Lecture 7
Lossy Source Coding

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

December 9, 2015
Recall: in Lecture 03, we investigated the fundamental limit of (almost) lossless block-to-block (or fixed-to-fixed) source coding.

- The recovery criterion is vanishing probability of error:

\[ \lim_{N \to \infty} P \left\{ S^N \neq \hat{S}^N \right\} = 0. \]

- The minimum compression ratio to fulfill lossless reconstruction is the entropy rate of the source:

\[ R^* = H(\{S_i\}), \quad \text{for stationary and ergodic } \{S_i\}. \]
In this lecture, we turn our focus to **lossy** block-to-block source coding, where the setting is the same as before, except

- The recovery criterion is **reconstruction to within a given distortion** $D$:

$$\limsup_{N \to \infty} \mathbb{E} \left[ d \left( S^N, \hat{S}^N \right) \right] \leq D.$$ 

- The minimum compression ratio to fulfill reconstruction to within a given distortion $D$ is the **rate-distortion function**:

$$R(D) = \min_{p_{S|\hat{S}}: \mathbb{E}[d(S,\hat{S})] \leq D} I \left( S; \hat{S} \right), \quad \text{for DMS} \ \{S_i\}.$$
Why lossy source coding?

- Sometimes it might be too expensive to reconstruct the source in a lossless way.

- Sometimes it is impossible to reconstruct the source losslessly. For example, if the source is continuous-valued, the entropy rate of the source is usually infinite!

- Lossy source coding has wide range of applications, including quantization/digitization of continuous-valued signals, image/video/audio compression, etc.

In this lecture, we first focus on discrete memoryless sources (DMS). Then, we employ the discretization technique to extend the coding theorems from the discrete-source case to the continuous-source case. In particular, Gaussian sources will be our main focus.
Lossless vs. Lossy Source Coding

The general lossy source coding problem involves quantizing all possible source sequences \( s^N \in S^N \) into \( 2^K \) reconstruction sequences \( \hat{s}^N \in \hat{S}^N \), which can be represented by \( K \) bits.

The goal is to design the correspondence between \( s^N \) and \( \hat{s}^N \) so that the distortion (quantization error) is below a prescribed level \( D \).

Lossy source coding has a couple of notable differences from lossless source coding:

- Source alphabet \( S \) and the reconstruction alphabet \( \hat{S} \) could be different in general.
- Performance is determined by the chosen distortion measure.
1 Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2 Proof of the Coding Theorem
   - Converse Proof
   - Achievability
1 Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2 Proof of the Coding Theorem
   - Converse Proof
   - Achievability
Distortion Measures

We begin with the definition of the distortion measure per symbol.

**Definition 1 (Distortion Measure)**

A per-symbol distortion measure is a mapping \( d(s, \hat{s}) \) that maps from \( S \times \hat{S} \) to \([0, \infty)\), and it is understood as the cost of representing \( s \) by \( \hat{s} \).

For two length \( N \) sequences \( s^N \) and \( \hat{s}^N \), the distortion between them is defined as the average of the per-symbol distortion:

\[
d( s^N, \hat{s}^N ) \triangleq \frac{1}{N} \sum_{i=1}^{N} d(s_i, \hat{s}_i).
\]

**Examples**: below are two widely used distortion measures:

- **Hamming distortion**: \( S = \hat{S}, \ d(s, \hat{s}) \triangleq 1 \{ s \neq \hat{s} \}. \)
- **Squared-error distortion**: \( S = \hat{S} = \mathbb{R}, \ d(s, \hat{s}) \triangleq (s - \hat{s})^2. \)
Lossy Source Coding: Problem Setup

1. A \((2^{NR}, N)\) source code consists of
   - an encoding function (encoder) \(\text{enc}_N : S^N \rightarrow \{0, 1\}^K\) that maps each source sequence \(s^N\) to a bit sequence \(b^K\), where \(K \triangleq \lfloor NR \rfloor\).
   - a decoding function (decoder) \(\text{dec}_N : \{0, 1\}^K \rightarrow \hat{S}^N\) that maps each bit sequence \(b^K\) to a reconstructed source sequence \(\hat{s}^N\).

2. The expected distortion of the code \(D^{(N)} \triangleq \mathbb{E} \left[ d \left( S^N, \hat{S}^N \right) \right]\).

3. A rate-distortion pair \((R, D)\) is said to be achievable if there exist a sequence of \((2^{NR}, N)\) codes such that \(\limsup_{N \rightarrow \infty} D^{(N)} \leq D\).

   The optimal compression rate \(R(D) \triangleq \inf \{ R \mid (R, D) : \text{achievable} \}\).
**Rate Distortion Trade-off**

\[
D_{\text{min}} \triangleq \min_{\hat{s}(s)} \mathbb{E} [d(S, \hat{s}(S))]
\]

It denotes the minimum possible target distortion so that the rate is finite.

Even the decoder knows the entire \(s^N\) and finds a best representative \(\hat{s}^N(N)\), the expected distortion is still \(D_{\text{min}}\).

\[
D_{\text{max}} \triangleq \min_{\hat{s}} \mathbb{E} [d(S, \hat{s})]
\]

Let \(\hat{s}^* \triangleq \arg \min_{\hat{s}} \mathbb{E} [d(S, \hat{s})]\). Then for target distortion \(D \geq D_{\text{max}}\), we can use a single representative \(\hat{s}^* \triangleq (\hat{s}^*, \hat{s}^*, \ldots, \hat{s}^*)\) to reconstruct all \(s^N \in S^N\) (rate is 0!), and

\[
D^{(N)} = \mathbb{E} [d(S^N, \hat{s}^*)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} [d(S_i, \hat{s}^*)] = D_{\text{max}} \leq D.
\]

Hence, \(R(D) = 0\) for all \(D \geq D_{\text{max}}\).
Theorem 1 (A Lossy Source Coding Theorem for DMS)

For a discrete memoryless source \( \{S_i | i \in \mathbb{N}\} \),

\[
R(D) = \min_{p_{\hat{S}|S} : \mathbb{E}[d(S,\hat{S})] \leq D} I(S;\hat{S}).
\] (1)

Interpretation:

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

Uncertainty of source \( S \)  
Uncertainty of \( S \) after learning \( \hat{S} \)  
The rate used in compressing \( S \) to \( \hat{S} \)
1 Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2 Proof of the Coding Theorem
   - Converse Proof
   - Achievability
Properties of Rate Distortion Function

A rate distortion function \( R(D) \) satisfies the following properties:

1. Nonnegative
2. Non-increasing in \( D \).
3. Convex in \( D \).
4. Continuous in \( D \).
5. \( R(D_{\text{min}}) \leq H(S) \).
6. \( R(D) = 0 \) if \( D \geq D_{\text{max}} \).

These properties are all quite intuitive.

Below we sketch the proof of these properties.
**Monotonicity**  Clear from the definition.

**Convexity**

- The goal is to prove that $D_1, D_2 \geq D_{\min}$ and $\lambda \in (0, 1)$, $\bar{\lambda} \triangleq 1 - \lambda$,

$$R \left( \lambda D_1 + \bar{\lambda} D_2 \right) \leq \lambda R(D_1) + \bar{\lambda} R(D_2).$$

- Let $p_i(\hat{s}|s) \triangleq \arg \min_{p_{\hat{s}|S}: \mathbb{E}[d(S, \hat{S})] \leq D_i} I(S; \hat{S})$, the optimizing conditional distribution that achieves distortion $D_i$, for $i = 1, 2$. Let $p_\lambda \triangleq \lambda p_1 + \bar{\lambda} p_2$.

- Under $p_\lambda(\hat{s}|s)$, the expected distortion between $S$ and $\hat{S} \leq \lambda D_1 + \bar{\lambda} D_2$,

$$\mathbb{E}_{p(S)p_\lambda(\hat{s}|s)} \left[ d(S, \hat{S}) \right] = \sum_s \sum_{\hat{s}} p(s) \left[ \lambda p_1(\hat{s}|s) + \bar{\lambda} p_2(\hat{s}|s) \right] d(s, \hat{s}).$$

- Proof is complete since $I(S; \hat{S})$ is convex in $p_{S|\hat{S}}$ with a fixed $p_S$:

$$R \left( \lambda D_1 + \bar{\lambda} D_2 \right) \leq I(S; \hat{S})_{p_\lambda} \leq \lambda I(S; \hat{S})_{p_1} + \bar{\lambda} I(S; \hat{S})_{p_2} = \lambda R(D_1) + \bar{\lambda} R(D_2).$$
Nonnegativity Clear from the definition.

Continuity It is well-known that convexity within an open interval implies continuity within that open interval.

\[
\begin{align*}
R & \uparrow \\
D_{\text{min}} & \quad D_{\text{max}} \\
\end{align*}
\]

\[
\begin{align*}
R & \uparrow \\
D_{\text{min}} & \quad D_{\text{max}} \\
\end{align*}
\]

violates convexity

violates convexity

discontinuity is possible here

Hence, the only point where \( R(D) \) might be discontinuous is at the boundary \( D = D_{\text{min}} \). The proof is technical and can be found in Gallager[2].
Example: Bernoulli Source with Hamming Distortion

Source (binary) \( S_i \in \mathcal{S} = \{0, 1\} \), and \( S_i \overset{i.i.d.}{\sim} \text{Ber}(p) \ \forall \ i. \)

Distortion (Hamming) \( d(s, \hat{s}) = 1 \{s \neq \hat{s}\}. \)

Example 1

Derive the rate distortion function of the Bernoulli \( p \) source with Hamming distortion and show that it is given by

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

This is the first example about how to compute the rate distortion function, that is, how to solve (1) in the lossy source coding theorem.
sol. The first step is to identify $D_{\text{min}}$ and $D_{\text{max}}$.

$$D_{\text{min}} = 0$$ because one can choose $\hat{s}(s) = s$.

$$D_{\text{max}} = \min(p, 1 - p)$$ because one can choose $\hat{s} = \begin{cases} 0 & p \leq \frac{1}{2} \\ 1 & p \geq \frac{1}{2} \end{cases}$.

The next step is to lower bound $I(S; \hat{S}) = H(S) - H(S|\hat{S})$.

It is equivalent to upper bounding $H(S|\hat{S})$:

$$H(S|\hat{S}) = H(S \oplus \hat{S}|\hat{S}) \leq H(S \oplus \hat{S}) = H_b(q),$$

where we assume that $S \oplus \hat{S} \sim \text{Ber}(q)$ for some $q \in [0, 1]$.

Observe that $d(S, \hat{S}) \equiv S \oplus \hat{S}$. Hence, $\mathbb{E}[d(S, \hat{S})] \leq D \implies q \leq D$.

Since $D \leq D_{\text{max}} \leq \frac{1}{2}$, we see that $H_b(q)$ is maximized when $q = D$.

Hence, $I(S; \hat{S}) \geq H_b(p) - H_b(D)$. 
Final step: show that the lower bound $H_b(p) - H_b(D)$ can be attained.

The goal is to find a probability transition matrix $p(\hat{s}|s)$ such that

$$\hat{S} \perp S \oplus \hat{S} \quad \text{so that} \quad H(S \oplus \hat{S}|\hat{S}) = H(S \oplus \hat{S}) \quad \text{and} \quad P\{S \oplus \hat{S} = 1\} = D.$$

At first glance this looks hard.

The difficulty can be resolved via an auxiliary reverse channel.

Consider a channel with input $\hat{S}$, output $S$, additive noise $Z \sim \text{Ber}(D) \perp \hat{S}$.

$$S = \hat{S} \oplus Z \implies Z = S \oplus \hat{S}.$$

The reverse channel specifies the joint distribution $p(s, \hat{s})$ and hence $p(\hat{s}|s)$!

$$p = (1 - \alpha)D + \alpha(1 - D) \implies \alpha = \frac{p - D}{1 - 2D}$$
Example: Gaussian Source with Squared Error Distortion

**Source** (Gaussian) \( S_i \in \mathcal{S} = \mathbb{R} \), and \( S_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \ \forall \ i. \)

**Distortion** (Squared Error) \( d(s, \hat{s}) = |s - \hat{s}|^2. \)

**Example 2**

Derive the rate distortion function of the Gaussian source with squared error distortion and show that it is given by

\[
R(D) = \begin{cases} 
\frac{1}{2} \log \left( \frac{\sigma^2}{D} \right) , & 0 \leq D \leq \sigma^2 \\
0 , & D > \sigma^2 
\end{cases}
\]

**Remark:** Although the source is continuous, one can use weak typicality or the discretization method used in channel coding to extend the lossy source coding theorem from discrete memoryless sources to continuous.

**Note:** In particular, note that \( R(0) = \infty \), which is quite intuitive!
**sol.** First step: identify $D_{\text{min}}$ and $D_{\text{max}}$.

\[
D_{\text{min}} = 0 \quad \text{because one can choose } \hat{s}(s) = s.
\]
\[
D_{\text{max}} = \sigma^2 \quad \text{because one can choose } \hat{s} = \mu, \text{ the mean of } S.
\]

Next step: lower bound $I\left(S; \hat{S}\right) = h(S) - h\left(S \mid \hat{S}\right)$.

It is equivalent to upper bounding $h\left(S \mid \hat{S}\right)$:

\[
h\left(S \mid \hat{S}\right) = h\left(S - \hat{S} \mid \hat{S}\right) \leq h\left(S - \hat{S}\right) \leq \frac{1}{2} \log (2\pi e D),
\]

where the last inequality holds since $\text{Var} \left[ S - \hat{S} \right] \leq \mathbb{E} \left[ |S - \hat{S}|^2 \right] \leq D$.

Hence, $I\left(S; \hat{S}\right) \geq \frac{1}{2} \log (2\pi e \sigma^2) - \frac{1}{2} \log (2\pi e D) = \frac{1}{2} \log \left( \frac{\sigma^2}{D} \right)$.
Final step: show that the lower bound $\frac{1}{2} \log \left( \frac{\sigma^2}{D} \right)$ can be attained.

The goal is to find a conditional distribution $p(\hat{s}|s)$ such that

$$\hat{S} \perp (S - \hat{S})$$

so that $h(S - \hat{S}|\hat{S}) = h(S - \hat{S})$ and $(S - \hat{S}) \sim \mathcal{N}(0, D)$.

Again, this can be done via an auxiliary reverse channel.

Consider a channel with input $\hat{S}$, output $S$, additive noise $Z \sim \mathcal{N}(0, D) \perp \hat{S}$.

$$S = \hat{S} + Z \implies Z = S - \hat{S}.$$  

The reverse channel specifies the joint distribution $p(s, \hat{s})$ and hence $p(\hat{s}|s)$!

\[
\begin{align*}
Z & \sim \mathcal{N}(0, D) \\
\hat{X} & \sim \mathcal{N}(\mu, \sigma^2 - D) \perp (X - \hat{X}) \\
X & \sim \mathcal{N}(\mu, \sigma^2)
\end{align*}
\]
Example: Source Alphabet $\neq$ Reconstruction Alphabet

**Source** (ternary) \[ S_i \in \mathcal{S} = \{0, *, 1\}, \text{ and } S_i \overset{\text{i.i.d.}}{\sim} p_S \forall i, \text{ where } p_S(0) = p_S(1) = \varepsilon \leq \frac{1}{2}. \]

**Reconstruction** (binary) \[ \hat{\mathcal{S}} = \{0, 1\}. \]

**Distortion** \[ d(s, \hat{s}) = \begin{cases} 1 & \text{if } s \neq * \text{ and } s \neq \hat{s} \\ 0 & \text{if } s = * \text{ or } \hat{s} \end{cases}. \]

In other words, there is a *don’t-care* symbol $*$, and $\mathcal{S} \neq \hat{\mathcal{S}}$.

---

**Example 3**

(HW5) Derive the rate distortion function and show that it is given by

\[ R(D) = \begin{cases} 2\varepsilon \left(1 - H_b\left(\frac{D}{2\varepsilon}\right)\right), & 0 \leq D \leq \varepsilon \\ 0, & D > \varepsilon \end{cases}. \]
1 Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2 Proof of the Coding Theorem
   - Converse Proof
   - Achievability
1 Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2 Proof of the Coding Theorem
   - Converse Proof
   - Achievability
Proof of the Converse of Theorem 1

We aim to show that for any sequence of \((2^{NR}, N)\) source codes with 
\(\limsup_{N \to \infty} D(N) \leq D\), the rate \(R\) must satisfy 
\(R \geq R(D)\) (defined in (1)).

We begin with similar steps as in lossless source coding (cf. Lecture 03).

\textbf{pf:} Note that \(B^K\) is a r.v. because it is generated by another r.v, \(S^N\).

\[
K = NR \geq H(B^K) \geq I(B^K; \hat{S}^N) \overset{(a)}{=} I(S^N; \hat{S}^N) \\
\overset{(b)}{=} \sum_{i=1}^{N} I(S_i; \hat{S}^N | S^{i-1}) \overset{(c)}{=} \sum_{i=1}^{N} I(S_i; \hat{S}^N, S^{i-1}) \geq \sum_{i=1}^{N} I(S_i; \hat{S}_i)
\]

- (a) is due to \(S^N - B^K - \hat{S}^N\) and the data processing inequality.
- (b) is due to Chain Rule. (c) is due to \(S_i \perp S^{i-1}\) (memoryless source).

So far, we have not yet used the condition on distortion.
Further working on the inequality:

\[ NR \geq \sum_{i=1}^{N} I\left(S_i; \hat{S}_i\right) \]

\[ \geq \sum_{i=1}^{N} R\left(\mathbb{E}\left[d\left(S_i, \hat{S}_i\right)\right]\right) = N \sum_{i=1}^{N} \frac{1}{N} R\left(\mathbb{E}\left[d\left(S_i, \hat{S}_i\right)\right]\right) \]

\[ \geq NR \left(\sum_{i=1}^{N} \frac{1}{N} \mathbb{E}\left[d\left(S_i, \hat{S}_i\right)\right]\right) = NR \left(\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} d\left(S_i, \hat{S}_i\right)\right]\right) \]

\[ = NR \left(\mathbb{E}\left[d\left(S^N, \hat{S}^N\right)\right]\right) = NR \left(D^{(N)}\right). \]

- (d) is due to the definition of \( R(D) \) in (1).
- (e) is due to the convexity of \( R(D) \) and Jensen’s inequality.

Hence, \( R \geq \limsup_{N \to \infty} R\left(D^{(N)}\right) \geq R\left(\limsup_{N \to \infty} D^{(N)}\right) \geq R(D). \)

- (f) is due to continuity of \( R(D) \).
- (g) is due to \( \limsup_{N \to \infty} D^{(N)} \leq D \) and \( R(D) \) is non-increasing.
You might note that in the previous proof of converse, we do not make use of lower bounds on error probability such as Fano’s inequality. This is because that in our formulation of the lossy source coding problem, the reconstruction criterion is laid on the expected distortion.

Instead of the criterion $\limsup_{N \to \infty} D^{(N)} \leq D$ where $D^{(N)} \triangleq \mathbb{E} \left[ d \left( S^N, \hat{S}^N \right) \right]$, we could use a stronger criterion as follows:

$$P_{e}^{(N,\delta)} \triangleq \mathbb{P} \left\{ d \left( S^N, \hat{S}^N \right) > D + \delta \right\}, \quad \delta > 0 \quad \text{(Probability of Error)}$$

$$\lim_{N \to \infty} P_{e}^{(N,\delta)} = 0, \quad \forall \delta > 0 \quad \text{(Reconstruction Criterion)}$$

Under this stronger criterion, we can then give a new operational definition of the rate distortion function.

It turns out Theorem 1 remains the same! (converse is implied by our converse)
1. Lossy Source Coding Theorem for Memoryless Sources
   - Lossy Source Coding Theorem
   - Rate Distortion Function

2. Proof of the Coding Theorem
   - Converse Proof
   - Achievability
Idea of Constructing Good Source Code

Key in source coding:
1. Find a good set of representatives (quantization codewords).
2. For each source sequence, determine which codeword to be used.

Main tools we use so far in developing achievability of coding theorems:
1. Random coding: construct the codebook randomly and show that at least one realization can achieve the desired target performance.
2. Typicality: help give bounds in performance analysis.

In the following, we prove the achievability part of Theorem 1 by
1. Random coding – show existence of good quantization codebook.
2. Typicality encoding – determine which codeword to be used.
Lossy Source Coding Theorem for Memoryless Sources

Proof of the Coding Theorem

Converse Proof

Achievability

\[ S^N \]

Random Coding

Typicality

Encoding

\[ \hat{S}^N \]
Proof Program

1 Random Codebook Generation:
Generate a random ensemble of quantization codebooks, each of which contains $2^K$ codewords.

2 Analysis of Expected Distortion:
Goal: Show that $\limsup_{N \to \infty} \mathbb{E}_{C, S^N} \left[ d \left( S^N, \hat{S}^N \right) \right] \leq D$, and conclude that there must exist a codebook $C$ such that the expected distortion satisfies $\limsup_{N \to \infty} \mathbb{E}_{S^N} \left[ d \left( S^N, \hat{S}^N \right) \right] \leq D$.

Note that for a source sequence $s^N$, the optimal encoder chooses an index $w \in [1 : 2^K]$, that is, a codeword $\hat{s}^N(w)$ in the codebook, so that $d \left( s^N, \hat{s}^N(w) \right)$ is minimized.

However, similar to ML decoding in channel coding, such optimal encoder is hard to analyze. To simplify analysis, we shall introduce a suboptimal encoder based on typicality.
Random Codebook Generation

Fix the conditional p.m.f. that attains $R \left( \frac{D}{1+\varepsilon} \right)$:

$$q_{\hat{S}|S} = \arg \min_{p_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \leq \frac{D}{1+\varepsilon}} I(S;\hat{S}) \tag{2}$$

Based on the chosen $q_{\hat{S}|S}$ and the source distribution $p_S$, calculate $p_{\hat{S}}$, the marginal distribution of the reconstruction $\hat{S}$.

Generate $2^K$ codewords $\{\hat{s}^N(w) \mid w = 1, 2, \ldots, 2^K\}$, i.i.d. according to $p(\hat{s}^N) = \prod_{i=1}^{N} p_{\hat{S}}(\hat{s}_i)$.

In other words, if we think of the quantization codebook as a $2^K \times N$ matrix $C$, the elements of $C$ will be i.i.d. distributed according to $p_{\hat{S}}$.

Remark: observe the resemblance with the channel coding achievability.
Encoding and Decoding

**Encoding**: unlike channel coding, the encoding process in source coding problem is usually much involved.

We use **typicality encoding**: (resembling **typicality decoding** in channel coding)

- Given a source sequence $s^N$, find an index $w \in [1 : 2^K]$ such that

  $$(s^N, \hat{s}^N(w)) \in \mathcal{T}_\varepsilon^{(N)} \left(p_{S,\hat{S}}\right).$$

  Recall the joint distribution $p_{S,\hat{S}} = p_{S} \times q_{\hat{S}|S}$ as defined in (2).

- If there is no or more than one such index, randomly pick one $w \in [1 : 2^K]$.

- Send out the bit sequence that represent the chosen $w$.

**Decoding**: Upon receiving the bit sequence representing $w$, generate the reconstructed $\hat{s}^N(w)$ by looking up the quantization codebook.
Analysis of Expected Distortion

Why typicality encoder? Typical average lemma (Lemma 2, Lecture 04):

For any nonnegative function $g(x)$ on $\mathcal{X}$, if $x^n \in T_\epsilon^{(n)}(X)$, then

$$(1 - \epsilon) \mathbb{E}[g(X)] \leq \frac{1}{n} \sum_{i=1}^{n} g(x_i) \leq (1 + \epsilon) \mathbb{E}[g(X)].$$

In analyzing $\mathbb{E}_{C,S^N} \left[d \left( S^N, \hat{S}^N \right) \right]$, we can then distinguish into two cases:

$\mathcal{E} \triangleq \left\{ \left( S^N, \hat{S}^N \right) \notin T_\epsilon^{(N)} \right\}$ and $\mathcal{E}^c \triangleq \left\{ \left( S^N, \hat{S}^N \right) \in T_\epsilon^{(N)} \right\}$:

$$\mathbb{P} \{ \mathcal{E} \} \mathbb{E}_{C,S^N} \left[ d \left( S^N, \hat{S}^N \right) \middle\vert \mathcal{E} \right] + \mathbb{P} \{ \mathcal{E}^c \} \mathbb{E}_{C,S^N} \left[ d \left( S^N, \hat{S}^N \right) \middle\vert \mathcal{E}^c \right]$$

$$\leq \mathbb{P} \{ \mathcal{E} \} \max_{s,\hat{s}} d(s,\hat{s}) + \mathbb{P} \{ \mathcal{E}^c \} (1 + \epsilon) \frac{D}{1 + \epsilon} \leq \mathbb{P} \{ \mathcal{E} \} \max_{s,\hat{s}} d(s,\hat{s}) + D.$$

Hence, as long as $\mathbb{P} \{ \mathcal{E} \}$ vanishes as $N \to \infty$, we are done.
Analysis of Expected Distortion → Analysis of $\mathbb{P}\{E\}$

With typicality encoding, analysis of expected distortion is made easy: just need to control $\mathbb{P}\{E\}$, where $E \triangleq \left\{ \left( S^N, \hat{S}^N \right) \notin \mathcal{T}_{\varepsilon}^{(N)} \right\}$.

Let us look at event $E$: it is the event that the reconstructed $\hat{S}^N$ is not jointly typical with $S^N$, which can only happen when none of the quantization codewords in the codebook is jointly typical with $S^N$.

Hence, $E \subseteq \cap_{w=1}^{2^K} A_w^c$, where $A_w \triangleq \left\{ \left( S^N, \hat{S}^N(w) \right) \in \mathcal{T}_{\varepsilon}^{(N)} \right\}$.

$$\implies \mathbb{P}\{E\} \leq \mathbb{P}\left\{ \cap_{w=1}^{2^K} A_w^c \right\}.$$

Unfortunately, the events $\left\{ A_w^c \mid w = 1, \ldots, 2^K \right\}$ may not be mutually independent, because they all involve a common random sequence $S^N$.

However, for fixed $S^N$, the events $A_w^c \left( S^N \right) \triangleq \left\{ \left( S^N, \hat{S}^N(w) \right) \notin \mathcal{T}_{\varepsilon}^{(N)} \right\}$, $w = 1, \ldots, 2^K$, are indeed mutually independent!
Motivated by the above observation, we give an alternative upper bound:

\[
\mathbb{P}\{\mathcal{E}\} \leq \sum_{s^N \in \mathcal{S}^N} p(s^N) \mathbb{P}\left\{\bigcap_{w=1}^{2^K} A_w^c(s^N)\right\}
\]

\[
= \sum_{s^N \in \mathcal{S}^N} p(s^N) \prod_{w=1}^{2^K} \mathbb{P}\{A_w^c(s^N)\}
\]

\[
= \sum_{s^N \in \mathcal{S}^N} p(s^N) \prod_{w=1}^{2^K} (1 - \mathbb{P}\{A_w(s^N)\})
\]

**Question:** Is there a way to lower bound

\[
\mathbb{P}\{A_w(s^N)\} \triangleq \mathbb{P}\left\{(s^N, \hat{\mathcal{S}}^N(w)) \in \mathcal{T}_{\varepsilon'}^{(N)}(p_{\mathcal{S},\hat{\mathcal{S}}})\right\} ?
\]

**Yes** – As long as \(s^N \in \mathcal{T}_{\varepsilon'}^{(N)}(p_{\mathcal{S}})\) for some \(\varepsilon' < \varepsilon\), Lemma 1 (next slide) guarantees that \(\mathbb{P}\{A_w(s^N)\} \geq 2^{-N(I(S;\hat{S}) + \delta(\varepsilon))}\) for sufficiently large \(N\), where \(\lim_{\varepsilon \to 0} \delta(\varepsilon) = 0\).
Joint Typicality Lemma

The following lemma formally states the bounds.

(Proof is omitted – see Section 2.5 of ElGamal & Kim[6])

**Lemma 1 (Joint Typicality Lemma)**

Consider a joint p.m.f. $p_{X,Y} = p_X \cdot p_{Y|X} = p_Y \cdot p_{X|Y}$. Then, there exist $\delta(\varepsilon) > 0$ with $\lim_{\varepsilon \to 0} \delta(\varepsilon) = 0$ such that:

1. For an arbitrary sequence $x^n$ and random $Y^n \sim \prod_{i=1}^{n} p_Y(y_i)$,

$$
\mathbb{P} \left\{ (x^n, Y^n) \in \mathcal{T}_{\varepsilon}^{(n)} (p_X, Y) \right\} \leq 2^{-n(I(X;Y) - \delta(\varepsilon))}.
$$

2. For an $\varepsilon'$-typical sequence $x^n \in \mathcal{T}_{\varepsilon'}^{(n)} (p_X)$ with $\varepsilon' < \varepsilon$, and random $Y^n \sim \prod_{i=1}^{n} p_Y(y_i)$, for sufficiently large $n$,

$$
\mathbb{P} \left\{ (x^n, Y^n) \in \mathcal{T}_{\varepsilon}^{(n)} (p_X, Y) \right\} \geq 2^{-n(I(X;Y) + \delta(\varepsilon))}.
$$
Finalizing the Proof

Revoking Lemma 1, the additional condition that $s^N \in \mathcal{T}_{\varepsilon'}^{(N)}(p_S)$ for some $\varepsilon' < \varepsilon$ motivates us to split the upper bound on $\mathbb{P}\{\mathcal{E}\}$ as follows:

$$
\mathbb{P}\{\mathcal{E}\} \leq \sum_{s^N \in \mathcal{S}^N} p(s^N) \prod_{w=1}^{2^K} (1 - \mathbb{P}\{\mathcal{A}_w(s^N)\})
$$

$$
\leq \sum_{s^N \notin \mathcal{T}_{\varepsilon'}^{(N)}(p_S)} p(s^N) + \sum_{s^N \in \mathcal{T}_{\varepsilon'}^{(N)}(p_S)} p(s^N) \prod_{w=1}^{2^K} (1 - \mathbb{P}\{\mathcal{A}_w(s^N)\})
$$

$$
\leq \mathbb{P}\left\{S^N \notin \mathcal{T}_{\varepsilon'}^{(N)}(p_S)\right\} + \sum_{s^N \in \mathcal{T}_{\varepsilon'}^{(N)}(p_S)} p(s^N) \left(1 - 2^{-N(I(S;\hat{S})+\delta(\varepsilon))}\right)^{2^K}
$$

$$
\leq \mathbb{P}\left\{S^N \notin \mathcal{T}_{\varepsilon'}^{(N)}(p_S)\right\} + \left(1 - 2^{-N(I(S;\hat{S})+\delta(\varepsilon))}\right)^{2^K}
$$

$$
\leq \mathbb{P}\left\{S^N \notin \mathcal{T}_{\varepsilon'}^{(N)}(p_S)\right\} + \exp\left(-2^K \times 2^{-N(I(S;\hat{S})+\delta(\varepsilon))}\right).
$$

The last step is due to $(1 - x)^r \leq e^{-rx}$ for $x \in [0, 1]$ and $r \geq 0$. 
We obtain a nice upper bound

\[ \mathbb{P} \{ \mathcal{E} \} \leq \mathbb{P} \left\{ S^N \notin \mathcal{T}^{(N)}_{\varepsilon} (p_S) \right\} + \exp \left( -2^K \times 2^{-N(I(S; \hat{S}) + \delta(\varepsilon))} \right). \]

The first term vanishes as \( N \to \infty \) due to AEP. The second term vanishes as \( N \to \infty \) if

\[ R > I(S; \hat{S}) + \delta(\varepsilon) = R \left( \frac{D}{1+\varepsilon} \right) + \delta(\varepsilon). \]

Hence, for any \( R > R \left( \frac{D}{1+\varepsilon} \right) + \delta(\varepsilon) \) can achieve average distortion \( \leq D \).

Finally, due to the continuity of rate-distortion function, we take \( \varepsilon \to 0 \) and complete the proof.
Summary
For a memoryless source \( \{S_i \mid i \in \mathbb{N}\} \),

\[
R(D) = \min_{p_{\mathcal{S}|\mathcal{S}}: \mathbb{E}[d(S,\hat{S})] \leq D} I(S;\hat{S}).
\]

Properties of a rate distortion function \( R(D) \):

1. Nonnegative; Non-increasing in \( D \).
2. Convex in \( D \); Continuous in \( D \).
3. \( R(D_{\min}) \leq H(S) \).
4. \( R(D) = 0 \) if \( D \geq D_{\max} \).

Bernoulli Source with Hamming Distortion:

\[
R(D) = \max \left\{ H_b(p) - H_b(D), 0 \right\}
\]

Gaussian Source with Squared Error Distortion:

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