hat{X})^2)) \quad (23.26)$$

$$\geq \frac{1}{2} \log((2\pi e)\sigma^2) - \frac{1}{2} \log((2\pi e)D) \quad (23.27)$$

$$= \frac{1}{2} \log(\sigma^2/D) \quad (23.28)$$
Proof of Theorem 23.4.1

Proof.

- Thus, \( R(D) \geq \frac{1}{2} \log(\sigma^2/D) \)
- Like before, we construct \( f(x) \) and \( f(\hat{x}|x) \) to achieve equality/tightness in the lower bound & distortion \( D \).
- We define it as follows, where \( \hat{X} \perp\!\!\!\!\perp Z \):

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

\( \hat{X} \sim \mathcal{N}(0, \sigma^2 - D) \quad \rightarrow \quad X \sim \mathcal{N}(0, \sigma^2) \)

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

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

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

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

If \( D > \sigma^2 \) we can choose \( \hat{x} = 0 \) w.p.1 for a zero rate code.

To summarize, we then get:

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

(23.31)

As always to keep in mind: if \( D > \sigma^2 \) then we can use a rate of \( R = 0 \). If \( D < \sigma^2 \) then we need to allocate some bits.
Example: Multiple Gaussians Unequal Noise

- What would be the rate for multiple Gaussians with different noise? I.e., given $X_{1:m}$ with $X_i \sim \mathcal{N}(0, \sigma_i^2)$ and with $X_i \perp X_j$ for all $i \neq j$, and no requirement for the $\{\sigma_i^2\}_i$’s to be equal.
- Overall distortion is of the form $d(x_{1:m}, \hat{x}_{1:m}) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$ with $E_p(x_{1:m}, \hat{x}_{1:m})[d(X_{1:m}, \hat{X}_{1:m})] \leq D$ where $D$ is overall distortion constraint.
- Information rate distortion function has form:

$$R(D) = \min_{f(\hat{x}_{1:m}|x_{1:m}):E[d(X_{1:m}, \hat{X}_{1:m})] \leq D} I(X_{1:m}; \hat{X}_{1:m}) \quad (23.32)$$

- We need to know how many bits to allocate to each source symbol (and how much “local distortion to use”) to achieve given overall distortion $D$. Any guesses?

We expand MI as follows:

$$I(X_{1:m}; \hat{X}_{1:m}) = h(X_{1:m}) - h(X_{1:m}|\hat{X}_{1:m}) \quad (23.33)$$

$$= \sum_i h(X_i) - \sum_i h(X_i|\hat{X}_{1:m}, X_{1:i-1}) \quad (23.34)$$

$$\geq \sum_i h(X_i) - \sum_i h(X_i|\hat{X}_i) \quad (23.35)$$

$$= \sum_i I(X_i; \hat{X}_i) \quad (23.36)$$

$$\geq \sum_i R(D_i) = \sum_i \max \left\{ \frac{1}{2} \log \frac{\sigma_i^2}{D_i}, 0 \right\} \quad (23.37)$$

where $D_i = E(X_i - \hat{X}_i)^2$. 

Example: Multiple Gaussians Unequal Noise

- To achieve equality, we set
  \[ f(\hat{x}_{1:m}|x_{1:m}) = \prod_{i} f(\hat{x}_i|x_i) \]  

(23.38)

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

(23.39)

- Thus, the problem becomes:

The problem becomes:

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

(23.40)

- This is a convex minimization problem and can be written as:

minimize over \( \{R_i\}, \{D_i\} \)

\[ \sum_{i=1}^{m} R_i \]  

(23.41)

subject to

\[ \sum_{i} D_i = D \]  

(23.42)

\[ R_i \geq \frac{1}{2} \log \frac{\sigma_i^2}{D_i} \quad \forall i \]  

(23.43)

\[ R_i \geq 0 \quad \forall i \]  

(23.44)
Example: Multiple Gaussians Unequal Noise

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

$$J(D) = \sum_{i=1}^{m} \left( \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda D_i \right) \quad (23.45)$$

$$\Rightarrow \frac{\partial J}{\partial D} = -\frac{1}{2} \frac{1}{D_i} + \lambda = 0 \quad (23.46)$$

$$\Rightarrow D_i = \frac{1}{2\lambda} = \lambda' \forall i \quad (23.47)$$

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

- That is, when $\lambda' = D/m < \sigma_{\text{min}}^2$ where $\sigma_{\text{min}}^2 = \min_i \sigma_i^2$, then $\lambda < \sigma_i^2$ for all $i$, and we can have $R_i > 0$ for all $i$.

- As total distortion $D$ increases, $\lambda$ will also increase eventually hitting one or more of the $\sigma_i^2$ values (i.e., $\lambda = \min_i \sigma_i^2$). This will shut off some source symbols as we’ll get $R_i = 0$ for them.

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

- We get

**Theorem 23.4.2**

Given parallel Gaussian source $X_i \sim \mathcal{N}(0, \sigma_i^2)$ i.i.d., under squared loss $d(x_1:m, \hat{x}_1:m) = \sum_i (x_i - \hat{x}_i)^2$, we have

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i} = \sum_{i=1}^{m} R_i \quad (23.48)$$

where

$$D_i = \begin{cases} \lambda & \text{if } \lambda < \sigma_i^2 \ (\Rightarrow R_i > 0) \\ \sigma_i^2 & \text{if } \lambda \geq \sigma_i^2 \ (\Rightarrow R_i = 0) \end{cases} = \min(\lambda, \sigma_i^2) \quad (23.49)$$

and where $\lambda$ is chosen so that $\sum_i D_i = D$. 

---

Prof. Jeff Bilmes  
EE515a/Winter 2014/Information Theory II – Lecture 23 - Jan 22nd, 2014  
L23 F35/59 (pg.35/59)
Example: Multiple Gaussians Unequal Noise

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

Rate-Distortion Theorem: Converse

- Converse of Theorem 23.3.2 states that if $\{X_i\}_i$ is an i.i.d. source with probability distribution $X_i \sim p(x)$, and $d(x, \hat{x})$ is a distortion measure, than any $(2^{nR}, n)$ code with average distortion

$$E[d(X^n, \hat{X}^n)] = \frac{1}{n} \sum_{i=1}^{n} E[d(X_i, \hat{X}_i)] \leq D \quad (23.50)$$

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 $P_e \rightarrow 0$, we can’t compress lower than the entropy.
Lemma 23.5.1

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

Proof.

- First, as $D \uparrow$, we are taking the minimum over a larger set so necessarily $R^{(I)}(D) \downarrow$ as $D \uparrow$.
- Now, consider $(R_1, D_1)$ and $(R_2, D_2)$ on $R$-$D$ curve of $R^{(I)}(D)$ with, respectively, $p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x)$ and $p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x)$ being two distributions that achieve those pairs respectively.
- Mix them, $p_\lambda = \lambda p_1 + (1 - \lambda)p_2$ which achieves distortion $D_\lambda = \lambda D_1 + (1 - \lambda)D_2 = \sum_{x, \hat{x}} p(x)p_\lambda(\hat{x}|x)d(x, \hat{x})$.
- Recall mutual information is convex in conditional distribution for fixed $p(x)$.
- Hence, $I_{p_\lambda}(X; \hat{X}) \leq \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda)I_{p_2}(X; \hat{X})$ (23.51)

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

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

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

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

Not surprising, shapes we’ve seen so far are of the form:
Proof of converse

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

proof of converse.

- Reminder: given a \((2^n R, n)\) code defined by functions \(f_n\) and \(g_n\), the reproduction of sequence \(X^n\) is given by:
  \[
  \hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n))
  \]  
  \[23.54\]

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

Proof of converse

proof of converse.

\[nR \geq H(\hat{X}^n) = H(\hat{X}^n) - H(\hat{X}^n|X^n) = I(\hat{X}^n; X^n) \]  
\[23.55\]

\[= H(X^n) - H(X^n|\hat{X}^n) = \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \]  
\[23.56\]

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

\[= n \sum_{i=1}^{n} \frac{1}{n} R^{(I)}(Ed(X_i, \hat{X}_i)) \geq n R^{(I)}(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \]  
\[23.58\]

\[= n R^{(I)}(Ed(X^n, \hat{X}^n)) = n R^{(I)}(D) \]  
\[23.59\]

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

**Theorem 23.5.2 (Achievability in 23.3.2)**

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

Typicality lives

**Definition 23.5.3 (distortion $\epsilon$-typical)**

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

\[
\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \quad \text{x-typical} \\
\left| -\frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \quad \hat{x}\text{-typical} \\
\left| -\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \quad \text{jointly typical} \\
|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| \leq \epsilon \quad \text{new, “distortion typical”}
\]

Any $x$ s.t. Equations (23.60)-(23.63) are true define the set $A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)}$. 
**Probability of typicality**

**Lemma 23.5.4**

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

**Proof.**

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

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

**Main Theorem: Achievability**

**Proof of achievability in 23.3.2.**

- We show that we can construct a random code, and use joint typicality to bound the probability of error as \(n \to \infty\).
- Fix \(p(\hat{x}|x)\) and then calculate \(p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)\).
- Chose \(\epsilon > 0\) and \(\delta > 0\).
- We will show that for any \(R > R^{(1)}(D)\), 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 23.3.2.

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 $\hat{x}^n(w)$.

Distortion:
- Average distortion over both codebooks and codewords:
  \[ \bar{D} = E_{X^n,\hat{C}}d(X^n, \hat{X}^n) = \sum_{\hat{C},x^n} \Pr(\hat{C})p(x^n)d(x^n, \hat{x}^n) \]  
  (23.64)
- In the above, we take expectation over both random choice of codebooks $\hat{C} = \{\hat{x}^n(1), \hat{x}^n(2), \ldots, \hat{x}^n(2^{nR})\}$ based on probability model $\Pr(\hat{C})$, and also random source strings based on $p(x^n)$. ...
Main Theorem: Achievability

...proof of achievability in 23.3.2.

- then, chose $\epsilon > 0$ and divide sequences $x^n$ into two categories, A and B as below:
  - Category A: $x^n : \exists \hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in A_{d,\epsilon}^{(n)}$ so that $d(x^n, \hat{x}^n(w)) < D + \epsilon$. The probability of these sequences is $\Pr(A_{d,\epsilon}^{(n)}) \to 1$.
  - 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

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

(23.65)

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 23.3.2.

General idea first:
- This gives
  \[ P_e \leq \epsilon + (e^2)^{-n(R-I(X;\hat{X}))-3\epsilon} \]  
  (23.68)
- So for any \( \delta > 0 \) \( \exists \epsilon, n \) s.t. over all randomly chosen rate \( R \) codes of block length \( n \), the expected distortion \( < D + \delta \).
- This means there must be at least one code \( C^* \) with this rate, block-length, and distortion.
- \( \delta \) is arbitrary \( \Rightarrow (R, D) \) is achievable if \( R > R^{(I)}(D) \).

**Subsidiary Theorems**

**Theorem 23.5.5**

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

\[ p(\hat{x}^n) \geq p(x^n|\hat{x}^n)2^{-n(I(X;\hat{X})+3\epsilon)} \]  
(23.69)

**Proof.**

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

\[ p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = \frac{p(x^n) p(\hat{x}^n, x^n)}{p(x^n) p(\hat{x}^n)} \]  
(23.70)

\[ \leq p(\hat{x}^n) \frac{2^{-n(H(X;\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} \]  
(23.71)

\[ = p(\hat{x}^n)2^{n(I(X;\hat{X})+3\epsilon)} \]  
(23.72)
Subsidiary Theorems

**Theorem 23.5.6**

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

$$(1 - xy)^n \leq 1 - x + e^{-yn} \quad (23.73)$$

**Proof.**

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

... proof continued.

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

  $$(1 - xy)^n = g_y(x) = g_y((1 - x) \cdot 0 + x \cdot 1) \quad (23.74)$$

  $$\leq (1 - x)g_y(0) + xg_y(1) \quad (23.75)$$

  $$= (1 - x) \cdot 1 + x \cdot (1 - y)^n \quad (23.76)$$

  $$\leq 1 - x + xe^{-y} \quad (23.77)$$

  $$\leq 1 - x + e^{-yn} \quad (23.78)$$
Main Theorem: Achievability

...proof of achievability in 23.3.2.

- Next, we calculate $P_e$ for a randomly chosen source sequence and randomly chosen codebook where there exists no codeword that is distortion typical with the source sequence.
- The set of source sequences s.t. there is at least one codeword in $C$ that is distortion typical with it, is defined as:

$$J(C) = \left\{ x^n : \exists \hat{x}^n \in C \text{ s.t. } (x^n, \hat{x}^n) \in A^{(n)}_{d, \epsilon} \right\} \quad (23.79)$$

- Then, an expression for $P_e$ follows next ...

\[
P_e = \sum_{C} \Pr(C) \sum_{x^n : x^n \notin J(C)} p(x^n) \quad (23.80)
\]
\[
= \sum_{x^n} p(x^n) \sum_{C : x^n \notin J(C)} \Pr(C) \quad (23.81)
\]
\[
= \sum_{x^n} p(x^n) \left\{ \text{total prob of all } 2^nR \text{ current } C \text{ codewords not being distortion typical with current } x^n \text{ (i.e., prob. of choosing codebook not good for current } x^n) \right\} \quad (23.82)
\]
\[
= \sum_{x^n} p(x^n) q^{2^nR} \quad (23.83)
\]

where $q$ is the probability that a single random codeword is not jointly typical with the source sequence.
Main Theorem: Achievability

...proof of achievability in 23.3.2.

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

- Then

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

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

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

This last line follows from Theorem 23.5.5.

...
Main Theorem: Achievability

proof of achievability in 23.3.2.

Now

\[ 1 - \sum_{x^n, \hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) \]  
(23.91)

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

Also

\[ \exp(-2^n(R-I(X;\hat{X})-3\epsilon)) \to 0 \]  
(23.92)

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