Information Rate Distortion Function

Definition The *information rate distortion function* R$^{(I\ )}$*(D)* for a source *X* with distortion measure *d(x, ˆx)* is defined as

$$R^{(I)}(D) = \min_{p(\hat{x}|x):\sum_{(\hat{x},x)} p(x)\,p(\hat{x}|x)\,d(\hat{x},x)\le D} I(X;\hat{X})$$

where the minimization is over all conditional distributions *p(ˆx|x)* for which the joint distribution *p(x, ˆx) = p(x)p(ˆx|x)* satisfies the expected distortion constraint.

Note that we minimize considering *p(ˆx|x)* because H(X) is given while we are interested in *H(X|^X)*.

Rate Distortion Theorem

Theorem The rate distortion function for an i.i.d. source X with distribution p(x) and bounded distortion function d(x, ˆ x) is equal to the associated information rate distortion function. Thus,

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{(\hat{x},x)} p(x)p(\hat{x}|x)d(\hat{x},x) \le D} I(X;\hat{X})$$

is the minimum achievable rate at distortion D.

This theorem shows that the operational definition of the rate distortion function is equal to the information definition. Hence we will use *R(D)* from now on to denote both definitions of the rate distortion function.

Calculation of the Rate Distortion Function for a Binary Source

We now find the description rate *R(D)* required to describe a Bernoulli(*p*) source with an expected proportion of errors less than or equal to *D*.

Theorem The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by:

$$R(D) = \begin{cases} H(p) - H(D) & 0 \le D \le \min\{p, 1-p\} \\ 0 & D \ge \min\{p, 1-p\} \end{cases}$$

Remember that a Bernoulli(p) random variable is a binary variable with takes on 1 with probability p.

Proof

Consider a binary source X ~ Bernoulli*(p)* with a Hamming distortion measure. Without loss of generality, we may assume that $p < 1/2$.

We wish to calculate the rate distortion function,

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

Let ⊕ denote modulo 2 addition. Thus, X ⊕ ˆX = 1 is equivalent to X ≠ ˆX.

We do not minimize *I (X*; ˆ *X)* directly; instead, we find a lower bound and then show that this lower bound is achievable.

Proof: Lower Bound on R(D)

For any joint distribution satisfying the distortion constraint, we have

If ^X is given, there is not variation on H if we consider X ⊕ ^X instead of X

$$
\begin{aligned}
& I(X;\hat{X}) = H(X) - H(X \mid \hat{X}) \\
& = H(p) - H(X \oplus \hat{X} \mid \hat{X}) \\
& \geq H(p) - H(X \oplus \hat{X}) \\
& \geq H(p) - H(D)
\end{aligned}
$$

since Pr*(X ≠ ˆX)* ≤ *D* and *H(D)* increases with *D* for *D* ≤ 1/2 . Thus,

$$R(D) \geq H(p) - H(D)$$

Proof: Achievability

We now show that the lower bound is actually the rate distortion function by finding a joint distribution that meets the distortion constraint and has
I (X; *ˆX) = R(D).*

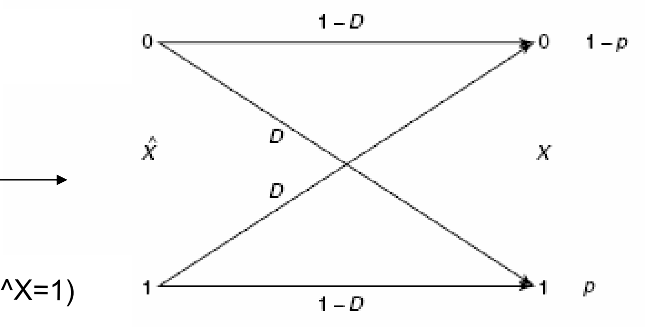
For 0 ≤ *D* ≤ *p*, we can achieve the value of the rate distortion function by choosing *(X, ˆX)* to have the joint distribution given by the binary symmetric channel shown in Figure

We would like to find a D such that H(p)-H(D)=I(X;^X).

Recall that for a BSC I(X;Y)=H(Y)-H(p). Here p corresponds to D and Y to p.

Then we impose that X follows Bernoulli

r = Pr(^X=1)



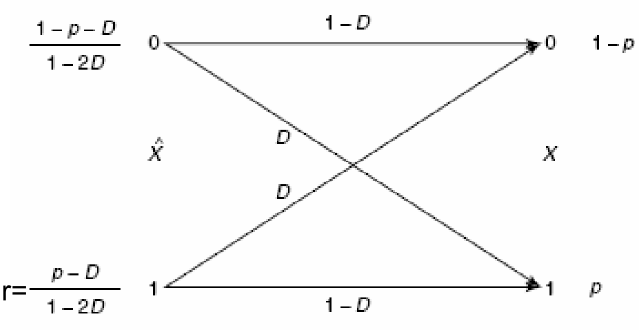
Proof: Achievability

We choose the distribution of ˆ*X* at the input of the channel so that the output distribution of *X* is the specified distribution. Let *r* = Pr*(* ˆ*X* = 1*)*.

Then choose *r* so that:

$$r(1-D)+(1-r)D = p$$

$$r = \frac{p-D}{1-2D}$$



If *D* ≤ *p* ≤ 1/2, then Pr*(* ˆ*X* = 1*)* ≥ 0 and Pr*(* ˆ*X* = 0*)* ≥ 0. We then have

$$I(X;\hat{X}) = H(X) - H(X \mid \hat{X}) = H(p) - H(D)$$

and the expected distortion is Pr*(X* ≠ ˆ*X)* = *D* as can be noted from the Figure. Indeed, the uncertainty of X when if ^X is known is D, hence H(X|^X)=H(D).

Proof: Achievability

If $D \geq p$, we can achieve $R(D) = 0$ by letting $\hat{}X = 0$ with probability 1.

In this case, $I\,(X;\ \hat{}\ X) = 0$ and $D = p$.

Similarly, if $D \geq 1 - p$, we can achieve $R(D) = 0$ by setting $\hat{}X = 1$ with probability 1.

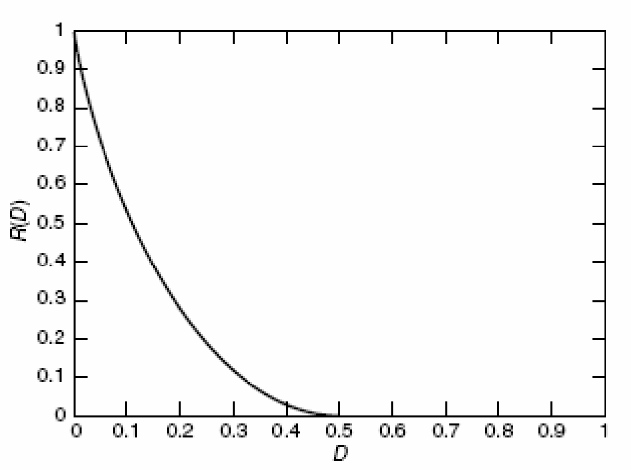
Hence, the rate distortion function for a binary source is

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

Note that for D≥p, if we set ^X=0, then H(X|^X)=H(X) since ^X is independent from X, then I(X;^X) = 0. The uncertainty on X is D also in this case.

R(D)

This function is illustrated in Figure



Rate Distortion of a Gaussian Source

We calculate the rate distortion function for a Gaussian source with squared-error distortion.

Theorem The rate distortion function for a N(0, σ^2) source with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & 0 \le D \le \sigma^2 \\ 0 & D > \sigma^2 \end{cases}$$

Proof: Lower Bound

Let X be ~ $N(0, \sigma^2)$. By the rate distortion theorem extended to continuous alphabets, we have

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

As in the preceding example, we first find a lower bound for the rate distortion function and then prove that this is achievable.

Since $E(X - \hat{X})^2 \le D$, we observe that

Proof: Lower Bound

$$I(X;\hat{X}) = h(X) - h(X \mid \hat{X})$$

= if X-^X=Z indep from ^X. Then h(X-^X|^X)=h(Z)

conditioning reduces entropy

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

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

= if Z~N(0,D)

The normal distribution maximizes entropy for a given variance

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

$$= \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)E(X-\hat{X})^2$$

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

= if E(X-^X)^2=D

$$= \frac{1}{2}\log\frac{\sigma^2}{D}$$

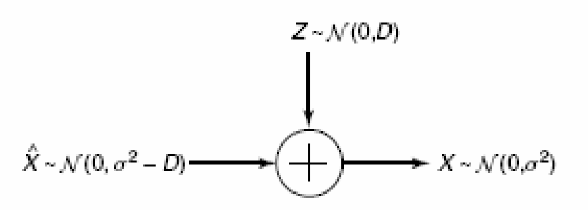
Proof: Achievability

Hence

$$R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$$

To find the conditional density *f (ˆx|x)* that achieves this lower bound, it is usually more convenient to look at the conditional density *f (x|ˆx)*, which is sometimes called the *test channel* (thus emphasizing the duality of rate distortion with channel capacity).

As in the binary case, we construct *f (x|ˆx)* to achieve equality in the bound. We choose the joint distribution as shown in Figure.


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

Proof: Achievability

If $D \leq \sigma^2$, we choose

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

where $\hat{}X$ and Z are independent. For this joint distribution, we calculate

$$I(X;\hat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}$$

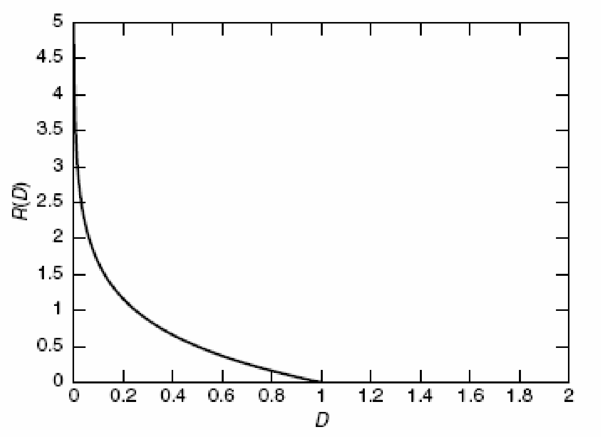
and $E(X - \hat{}X)^2 = D$, thus achieving the bound. If $D > \sigma^2$, we choose $\hat{}X = 0$ with probability 1, achieving $R(D) = 0$. Hence, the rate distortion function for the Gaussian source with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & 0 \leq D \leq \sigma^2 \\ 0 & D > \sigma^2 \end{cases}$$

R(D)

as illustrated in Figure. We can also express in terms of the rate.

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



Comments

Each bit of description reduces the expected distortion by a factor of 4.

With a 1-bit description, the best expected square error is $\sigma^2/4$.

We can compare this with the result of simple 1-bit quantization of a $N(0, \sigma^2)$ random variable

In this case, using the two regions corresponding to the positive and negative real lines and reproduction points as the centroids of the respective regions, the expected distortion is $((\pi-2)/\pi)\,\sigma^2 = 0.3633\sigma^2$.

Converse to the Rate Distortion Theorem

It is possible to prove that we cannot achieve a distortion of less than *D* if we describe *X* at a rate less than *R(D)*, where

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

Lemma (Convexity of R(D)) The rate distortion function R(D) given in is a nonincreasing convex function of D.

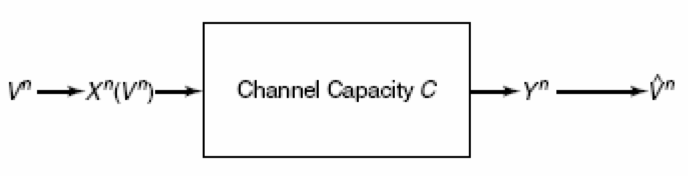
This shows that the rate *R* of any rate distortion code exceeds the rate distortion function *R(D)* evaluated at the distortion level $D = Ed(X^n, \hat{}X^n)$ achieved by that code.

Source-Channel Separation Theorem with Distortion

A similar argument can be applied when the encoded source is passed through a noisy channel and hence we have the equivalent of the source channel separation theorem with distortion:

Theorem Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel with capacity C.

The output of the channel Y^n is mapped onto the reconstruction alphabet ˆV^n = g(Y^n). Let D = Ed(V^n, ˆV^n) = $1/n\sum_{i=1}^{n} Ed(V_i, \hat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme. Then distortion D is achievable if and only if C > R(D).



Channel Coding for the Gaussian Channel: A Sphere Packing Problem

Consider a Gaussian channel, $Yi = Xi + Zi$, where the Zi are i.i.d. ~ $N(0,N)$ and there is a power constraint P on the power per symbol of the transmitted codeword.

Consider a sequence of n transmissions. The power constraint implies that the transmitted sequence lies within a sphere of radius $\surd nP$ in Rn.

The coding problem is equivalent to finding a set of $2nR$ sequences within this sphere such that the probability of any of them being mistaken for any other is small—the spheres of radius $\surd nN$ around each of them are almost disjoint.

This corresponds to filling a sphere of radius $\surd n(P + N)$ with spheres of radius $\surd nN$.

One would expect that the largest number of spheres that could be fit would be the ratio of their volumes, or, equivalently, the *n*th power of the ratio of their radii.

Thus, if *M* is the number of codewords that can be transmitted efficiently, we have

$$M \leq \frac{(\sqrt{n(P+N)})^n}{(\sqrt{nN})^n} = \left(\frac{P+N}{N}\right)^{n/2}$$

The results of the channel coding theorem show that it is possible to do this efficiently for large *n*; it is possible to find approximately

$$2^{nC} = \left(\frac{P+N}{N}\right)^{n/2}$$

codewords such that the noise spheres around them are almost disjoint (the total volume of their intersection is arbitrarily small).

Rate Distortion for the Gaussian Source: A Sphere Covering Problem

Consider a Gaussian source of variance σ^2. A *$(2^{nR}, n)$* rate distortion code for this source with distortion D is a set of 2^{nR} sequences in R^n such that most source sequences of length n (all those that lie within a sphere of radius $\surd n\sigma^2$) are within a distance $\surd nD$ of some codeword.

Again, by the sphere-packing argument, it is clear that the minimum number of codewords required is

$$2^{nR(D)} = \left(\frac{\sigma^2}{D} \right)^{n/2}$$

The rate distortion theorem shows that this minimum rate is asymptotically achievable (i.e., that there exists a collection of spheres of radius $\surd nD$ that cover the space except for a set of arbitrarily small probability).

Putting it All Together

- The above geometric arguments also enable us to transform a good code for channel transmission into a good code for rate distortion. In both cases, the essential idea is to fill the space of source sequences.

- In channel transmission, we want to find the largest set of codewords that have a large minimum distance between codewords (given a bound on P)
- In rate distortion, we wish to find the smallest set of codewords that covers the entire space (given a bound on D)

- In the Gaussian case, choosing the codewords to be Gaussian with the appropriate variance is asymptotically optimal for both rate distortion and channel coding.