

Topic 2:

Mathematical Models of Compression

Assumptions:

- Signal is discrete time or space (e.g., already sampled).
- Do not separate out signal decompositions, i.e., assume either done already or to be done as part of the code.
- Consider a code structure that maps blocks or vectors of input data into possibly variable length binary strings.

Will later consider other code structures, which can be recursive and have memory. The fixed block structures can, however, model such codes by allowing local behavior that is recursive.

Input Signal

Assume discrete-time random process input signal: $\{X_n\}$ e.g., each X_n might be a real number, a block of pixels, 50 speech samples, a single binary symbol or N binary symbols. We will assume that each X_n is a k -dimensional vector with alphabet $A \subset \Re^k$.

Implies know distributions P_{X^k} for all vectors $X^k = (X_0, X_1, \dots, X_{k-1})$, $k = 1, 2, \dots$

Use superscript notation when dimension not clear from context.

Usually assume some form of stationarity (strict, asymptotic, etc.)

Arguing over “stationarity” of real data is a red herring, the theory can handle very general processes (with more complicated “universal” or “adaptive” codes).

For simplicity, assume strictly stationary. So all $X_n^k = (X_{nk}, X_{nk+1}, \dots, X_{(n+1)k-1})$ are distributed as a generic random vector X^k , say with probability distribution P_{X^k} .

Drop the superscript k if it is clear from context.

Encoder

An *encoder* (or *source encoder*) α is a mapping of the input vectors into a collection \mathcal{W} of binary sequences.

$\mathcal{W} \subset \{0, 1\}^*$, the space of all possible binary sequences.

\mathcal{W} is called the *channel codebook* and its members are referred to as *channel codewords*. It is the set of binary sequences that will be stored or transmitted and used by the decoder to reconstruct the source as well as it can.

Thus an encoder is a mapping $\alpha : A \rightarrow \mathcal{W}$. Applying the encoder to an input X yields a codeword $i = \alpha(X)$.

Given an $i \in \{0, 1\}^*$, define

$$l(i) = \text{length of binary vector } i$$

e.g.,

$$l(0) = 1, l(101) = 3, l(1011000) = 7$$

Define the *instantaneous rate* of a binary vector i by

$$r(i) = \frac{i}{k}$$

in bits per input symbol.

The *average rate* or *average codeword length* of the encoder applied to the source is defined by

$$R(\alpha, \mathcal{W}) = E[r(\alpha(X))].$$

An encoder is said to be *fixed length* or *fixed rate* if all channel codewords have the same length, i.e., if $l(i) = Rk$ for all $i \in \mathcal{W}$.

The property of fixed or variable rate codes can have important implications in practice.

- Variable rate codes can cost more as they may require data buffering if the encoded data is to be transmitted over a fixed rate communication channel.
- Harder to synchronize variable-rate codes. Single errors in decoding or lost or added bits on the channel can have catastrophic effects.
- Buffers can overflow (causing data loss) or underflow (causing wasted time or bandwidth).

But variable rate codes can provide superior rate/distortion tradeoffs.

E.g., in image compression can use more bits for edges, fewer for flat areas. In voice compression, more bits for plosives, fewer for vowels.

Define a source *decoder* $\beta : \mathcal{W} \rightarrow \hat{A}$
usually $\hat{A} = A$

Decoder is a table lookup.

Define the *reproduction codebook*

$$\mathcal{C} \equiv \{\beta(i); i \in \mathcal{W}\}$$

members of \mathcal{C} called *reproduction codewords* or
templates.

Often convenient to reindex codebook using
integers as

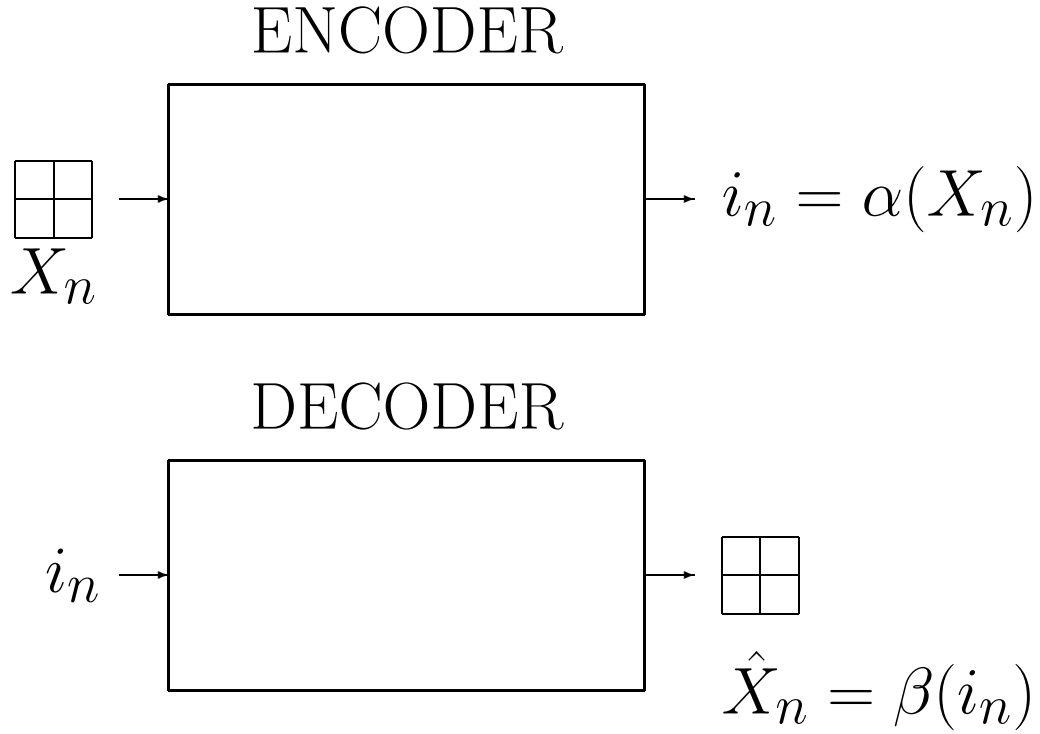
$$\mathcal{C} \equiv \{\beta_l; l = 0, 1, \dots, M - 1\}$$

where M is the number of reproduction
codewords:

$$M = ||\mathcal{W}||$$

A *source code* or *compression code* for the
source $\{X_n\}$ consists of a triple $(\alpha, \mathcal{W}, \beta)$ of
encoder, channel codebook, and decoder.

Will see other, equivalent, representations.



$$A \xrightarrow{\alpha} \mathcal{W} \xrightarrow{\beta} \mathcal{C} \quad (1)$$

or, equivalently,

$$X \xrightarrow{\alpha} i = \alpha(X) \xrightarrow{\beta} \hat{X} = \beta(\alpha(X)). \quad (2)$$

General block memoryless source code

Later consider codes with memory, but general block might operate in local nonmemoryless fashion.

A source code is *invertible* or *noiseless* or *lossless* if

$$\beta(\alpha(x)) = x$$

(at least with probability 1). Alternatively, the code is invertible if

$$\beta = \alpha^{-1}$$

A code is *lossy* if it is not lossless. In this case a measure of distortion d between input vector and reconstruction is required to measure the seriousness of the loss.

Quality and Cost

Distortion

Distortion measure $d(x, \hat{x})$ measures the distortion or loss resulting if an original input x is reproduced as \hat{x} .

Mathematically: A distortion measure satisfies

$$d(x, \hat{x}) \geq 0$$

To be useful, d should be

- easy to compute
- tractable
- meaningful for perception or application.

No single distortion measure accomplishes all of these goals, although the ubiquitous squared error distortion defined by

$$d(x, y) = ||x - y||^2 = \sum_{l=0}^{k-1} |x_l - y_l|^2$$

where $x = (x_0, x_1, \dots, x_{k-1})^t$,
 $y = (y_0, y_1, \dots, y_{k-1})^t$ accomplishes the first two goals and occasionally correlates with the third.

Weighted or transform/weighted versions are used for perceptual coding.

$$d(X, \hat{X}) = (X - \hat{X})^* B_X (X - \hat{X}),$$

B_X positive definite.

most common $B_X = I$,

$$d(X, \hat{X}) = \|X - \hat{X}\|^2 \text{ (MSE)}$$

Other distortion measures of interest:

Hamming distortion:

$$d_H(x, \hat{x}) = 1 - \delta_{x-\hat{x}} = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{otherwise} \end{cases}$$

Hamming weight or average Hamming distance:

For vectors $x = (x_0, x_1, \dots, x_{k-1})$

$$d(x, \hat{x}) = \sum_{l=0}^{k-1} d_H(x_l, \hat{x}_l)$$

l_p norms

$$d(x, y) = \|x - y\|^2 = \left| \sum_{l=0}^{k-1} |x_l - y_l|^p \right|^{1/p}$$

Common assumption:

d is an *additive* distortion measure: have a distortion measure $d_1(a, b)$ defined on “scalars”. For vector $x^k = (x_0, x_1, \dots, x_{k-1})$ define

$$d_k(x^k, \hat{x}^k) = \sum_{l=0}^{k-1} d_1(x_l, \hat{x}_l)$$

Often normalize.

If $d(x, y) = 0$ iff $x = y$, then zero distortion coding is equivalent to lossless coding. We make this assumption for convenience (if source discrete).

If apply $(\alpha, \mathcal{W}, \beta)$ to x , distortion is $d(x, \beta(\alpha(x)))$, does not depend explicitly on \mathcal{W} .

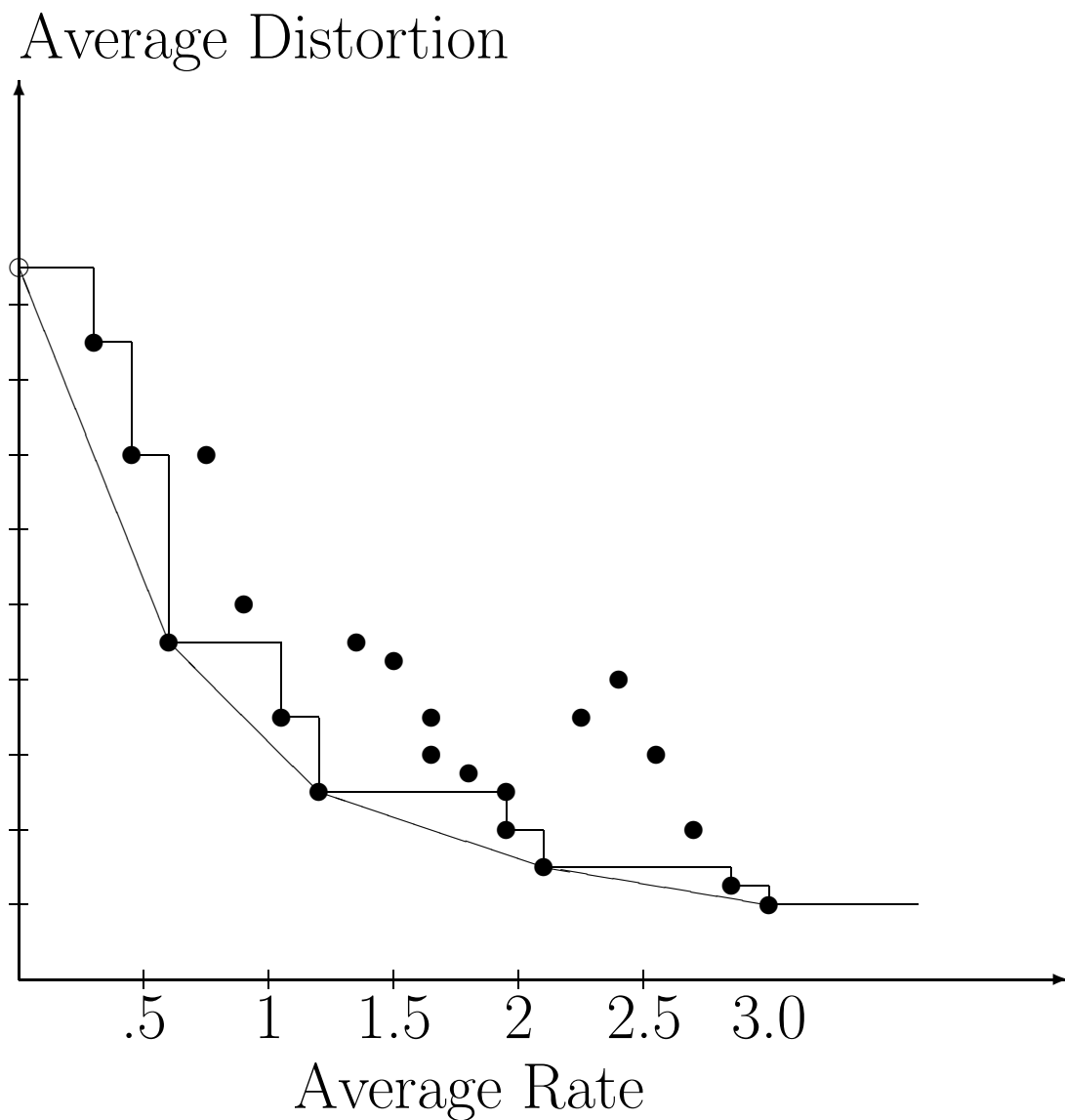
Performance of a compression system measured by expected values of the distortion and rate.

$$D(\alpha, \mathcal{W}, \beta) = D(\alpha, \beta) = E[d(X, \beta(\alpha(X)))]$$

$$R(\alpha, \mathcal{W}) = E(r(X)) = \frac{1}{k} E(l(\alpha(X)))$$

The unit of rate is bits per input symbol.
Occasionally drop normalization and unit becomes bits per input vector.

Every code yields point in distortion/rate plane:
 (R, D) .



$D(\alpha, \beta)$ and $R(\alpha, \mathcal{W})$ measure costs, want to
 minimize both \Leftrightarrow Tradeoff

Interested in undominated points in D-R plane:
For given rate (distortion) want smallest possible
distortion (rate) \Leftrightarrow optimal codes

Optimization problem:

- Given R , what is smallest possible D ?
- Given D , what is smallest possible R ?
- Lagrangian formulation: What is smallest possible $D + \lambda R$?

Lossless codes correspond to $(0, R)$ points.

An extreme point of above general optimization.
What is smallest R giving a lossless code?

Another extreme point that will prove of
interest: best $(D, 0)$ point.

What is the best possible 0 rate code?

(Useless in practice, but provides key ideas for
general case in very simple setting.)

Make these ideas more precise and carefully describe the various optimization problems.

- Rate-distortion approach: Constrain $R(\alpha, \mathcal{W}) \leq R$. Then optimal code $(\alpha, \mathcal{W}, \beta)$ minimizes $D(\alpha, \beta)$ over all allowed codes.

operational rate-distortion function

$$\hat{R}(D) = \inf_{\alpha, \mathcal{W}, \beta: D(\alpha, \beta) \leq D} R(\alpha, \mathcal{W}).$$

- Distortion-rate approach: Constrain $D(\alpha, \beta) \leq D$. Then optimal code $(\alpha, \mathcal{W}, \beta)$ minimizes $R(\alpha, \mathcal{W})$ over all allowed codes.

operational distortion-rate function

$$\hat{D}(R) = \inf_{\alpha, \mathcal{W}, \beta: R(\alpha, \mathcal{W}) \leq R} D(\alpha, \beta).$$

- Lagrangian approach: Fix Lagrangian multiplier λ .

Optimal code $(\alpha, \mathcal{W}, \beta)$ minimizes

$$J_\lambda(\alpha, \mathcal{W}, \beta) = D(\alpha, \beta) + \lambda R(\alpha, \mathcal{W})$$

over all allowed codes.

operational Lagrangian distortion function

$$\begin{aligned} \hat{J}_\lambda &= \inf_{\alpha, \mathcal{W}, \beta} E\rho_\lambda(X, \beta(\alpha(X))) = \\ &\inf_{\alpha, \mathcal{W}, \beta} [D(\alpha, \beta) + \lambda R(\alpha, \mathcal{W})]. \end{aligned}$$

Each viewpoint has its uses.

- DRF good for proving coding theorems
- RDF useful for computing and bounding Shannon limits for known sources
- Lagrangian good for formulating code design descent algorithms

First two problems are duals, all three are equivalent. E.g., Lagrangian approach yields R-D for some D or D-R for some R.

Lagrangian approach effectively unconstrained minimization of modified distortion

$J_\lambda = E\rho_\lambda(X, \alpha(X))$ where

$$\rho_\lambda(X, \alpha(X)) = d(X, \beta(\alpha(X))) + \lambda l(\mathcal{W}(\alpha(X)))$$

As $\lambda \rightarrow 0$ distortion $\rightarrow 0$, rate $\rightarrow \hat{R}(0) = ??$

As $\lambda \rightarrow \infty$ rate $\rightarrow 0$ distortion $\rightarrow \hat{D}(0) = ??$

These extreme points to be studied

Easy: $\hat{D}(R)$ and $\hat{R}(D)$ are monotonically nonincreasing in their arguments.

Note: usually wish to optimize over constrained subset of computationally reasonable codes, or implementable codes.

Examples: Fixed rate codes, tree-structured codes, product codes

Introduce these structures later, but mention fixed rate codes now.

Fixed Rate Codes

Important special case: Fixed rate (length) codes.

Require all words in \mathcal{W} = range space of \mathcal{W} to have equal length.

Then $r(X)$ constant and problem simplified.

Eases buffering requirements when use fixed-rate transmission media

Eases effects of channel errors

In fixed rate case, minimizing modified distortion equivalent to minimizing ordinary distortion.

Lagrangian and distortion-rate identical for fixed R .

Shannon Theory

$X = X^k$, k allowed to vary.

Can define optima for increasing dimensions:

$\hat{D}_k(R)$, $\hat{R}_k(D)$, and $\hat{J}_{\lambda,k}$

Quantities are subadditive & can define asymptotic optimal performance

$$\bar{D}(R) = \inf_k \hat{D}_k(R) = \lim_{k \rightarrow \infty} \hat{D}_k(R) \quad (3)$$

$$\bar{R}(D) = \inf_k \hat{R}_k(D) = \lim_{k \rightarrow \infty} \hat{R}_k(D) \quad (4)$$

$$\bar{J}_{\lambda} = \inf_k \hat{J}_{\lambda,k} = \lim_{k \rightarrow \infty} \hat{J}_{\lambda,k}. \quad (5)$$

Shannon coding theorems relate these operationally optimal performances (impossible to compute) to information theoretic minimizations.

$$\bar{\hat{D}}(R) = \bar{D}(R), \quad \bar{\hat{R}}(D) = \bar{R}(D)$$

i.e., operational DRF (RDF) = Shannon DRF (RDF)

To define these Shannon quantities need some definitions from information theory:

The *entropy* of a discrete random variable X is defined as

$$H(X) = - \sum_x \Pr(X = x) \log_2 \Pr(X = x)$$

and the *entropy rate* of a random process $\{X_n\}$ is defined as

$$\overline{H}(X) = \lim_{k \rightarrow \infty} \frac{1}{k} H(X^k)$$

Average mutual information between two discrete random variables X and Y is

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = \sum_{x,y} \Pr(X = x, Y = y) \log_2 \frac{\Pr(X = x, Y = y)}{\Pr(X = x) \Pr(Y = y)}$$

Definitions extend to continuous alphabets by maximizing over all quantized versions:

$$H(X) = \sup_{\alpha} H(\alpha(X))$$

$$I(X; Y) = \sup_{\alpha_1, \alpha_2} I(\alpha_1(X); \alpha_2(Y))$$

Warning: If X is continuous, entropy $H(X)$ is infinite. Do not confuse with differential entropy.

Shannon channel capacity: Channel described by family of conditional probability distributions

$$P_{Y^k|X^k}, k = 1, 2, \dots$$

$$C = \lim_{k \rightarrow \infty} \sup_{P_{X^k}} \frac{1}{k} I(X^k; Y^k)$$

Shannon distortion-rate function: Source described by family of source probability distributions $P_{X^k}, k = 1, 2, \dots$

$$D(R) = \lim_{k \rightarrow \infty} \inf_{P_{Y^k|X^k}: I(X^k; Y^k) \leq kR} \frac{1}{k} E[d_k(X^k, Y^k)]$$

We shall see that for a discrete source:

$$\bar{\hat{R}}(0) = \bar{R}(0) = \bar{H}(X)$$

the entropy rate of the source.

This is the lowest achievable rate for a lossless source code of the given process.

Lagrangian a bit more complicated, equals $D + \lambda R$, where $D = \bar{D}(R)$ at point where λ is the magnitude of the slope of the DRF.

(See, e.g., *Entropy and Information Theory*, Springer, 1992.)

Comments

We will not attempt to approve all the Shannon theory stuff as the general results can be quite complicated and this is not a course in information theory.

The point is to show the form of the results and consider some implications.

We will prove some of the results, including the $R(0)$ result for lossless codes, the $D(0)$ result for zero rate codes, and the fact that the DRF provides a general lower bound to the operational DRF.

Before turning to lossless codes, however, we mention another theoretical approach to source coding

Shannon's distortion rate-theory is asymptotic in that its positive results are for a fixed rate R and asymptotically large block size k (and hence large coding delay)

Another approach:

Bennett/Zador/Gersho asymptotic quantization theory.

Fixed k , asymptotically large R .

High rate or low distortion theory. We will develop some results from this point of view.

Theories consistent when both R and k asymptotically large.

But keep in mind: the real world is not asymptopia!

Example: Suppose that $\{X_n\}$ is iid Gaussian with mean $E(X_n) = 0$ and variance $\sigma_{X_n}^2 = \sigma^2$ and the distortion measure is MSE. Then from basic rate-distortion theory it can be shown that for fixed R

$$\overline{D}(R) = \sigma^2 2^{-2R}$$

is the Shannon DRF. The operational DRF $\hat{D}_k(R)$ lies above this curve and approaches it as $k \rightarrow \infty$

The Bennett/Zador/Gersho result is that for a fixed dimension k and large R

$$\hat{D}_k(R) \approx G_k 2^{-2R}$$

Next: Lossless codes.