# Slide Set IV: Lossy Source Coding

- Lossy Source Code
- Rate vs. Distortion for Memoryless Sources
- Rate vs. Distortion for Sources with Memory
- Universal Lossy Source Coding
- Shannon Lower Bound



# Lossy Source Code

Assume $\mathcal{X}$, $\mathcal{Y}$ are, respectively, the finite source and reconstruction alphabet.

Assume a given single-letter distortion measure $\rho : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ which induces a block distortion:

$$\rho(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i).$$



# Lossy Source Code (cont.)

**Definition 1.** *A fixed-rate $n$-block code is a pair $(\mathcal{C}_n, \phi_n)$ where $\mathcal{C}_n \subseteq \mathcal{Y}^n$ is the code-book and $\phi_n : \mathcal{X}^n \to \mathcal{C}_n$. The rate of the block-code is given by $\frac{1}{n} \log |\mathcal{C}_n|$. $Y^n = Y^n(X^n) = \phi_n(X^n)$ will denote the reconstruction sequence when the $n$-block code is applied to the source sequence $X^n$. The expected distortion is $E\rho(X^n, Y^n)$.*

**Definition 2.** *A variable-rate code for $n$-blocks is a triple $(\mathcal{C}_n, \phi_n, l_n)$ with $\mathcal{C}_n$ and $\phi_n$ as in Definition 1. The code operates by mapping a source $n$-tuple $X^n$ into $\mathcal{C}_n$ via $\phi_n$, and then encoding the corresponding member of the code-book (denoted $Y^n$ in Definition 1) using a uniquely decodable binary code. Letting $l_n(X^n)$ denote the associated length function, the rate of the code is defined by $E\frac{1}{n}l_n(X^n)$. The expected distortion is $E\rho(X^n, Y^n)$.*



# Rate Distortion Function

The Rate Distortion function associated with the random variable $X$ is

$$R(D) = \min_{E\rho(X,Y) \le D} I(X;Y)$$

## Rate Distortion Theorem

**Theorem 1.** *Let $\mathbf{X} = \{X_i\}$ be an i.i.d. source $X_i \sim X$.*

*[Direct:] For any $\epsilon > 0$ and $n$ sufficiently large there exists an $n$-block code with rate $\leq R(D) + \epsilon$ and distortion $\leq D + \epsilon$.*

*[Converse:] Any code-book $\mathcal{C}_n$ with rate $\leq R(D)$ has distortion $\geq D$.*



## Strongly Typical Sequences and Rate Distortion

For $x^n \in \mathcal{X}^n$ let $N(a|x^n) = |\{1 \leq i \leq n : x_i = a\}|$. The *strongly typical set* $T_X^\epsilon$ is the set of all sequences $x^n \in \mathcal{X}^n$ satisfying

1. $\left|\frac{1}{n}N(a|x^n) - P_X(a)\right| \leq \frac{\epsilon}{|\mathcal{X}|}$ for all $a \in \mathcal{X}$ with $P_X(a) > 0$ and
2. $N(a|x^n) = 0$ for all $a \in \mathcal{X}$ with $P_X(a) = 0$.

$T_{X,Y}^\epsilon$ is defined analogously.

**Lemma 1.** *Let $\{Y_i\}$ be i.i.d.$\sim Y$. For $x^n \in T_X^\epsilon$*

$$2^{-n(I(X;Y)+\epsilon_1)} \leq \Pr\left((x^n, Y^n) \in T_{X,Y}^\epsilon\right) \leq 2^{-n(I(X;Y)-\epsilon_1)},$$

*where $\epsilon_1 \to 0$ as $\epsilon \to 0$ and $n \to \infty$.*



## Recalling Achievability Idea in Rate Distortion Theorem

- Let $(X, Y)$ achieve the rate distortion function.
- Generate codewords in code-book independently with components i.i.d. $\sim Y$
- By Lemma 1, for typical source realizations $x^n$, probability that $x^n$ be jointly typical with, say, the first codeword is $\approx 2^{-nI(X;Y)}$.
- So, if we generate $2^{n(I(X;Y)+\epsilon)}$ codewords at least one of them will be jointly typical with the source with high probability
- Distortion between this codeword and source will be $\approx E\rho(X, Y)$

Recall, e.g., [Berger71], [Gallager68], [CT91, Sec. 13.6].



## Remarkably Simple Converse

Fix a code-book with rate $R$ and distortion $\leq D$.

$$\begin{aligned} R &\geq \frac{1}{n}H(Y^n) \\ &\geq \frac{1}{n}I(X^n; Y^n) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} I(X_i; Y_i) \\ &\geq \sum_{i=1}^{n} R(E\rho(X_i, Y_i)) \\ &\geq \frac{1}{n}R\left(\sum_{i=1}^{n} E\rho(X_i, Y_i)\right) \\ &\geq R(D). \square \end{aligned}$$

## Rate Distortion Function for Sources with Memory

The Rate Distortion function associated with $X^n$ is

$$R(X^n, D) = \min_{E\rho(X^n,Y^n)\leq D} \frac{1}{n} I(X^n; Y^n).$$

The Rate Distortion function associated with the stationary source $\mathbf{X}$ is

$$R(\mathbf{X}, D) = \lim_{n\to\infty} R(X^n, D) = \inf_n R(X^n, D).$$

The existence of the limit, as well as the second equality, will be established in the HW exercise. You will also show that

$$R(\mathbf{X}, D) = \inf\{\overline{I}(\mathbf{X}; \mathbf{Y}) : E\rho(X_0, Y_0) \leq D\},$$

the infimum taken over jointly stationary processes consistent with the distribution of $\mathbf{X}$ and $\overline{I}(\mathbf{X}; \mathbf{Y}) \stackrel{\triangle}{=} \lim \frac{1}{n} I(X^n; Y^n)$ denotes the mutual information rate.



## Rate Distortion Theorem for Stationary Ergodic Sources

**Theorem 2.** *Let $\mathbf{X}$ be stationary ergodic.*

*[Direct:] For any $\epsilon > 0$ and $n$ sufficiently large there exists a code-book for $n$-blocks with rate $\leq R(\mathbf{X}, D) + \epsilon$ and distortion $\leq D + \epsilon$.*

*[Converse:] Any code-book $\mathcal{C}_n$ with rate $\leq R(\mathbf{X}, D)$ has distortion $\geq D$.*



## Achievability Idea

Achievability of $R(X_1, D)$: Same as i.i.d. case.

Achievability of $R(X^n, D)$:

- Similar to achievability of $R(X_1, D)$ by considering super alphabet $\mathcal{X}^n$
- Only problem is: "super process" may not be ergodic.
- But can be separated into, at most, $n$ "ergodic modes".
- Letting $P_n(y^n|x^n)$ achieve $R(X^n, D)$ and $i$ represent the mode

$$nR(X^n, D) = I(X^n; Y^n) \geq \frac{1}{n} \sum_{i=1}^{n} I_i(X^n; Y^n) \geq \sum_{i=1}^{n} R(X^n, D_i|i),$$

(notation $I_i(X^n; Y^n)$ and $R(X^n, D_i|i)$ to be explained in class) where $D_i = E[\rho(X^n, Y^n)|i]$ so $D = \frac{1}{n}\sum_{i=1}^{n} D_i$.



- Thus, for large $L$ can find codes of block length $nL$ attaining distortion essentially $D_i$ and rate $R(X^n, D_i|i)$ on the $i$-th mode. Average distortion is then $D = \frac{1}{n}\sum_{i=1}^{n} D_i$ and average rate is $\frac{1}{n}\sum_{i=1}^{n} R(X^n, D_i|i) \leq R(X^n, D)$.

Details can be found in [Berger71], [Gallager68].

## Universal Lossy Source Coding

David L. Neuhoff, Paul C. Shields; Simplistic Universal Coding, IEEE Trans. Info. Theory, vol. IT-44, pp. 778 - 781, March 1998.

[Ziv72] J. Ziv. Coding of sources with unknown statistics - part II: Distortion relative to a fidelity criterion. IEEE Trans. Inform. Theory, 18:389394, May 1972.

[YK96] E. H. Yang and J. Kieffer. Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm. IEEE Trans. Inform. Theory, 42:239245, 1996.

We follow [YK96]



## Universal Lossy Source Coding (cont.)

**A Simple Fixed-Rate Universal Lossy Coding Scheme:**

Let the code-book be given by

$$B_n = \{y^n \in \hat{\mathcal{X}}^n : l_{LZ}(y^n) \leq nR\}. \tag{1}$$

The encoding is done by giving the index of $y^n \in B_n$ that minimizes $\rho(x^n, y^n)$.

Note universality of code-book w.r.t. source distribution and also w.r.t. distortion measure !



## Universal Lossy Source Coding (cont.)

**A Simple Fixed-Distortion Universal Lossy Coding Scheme:**

Code-book: List the elements of $\hat{\mathcal{X}}^n$ in nondecreasing order of LZ codeword length.

Encoding: Encode source $n$-block into the LZ codeword of the first word in code-book for which the resulting distortion is $\leq D$.



## Universality Result

**Theorem 3. [YK96]** *For $\mathbf{X}$ stationary and ergodic*

$$E\rho(X^n, B_n) \to D(\mathbf{X}, R) \quad \textit{as} \quad n \to \infty.$$

## Proof of Theorem 3

Our proof will use the following lemma:

**Lemma 2.** *Let $m$ be a positive integer and $\sigma : \hat{\mathcal{X}}^m \to \{1, 2, \ldots\}$ be a length function. Then there exists a sequence of positive real numbers $\{\delta_n\}$ tending to zero such that for all $n$ and $y^n \in \hat{\mathcal{X}}^n$*

$$l_{LZ}(y^n) \le \min_{1 \le j \le m} \left\{ \sum_{i \equiv j (\text{mod } m), 1 \le i \le n-m+1} \sigma(y_i^{i+m-1}) \right\} + n\delta_n.$$

*Proof of Theorem 3:* See white-board.



## The Shannon Lower Bound

The Shannon lower bound we now derive will be key in our treatment of the compression-based approach to denoising.

Let $A$ be an alphabet where addition and subtraction of elements are well-defined.

Let $\rho$ be a difference distortion measure $\rho(x, y) = \rho(y - x)$.

The maximum-entropy function (associated with $\rho$) is defined by

$$\phi(D) = \max\{H(X) : E\rho(X) \le D\}.$$

The distribution attaining the maximum is of the form

$$P_{\rho,D}(x) = \frac{e^{-\beta\rho(x)}}{\sum_{x'} e^{-\beta\rho(x')}},$$

where $\beta$ is tuned so that the constraint is met with equality (when possible).



## The Shannon Lower Bound (cont.)

Let $R(X, D)$ (for the difference distortion measure $\rho$) be achieved by $(X, Y)$. Then

$$\begin{aligned} R(X, D) &= I(X; Y) \\ &= H(X) - H(X|Y) \\ &= H(X) - H(X - Y|Y) \\ &\ge H(X) - H(X - Y) \\ &\ge H(X) - \phi(E\rho(X - Y)) \\ &\ge H(X) - \phi(D). \end{aligned}$$

Thus we obtain

$$R(X, D) \ge H(X) - \phi(D)$$

with equality (assuming $\phi$ strictly increasing at $D$) if and only if $X$ can be represented as $X = Y \oplus N$ for some $Y$ and $N \sim P_{\rho,D}$ independent of $Y$.



## The Shannon Lower Bound (cont.)

Similarly can show that for each $n$

$$R(X^n, D) \ge \frac{1}{n} H(X^n) - \phi(D)$$

with equality if and only if $\exists Y^n, N^n$ independent random $n$-tuples with $N_i$ i.i.d. $\sim P_{\rho,D}$ such that $X_i = Y_i \oplus N_i$. This also leads to the following:

**Theorem 4.** *For any stationary ergodic process*

$$R(\mathbf{X}, D) \ge \overline{H}(\mathbf{X}) - \phi(D),$$

*with equality whenever $\mathbf{X}$ can be represented as $X_i = Y_i + N_i$ for some process $\mathbf{Y}$ independent of $\mathbf{N}$, with $N_i$ i.i.d.$\sim P_{\rho,D}$.*

## The Shannon Lower Bound (cont.)

Deceivingly many sources turn out to satisfy the Shannon lower bound with equality at least for small distortions:

- Various Hidden Markov processes
- Gaussian processes
- Certain Markov processes
- Certain Auto-regressive sources

Robert M. Gray; Information rates of autoregressive processes, IEEE Trans. Info. Theory, vol. IT-16, pp. 412 - 421, July 1970.

Robert M. Gray; Rate distortion functions for finite-state finite-alphabet Markov sources, IEEE Trans. Info. Theory, vol. IT-17, pp. 127 - 134, March 1971.