

Quantization, Compression, and Classification

Robert M. Gray

Information Systems Laboratory, Dept of Electrical Engineering

Stanford, CA 94305

rmgray@stanford.edu

Partitions and encoders

X a random object (variable, vector, field, signal) taking values
alphabet A (e.g., \mathbb{R}^k)

P_X a probability distribution describing X

E.g., described by pdf f_X

$$P_X(F) = \int_F f_X(x) dx$$

or pmf p_X

$$P_X(F) = \sum_{x \in F} p_X(x)$$

Partition $\mathcal{S} = \{S_i; i \in \mathcal{I}\}$ of A :

$$S_i \cap S_j = \emptyset, i \neq j$$

$$\bigcup_{i \in \mathcal{I}} S_i = A$$

Usually $\mathcal{I} = \mathcal{Z} = \{0, 1, 2, \dots\}$

partition *finite* if # atoms $|\mathcal{S}| = |\mathcal{I}| = N(\mathcal{S}) = N(\mathcal{I}) < \infty$.

Sometimes allow empty atoms for convenience, but count only includes nonempty atoms.

Quantizer *encoder*: $\mathcal{E} : A \rightarrow \mathcal{Z}$

Index set of the quantizer $\mathcal{I} = \mathcal{E}(A) \subset \mathcal{Z}$

Later consider also quantizer *decoder* $\mathcal{D} : \mathcal{I} \rightarrow \hat{A}$

Encoder and partition are equivalent concepts:

$\mathcal{E} \Rightarrow$ cells $S_i = \mathcal{E}^{-1}(i) = \{x : \mathcal{E}(x) = i\}$, $i \in \mathcal{I}$.

$\mathcal{S} \Rightarrow \mathcal{E}$, e.g.,

$$\mathcal{E}(x) = \sum_{i \in \mathcal{I}} i 1(x \in S_i), \quad (1)$$

where

$$1(x \in S) = \begin{cases} 1 & x \in S \\ 0 & \text{otherwise.} \end{cases}$$

\mathcal{E} or $\mathcal{S} \Rightarrow$ probability mass function (pmf) $p = \{p(i); i \in \mathcal{I}\}$:

$$p(i) = P_X(S_i) = \Pr(\mathcal{E}(X) = i).$$

Entropy

Shannon *entropy* of a partition and distribution

$$H(\mathcal{S}) = - \sum_i P(S_i) \log P(S_i), \quad (2)$$

Usually $\log = \log_2$ (bits) or \ln (nats)

By convention $0 \ln 0 = 0$.

Also denoted by $H(\mathcal{E}(X))$ and

$$H(p) = \sum_{i \in \mathcal{I}} p(i) \ln \frac{1}{p(i)}$$

Since $p(i) \in [0, 1] \Rightarrow H(p) \geq 0$.

Partition \mathcal{S}' refines \mathcal{S} or $\mathcal{S} \subset \mathcal{S}'$ if atoms of \mathcal{S} are unions of atoms of \mathcal{S}' .

\mathcal{S} is a *subpartition* of \mathcal{S}'

\mathcal{S}' is a *superpartition* of \mathcal{S}

$$\mathcal{S} \subset \mathcal{S}' \Rightarrow N(\mathcal{S}) \leq N(\mathcal{S}')$$

Similar property for entropy:

Lemma 1. *If $\mathcal{S} \subset \mathcal{S}'$, then $H(\mathcal{S}) \leq H(\mathcal{S}')$.*

Proof Each atom of \mathcal{S} has the form $S_i = \cup_{j:S'_j \subset S_i} S'_j$ and the contribution of the atom to the entropy is

$$\begin{aligned} -P(S_i) \ln P(S_i) &= -P(\cup_{j:S'_j \subset S_i} S'_j) \ln P(\cup_{j:S'_j \subset S_i} S'_j) \\ &= -\left(\sum_{j:S'_j \subset S_i} P(S'_j) \right) \ln \left(\sum_{j:S'_j \subset S_i} P(S'_j) \right) \quad (3) \end{aligned}$$

$$\begin{aligned} &= -\sum_{j:S'_j \subset S_i} P(S'_j) \ln \left(\sum_{l:S'_l \subset S_i} P(S'_l) \right) \\ &\leq -\sum_{j:S'_j \subset S_i} P(S'_j) \ln P(S'_j), \quad (4) \end{aligned}$$

so that

$$\begin{aligned} H(\mathcal{S}) &= - \sum_{i \in \mathcal{I}} P(S_i) \ln P(S_i) \\ &\leq - \sum_{i \in \mathcal{I}} \sum_{j: S'_j \subset S_i} P(S'_j) \ln P(S'_j) \\ &= - \sum_{j \in \mathcal{I}'} P(S'_j) \ln P(S'_j) = H(\mathcal{S}'). \end{aligned}$$

□

⇒ the finer the quantization, the larger the entropy and codebook size!

Relative entropy

relative entropy or *Kulback-Leibler divergence*

Defined in terms of *two* distributions, P and Q , and a partition \mathcal{S} :

$$H(P\|Q, \mathcal{S}) = \sum_i P(S_i) \ln \frac{P(S_i)}{Q(S_i)}. \quad (5)$$

Often expressed in terms of $p(i) = P(S_i)$ and $q(i) = Q(S_i)$

$$H(p\|q) = \sum_{i \in \mathcal{I}} p(i) \ln \frac{p(i)}{q(i)}. \quad (6)$$

Require that if $q(i) = 0$, then also $p_i = 0$ and define $p_i \ln(p_i/q_i) = 0$.

p is *absolutely continuous* with respect to q , $p \ll q$.

Lemma 2. The divergence inequality *For any pmfs p and q ,*

$$H(p||q) \geq 0, \quad (7)$$

with equality if and only if $p(i) = q(i)$ for all i .

Proof Follows from elementary inequality

$$\ln a \leq a - 1 \quad (8)$$

for positive a , with equality if and only if $a = 1$:

$$\sum_i p(i) \ln \frac{q(i)}{p(i)} \leq \sum_i p(i) \left(\frac{q(i)}{p(i)} - 1 \right) = 0,$$

with equality if and only if $p(i) = q(i)$ for all i .

Also follows from Jensen's inequality and the concavity of the \ln function:

$$-H(p||q) = \sum_i p(i) \ln \frac{q(i)}{p(i)} \leq \ln \left(\sum_i p(i) \frac{q(i)}{p(i)} \right) = 0.$$

□

Implication: If \mathcal{S} has $N = N(\mathcal{S}) < \infty$ atoms S_i ; $i = 0, 1, \dots, N - 1$, then setting $q(i) = 1/N$ for all i yields

$$H(\mathcal{S}) \leq \ln N(\mathcal{S}). \quad (9)$$

A finite partition will have the maximum possible entropy if all of its atoms have equal probability.

Decoders

Use encoder output $i = \mathcal{E}(X)$ to provide an estimate/approximation/classification/decision regarding the original input X : $\mathcal{D}(i)$

Often abbreviate overall operation to

$$q(x) = \mathcal{D}(\mathcal{E}(x)).$$

but also use q to mean collection of components $(\mathcal{E}, \mathcal{D})$.

The collection $\mathcal{C} = \{\mathcal{D}(i); i \in \mathcal{I}\}$ of possible decoder outputs is called the *reproduction codebook* or, simply, *codebook*

Decoder is equivalent to (ordered) codebook and can write $q = (\mathcal{S}, \mathcal{C})$.

Uniform quantization

Finite codebook size

Assume $f_X(x)$ has finite support $[a, b]$ and carve up interval into N equal pieces of width (or *bin width*) $\Delta = (b - a)/N$ each (10).

$$q(x) = \begin{cases} \text{quantization levels} & \text{quantization cells} \\ \mathcal{D}(N-1) = b - \frac{\Delta}{2}; & x \in S_{N-1} = [b - \Delta, b] \\ \mathcal{D}(i) = a + (i + \frac{1}{2})\Delta; & x \in S_i = [a + i\Delta, a + (i + 1)\Delta) \\ & i = 1, \dots, N-2 \\ \mathcal{D}(0) = a + \frac{\Delta}{2}\Delta; & x \in S_0 = [a, a + \Delta) \end{cases} \quad (10)$$

Equivalently, the quantizer can be expressed as

$$q(x) = \sum_{i=0}^{N-1} \left[a + \left(i + \frac{1}{2} \right) \Delta \right] 1_{[a+i\Delta, a+(i+1)\Delta)}(x).$$

Can extend to entire real line by extending top and bottom cells, \mathcal{S}_0 becomes $(-\infty, a + \Delta)$

but no longer genuinely uniform.

Infinite codebook size

If allow an infinity of levels, fix Δ and define

$$q(x) = \sum_{i=-\infty}^{\infty} \left[a + \left(i + \frac{1}{2} \right) \Delta \right] 1_{[a+i\Delta, a+(i+1)\Delta)}(x).$$

where a is a design parameter, e.g., $a = 0$.

Distortion

$$d(x, y)$$

$X \in \mathfrak{R}^k$, squared error distortion measure

$$d(X, \mathcal{D}(\mathcal{E}(X))) = \|X - \mathcal{D}(\mathcal{E}(X))\|^2, \text{ where}$$

$$\|x - y\|^2 = (x - y)^t(x - y) = \sum_{i=0}^{k-1} |x_i - y_i|^2,$$

$x = (x_0, \dots, x_{k-1})^t, y = (y_0, \dots, y_{k-1})^t \in \mathfrak{R}^k$ are column vectors and x^t denotes the transpose of x .

$$\text{Quantizer performance } D(q) = D(\mathcal{E}, \mathcal{D}) = E[\|X - \mathcal{D}(\mathcal{E}(X))\|^2]$$

Tractable, simple for analysis, can make perceptually meaningful by using weighted quadratic distortion measures.

The centroid condition

Using nested or iterated expectation:

$$D(\mathcal{E}, \mathcal{D}) = \sum_i E[\|X - \mathcal{D}(i)\|^2 | \mathcal{E}(X) = i] \Pr(\mathcal{E}(X) = i).$$

For each i the conditional expectation $E[\|X - \mathcal{D}(i)\|^2 | \mathcal{E}(X) = i]$ is minimized by the conditional expectation

$$\mathcal{D}(i) = E(X | \mathcal{E}(X) = i). \quad (11)$$

In general: $E(X|Y)$ is the minimum mean squared estimate of X given Y .

General Lloyd centroid: $\text{cent}(S) = \operatorname{argmin}_y E[d(X, y) | X \in S]$
(if minimum exists)

Centroidal quantizers

If a quantizer $(\mathcal{E}, \mathcal{D})$ is to be optimal, then necessarily the decoder \mathcal{D} must be optimal for the encoder \mathcal{E} . A quantizer satisfying the centroid condition is said to be a *centroidal quantizer*.

Recall that $q(x) = \sum_i \mathcal{D}(i) 1_{S_i}(x)$

$$\begin{aligned} E[q(X)] &= E\left[\sum_i \mathcal{D}(i) 1_{S_i}(X)\right] \\ &= \sum_i \mathcal{D}(i) E[1_{S_i}(X)] \\ &= \sum_i E(X|X \in S_i) P_X(S_i) \\ &= E(X), \end{aligned} \tag{12}$$

First moment is preserved by a centroidal quantizer.

Similarly, correlation:

$$\begin{aligned} E[X^t q(X)] &= E[X^t \sum_i \mathcal{D}(i) 1_{S_i}(X)] \\ &= \sum_i E[X^t 1_{S_i}(X)] \mathcal{D}(i) \\ &= \sum_i \mathcal{D}(i)^t \mathcal{D}(i) P_X(S_i) \\ &= E(\|q(X)\|^2), \end{aligned} \tag{13}$$

\Rightarrow the input and output are positively correlated.

Combining the two previous results produces

$$E[q(X)^t(q(X) - X)] = 0, \quad (14)$$

(orthogonality principal!)

Define the error vector $\epsilon = q(X) - X$, then

$$\begin{aligned} E(\|\epsilon\|^2) &= E((q(X) - X)^t(q(X) - X)) \\ &= E(q(X)^t(q(X) - X)) - E(X^t(q(X) - X)) \\ &= 0 - E(X^t q(X)) + E(\|X\|^2) \\ &= E(\|X\|^2) - E(\|q(X)\|^2). \end{aligned} \quad (15)$$

Thus $E(\|X\|^2) \geq E(\|q(X)\|^2)$ and

$$E(\|q(X)\|^2) = E(\|X\|^2) - E(\|\epsilon\|^2) \quad (16)$$

Equal means \Rightarrow

$$E(\|\epsilon\|^2) = \sigma_X^2 - \sigma_{q(X)}^2,$$

$$\Rightarrow \sigma_X^2 \geq \sigma_{q(X)}^2$$

(15) \Rightarrow

$$E(X^t \epsilon) = -E(\|\epsilon\|^2), \quad (17)$$

which contradicts the common assumption that the input and quantizer error are uncorrelated!

Partition cost and rate

Since an encoder implies an optimal decoder, question is how small the average distortion

$$D(\mathcal{E}) = \min_{\mathcal{D}} D(\mathcal{E}, \mathcal{D})$$

can be made by choosing the “best” encoder.

If no further constraints, this is generally 0 (except for remote source problem).

Constraint: there is a *cost* $R(\mathcal{E})$ associated with a partition \mathcal{E} .

Assume $R(\mathcal{E}) \geq 0$,

Also: If $\mathcal{S} \subset \mathcal{S}'$, then $R(\mathcal{E}) \leq R(\mathcal{E}')$.

In general, arbitrarily small average distortion requires arbitrarily high cost.

⇒ optimization goal is optimal *tradeoff*

Core of the subject of quantization theory, algorithms, and applications of optimal (or at least good) distortion-rate tradeoffs.

Two candidate notions of cost meeting these conditions are the size of the partition $N(\mathcal{S})$ and the entropy of the partition $H(\mathcal{S})$.

Fixed and variable rate codes

$R(\mathcal{E}) = \ln N(\mathcal{S})$ assigns fixed cost (called *instantaneous rate*) $r(i) = \ln N(\mathcal{S})$ to each index $i \Rightarrow$ average cost (*average rate*)

$$R(\mathcal{E}) = E[r(\mathcal{E}(X))] = \sum_i p(i)r(i) = \ln N(\mathcal{S}).$$

$R(\mathcal{E}) = H(\mathcal{S})$ assigns variable cost $r(i) = -\ln p(i)$ to each index $i \Rightarrow$ average cost

$$\begin{aligned} R(\mathcal{E}) &= E[r(\mathcal{E}(X))] = \sum_i p(i)r(i) \\ &= \sum_i p(i) \ln \frac{1}{p(i)} = H(\mathcal{S}). \end{aligned}$$

More general variable cost/rate assignment: $r(i) = \ell(i)$, ℓ is a *length function*, a positive sequence of numbers with the property that

$$\sum_i e^{-\ell(i)} \leq 1, \quad (18)$$

Kraft's inequality

The special case of $\ell(i) = -\ln p(i)$ known as the *Shannon codelengths*

For general length function

$$\begin{aligned} r(i) &= \ell(i) \\ R(\mathcal{E}, \ell) &= E[\ell(\mathcal{E}(X))] \\ &= \sum_i p(i)\ell(i). \end{aligned}$$

In this general form, $q = (\mathcal{E}, \mathcal{D}, \ell)$

Note: *fixed rate* a special case with $\ell(i) = 1/N$.

Alternative characterization: each codeword has a *weighting* or *importance weighting*

$$w(i) = e^{-\ell(i)}$$

Kraft's inequality for $\ell \Leftrightarrow w$ a sub-pmf:

$$w_i \geq 0 \text{ all } i; \sum_i w(i) \leq 1.$$

$\ell(i) = \infty \Leftrightarrow w(i) = 0$, index or codeword never used.

$$N(w) = |\{i : w(i) > 0\}| = N(\ell) = |\{i : \ell(i) \text{ finite}\}| \quad (19)$$

Combined constraint [Zador (1982)] for unified treatment:

$$\begin{aligned}r_{\eta}(i) &= (1 - \eta)\ell(i) + \eta \ln N(\ell) \\ &= (1 - \eta) \ln \frac{1}{w(i)} + \eta \ln N(w) \\ R_{\eta}(\mathcal{E}, \ell) &= (1 - \eta)E[\ell(\mathcal{E}(X))] + \eta \ln N(\ell) \\ &= (1 - \eta) \sum_i p(i) \ln \frac{1}{w(i)} + \eta \ln N(w) \\ &= R_{\eta}(\mathcal{E}, w)\end{aligned}\tag{20}$$

$\eta \in [0, 1]$. (fixed-rate is $\eta = 1$, pure variable-rate is $\eta = 0$)

Optimal quantization

Several possible definitions:

operational distortion-rate function

$$\delta_\eta(R) = \inf_{\mathcal{E}, \mathcal{D}, w: R(\mathcal{E}, w) \leq R} D(\mathcal{E}, \mathcal{D}),$$

operational rate-distortion function (dual)

$$r_\eta(D) = \inf_{\mathcal{E}, \mathcal{D}, w: D(\mathcal{E}, \mathcal{D}) \leq D} R(\mathcal{E}, w),$$

Lagrangian optimization,

Lagrangian multiplier $\lambda > 0$

$$\rho(\lambda, \eta) = \inf_{\mathcal{E}, \mathcal{D}, w} (D(\mathcal{E}, \mathcal{D}) + \lambda R(\mathcal{E}, w))$$

Effectively incorporates rate constraint into general Lagrangian distortion.

Traditional cases $\eta = 1, 0$

The distortion-rate and rate-distortion formulations are equivalent

Optimization theory: Lagrangian minimization yields distortion-rate points on convex hull of optimal distortion-rate pairs.

None of these problems are in general convex.

Optimality properties and the Lloyd algorithm: $\eta = 1$

Already seen: necessary condition for optimality that given \mathcal{E} ,
 $\mathcal{D}(i) = E[X | \mathcal{E}(X) = i]$

For $\eta = 1$, Lloyd showed similarly that encoder must be optimal for decoder:

$$\begin{aligned} D(\mathcal{E}, \mathcal{D}) &= \int dP(x) \|x - \mathcal{D}(\mathcal{E}(x))\|^2 \\ &\geq \int dP(x) \min_i \|x - \mathcal{D}(i)\|^2, \end{aligned}$$

bound achieved by *nearest neighbor* or *minimum distortion* encoder

$$\mathcal{E}(x) = \operatorname{argmin}_i \|x - \mathcal{D}(i)\|, \quad (21)$$

Break ties in arbitrary fashion, e.g., by lowest index value i

Implicit encoder in Shannon's development.

Lloyd argued also necessary condition for optimality is that $P(S_i) > 0$ for all i since otherwise can strictly reduce rate with no increase in distortion. (True for all $\eta \in (0, 1]$, but not necessary for $\eta = 0$.)

Necessary conditions for optimality: fixed-rate

- $\mathcal{D}(i) = E(X | \mathcal{E}(X) = i)$ for all $i \in \mathcal{I}$.
- $\mathcal{E}(x) = \operatorname{argmin}_i \|x - \mathcal{D}(i)\|$ for all x .
- $\Pr(\mathcal{E}(x) = i) > 0$ for all $i \in \mathcal{I}$.

Lloyd quantizer improvement algorithm: fixed-rate

Step 0 Let \mathcal{E}_0 be an initial encoder. Set $m = 0$.

Step 1 Optimize the decoder \mathcal{D}_m for the encoder \mathcal{E}_m : $\mathcal{D}_m(i) = E(X | \mathcal{E}_m(X) = i)$ for all $i \in \mathcal{I}$.

Step 2 Optimize the encoder \mathcal{E}_{m+1} for the decoder \mathcal{D}_m : $\mathcal{E}_{m+1}(x) = \operatorname{argmin}_i \|x - \mathcal{D}_m(i)\|$ for all $x \in A$.

Step 3 Prune useless codewords. If $\Pr(\mathcal{E}_{m+1}(X) = i) = 0$, remove i from \mathcal{I} and merge the corresponding partition cell into an arbitrary remaining partition cell.

Step 3 Set $m + 1 \rightarrow m$ and go to Step 1.

Variation: initialize with a decoder

$D(\mathcal{E}_m)$ is nonincreasing in $m \Rightarrow$ descent algorithm, distortion must converge.

In practice: test the improvement and stop when some threshold was reached, e.g., when

$$\frac{D(\mathcal{E}_{m-1}) - D(\mathcal{E}_m)}{D(\mathcal{E}_{m-1})} \leq \epsilon.$$

Lloyd algorithm is a *clustering algorithm*

Earliest example of *alternating optimization* (AO) algorithm

Also renamed k-means, principal points, . . .

Resulting partition is a *centroidal Voronoi* partition.

Simple example:

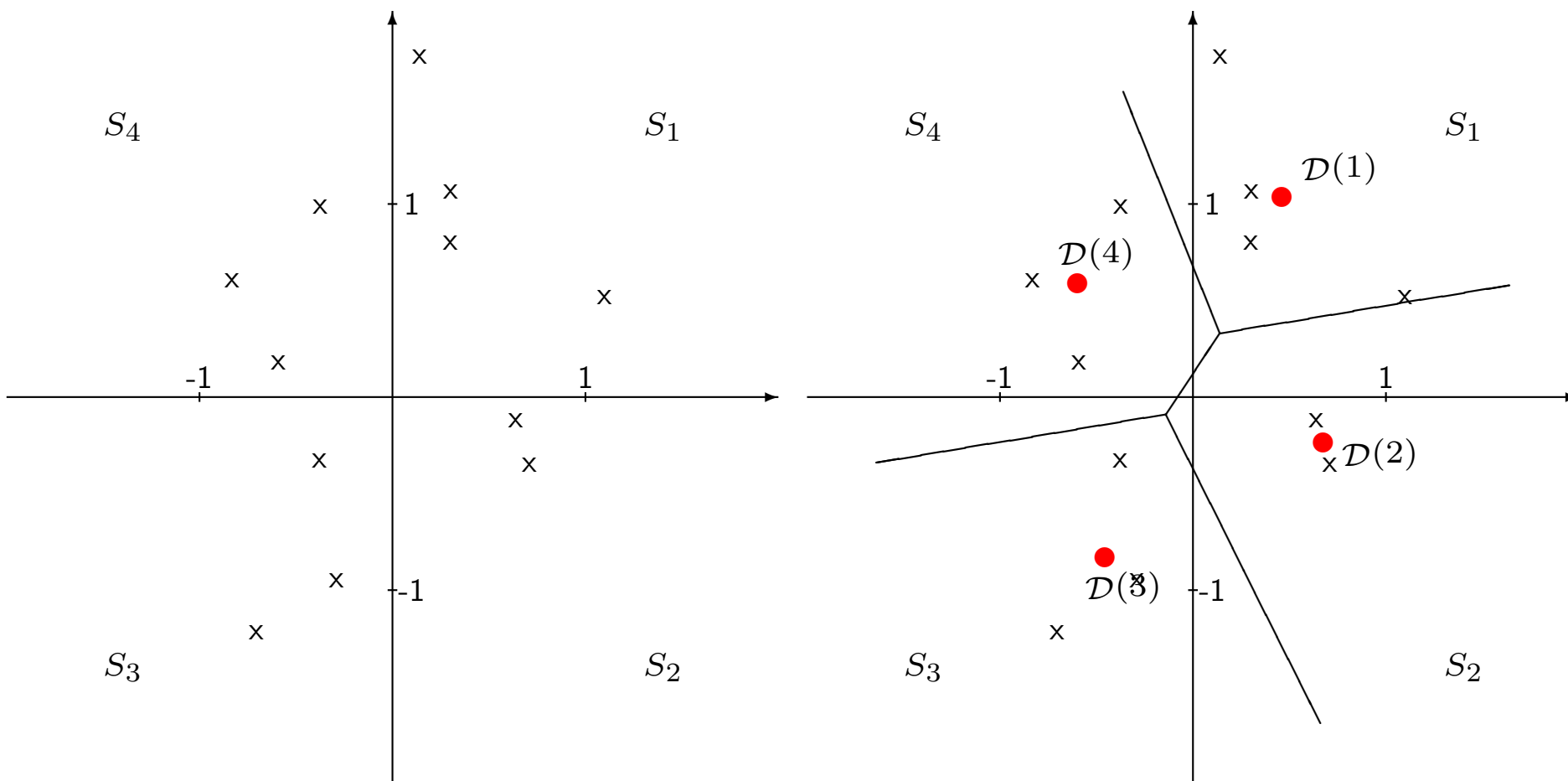


Figure 1: Lloyd algorithm

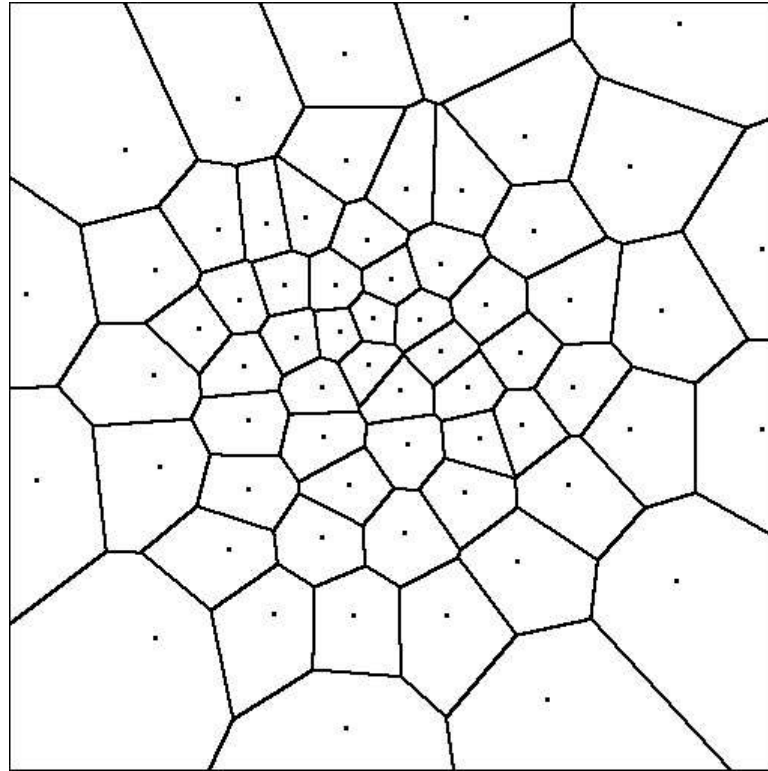


Figure 2: Two-dimensional centroidal Voronoi diagram

The Lloyd algorithm: entropy constraint

Lloyd iteration almost works in the entropy-constrained case with

$$R(\mathcal{E}) = - \sum_i p(i) \ln p(i) = H(\mathcal{S})$$

using a Lagrangian formulation:

$$\begin{aligned} D(\mathcal{E}, \mathcal{D}) + \lambda H(\mathcal{S}) &= \int dP(x) (\|x - \mathcal{D}(\mathcal{E}(x))\|^2 - \lambda \ln p(\mathcal{E}(x))) \\ &\geq \int dP(x) \min_i (\|x - \mathcal{D}(i)\|^2 - \lambda \ln p(i)) , \end{aligned}$$

Achieve lower bound with minimum Lagrangian distortion

$$\mathcal{E}(x) = \operatorname{argmin}_i \left(\|x - \mathcal{D}(i)\|^2 + \lambda \ln \frac{1}{p(i)} \right). \quad (22)$$

Problem: Definition is circular, encoder defined in terms of the probabilities $p(i) = \Pr(\mathcal{E}(X) = i)$ which depend upon the encoder.

Easy fix, introduce another pmf w which when optimized gives p , decouple encoder from rate.

Using divergence inequality,

$$\begin{aligned}
\inf_{\mathcal{E}, \mathcal{D}} (D(\mathcal{E}, \mathcal{D}) + \lambda H(\mathcal{E})) &= \inf_{\mathcal{E}, \mathcal{D}, w} (D(\mathcal{E}, \mathcal{D}) + \lambda [H(\mathcal{E}) + H(p||w)]) \\
&= \inf_{\mathcal{E}, \mathcal{D}, w} \left(D(\mathcal{E}, \mathcal{D}) + \lambda \sum_i p(i) \ln \frac{1}{w(i)} \right) \\
&= \inf_{\mathcal{E}} \left(\inf_{\mathcal{D}} D(\mathcal{E}, \mathcal{D}) + \lambda \inf_w \sum_i p(i) \ln \frac{1}{w(i)} \right).
\end{aligned}$$

infimum is over all $w: p \ll w$

Slight extension yields general variable-rate case:

Lemma 3. *Let \mathcal{E} be an encoder with index pmf $p = \{p(i); i \in \mathcal{I}\}$.
Then*

$$H(\mathcal{E}) = \inf_w \sum_i p(i) \ln \frac{1}{w(i)}$$

where the infimum is over all sub-pmf's w , that is, all nonnegative $w = \{w(i); i \in \mathcal{I}\}$ for which

$$\sum_{i \in \mathcal{I}} w(i) \leq 1. \tag{23}$$

Proof Divergence inequality & constraint on $w \Rightarrow$

$$\begin{aligned} \sum_i p(i) \ln \frac{1}{w(i)} &= \sum_i p(i) \ln \frac{p(i)}{w(i) / \sum_j w(j)} + H(\mathcal{E}) - \ln \left(\sum_j w(j) \right) \\ &\geq H(\mathcal{E}). \end{aligned}$$

achievable iff

$$\sum_{i \in \mathcal{I}} w(i) = 1 \quad (24)$$

and

$$p(i) = \frac{w(i)}{\sum_j w(j)} = w(i) \quad (25)$$

and hence

$$\min_w \left(\sum_i p(i) \ln \frac{1}{w(i)} \right) = H(\mathcal{E}) \quad \square$$

Extension to combined constraint case $R_\eta(\mathcal{E}, w)$

Corollary 1.

$$\inf_{\mathcal{E}, \mathcal{D}, w} (D(\mathcal{E}, \mathcal{D}) + \lambda R_\eta(\mathcal{E}, w)) \quad (26)$$

$$\begin{aligned} &= \inf_{\mathcal{E}} \left(\inf_{\mathcal{D}} D(\mathcal{E}, \mathcal{D}) + \lambda \inf_w R_\eta(\mathcal{E}, w) \right) \\ &= \inf_{\mathcal{E}} (D(\mathcal{E}) + \lambda [(1 - \eta)H(\mathcal{E}) + \eta \ln N(\mathcal{E})]), \quad (27) \end{aligned}$$

where $N(\mathcal{E})$ is the number of indices for which $p(i) = \Pr(X \in S_i) > 0$.

Proof: The lemma implies that

$$\begin{aligned} & D(\mathcal{E}, \mathcal{D}) + \lambda R(\mathcal{E}, w) \\ &= D(\mathcal{E}, \mathcal{D}) + \lambda \left((1 - \eta) \sum_i p(i) \ln \frac{1}{w(i)} + \eta N(w) \right) \\ &\geq D(\mathcal{E}, \mathcal{D}) + \lambda ((1 - \eta)H(\mathcal{E}) + \eta N(w)) \end{aligned}$$

with equality if $w = p$. Furthermore, since $p \ll w$, $N(\mathcal{E}) = N(p) \leq N(w)$, which is achieved if $w = p$. \square

The necessary conditions on \mathcal{E}, \mathcal{D} and w to minimize $D(\mathcal{E}, \mathcal{D}) + \lambda R(\mathcal{E}, w)$ now become

Necessary conditions for optimality of $(\mathcal{E}, \mathcal{D}, q)$

- $\mathcal{D}(i) = E(X | \mathcal{E}(X) = i)$ all $i \in \mathcal{I}$.
- $w(i) = \Pr(\mathcal{E}(X) = i)$, all $i \in \mathcal{I}$.
- $\mathcal{E}(x) = \operatorname{argmin}_i \left(\|x - \mathcal{D}_m(i)\|^2 + \lambda \ln \frac{1}{w(i)} \right)$ for all x .
- If $\eta \neq 0$, then $P_X(S_i) > 0$, all $i : w(i) \neq 0$.

Arguably a good idea to prune zero probability cells even if $\eta = 0$

Lloyd quantizer improvement algorithm entropy cost function

Step 0 Let \mathcal{E}_0 be an initial encoder. Set $m = 0$. Fix $\lambda > 0$.

Step 1 Optimize the decoder \mathcal{D}_m for the encoder \mathcal{E}_m :

$$\mathcal{D}(i) = E(X | \mathcal{E}(X) = i) \text{ for all } i \in \mathcal{I}.$$

Step 2 Optimize w_m for \mathcal{E} : set $w_m(i) = \Pr(\mathcal{E}_m(X) = i)$.

Step 3 Optimize the encoder \mathcal{E}_{m+1} for the decoder \mathcal{D}_m and

w_m : for all x

$$\mathcal{E}_{m+1}(x) = \operatorname{argmin}_i \left(\|x - \mathcal{D}_m(i)\|^2 + \lambda \ln \frac{1}{w_m(i)} \right)$$

Step 4 Set $m + 1 \rightarrow m$ and go to Step 1.

Variation: initialize weighted codebook instead of partition

Lloyd quantizer improvement algorithm combined constraints

Step 0 Let \mathcal{E}_0 be an initial encoder. Set $m = 0$. Fix $\lambda > 0$.

Step 1 Optimize the decoder \mathcal{D}_m for the encoder \mathcal{E}_m :
 $\mathcal{D}(i) = E(X | \mathcal{E}(X) = i)$ for all $i \in \mathcal{I}$.

Step 2 Optimize w_m for \mathcal{E} : set $w_m(i) = \Pr(\mathcal{E}_m(X) = i)$.

Step 3 Optimize the encoder \mathcal{E}_{m+1} for the decoder \mathcal{D}_m and
 w_m :

$$\mathcal{E}_{m+1}(x) = \operatorname{argmin}_i \left(\|x - \mathcal{D}_m(i)\|^2 + \lambda(1 - \eta) \ln \frac{1}{w_m(i)} \right)$$

Step 4 Set $m + 1 \rightarrow m$ and go to Step 1.

As in the entropy constrained case, the algorithm can be initialized with a reproduction codebook and codebook weighting instead of a partition.

Subcodes and supercodes: pruning and growing codes

When can a code be improved by removing or adding codewords?

Might want better code for fixed λ , or good code at smaller λ (bigger codebook) or larger λ (smaller codebook)

Suppose have quantizer q satisfying Lloyd conditions. Described by either partition \mathcal{S} or by weighted decoder \mathcal{D}, w . Can form a *subcode* of q either by using subpartition of or a subset of the weighted codebook, similarly can form a *supercode* by using a superpartition of \mathcal{S} or a superset of the weighted codebook.

Obvious Lloyd condition: a necessary condition for q to be optimal is that no subcode or supercode yields better performance.

Suggests additional Lloyd algorithm step: Find computationally efficient means of testing for better sub or supercodes (fixed λ) or for finding codes that provide best tradeoff if changing λ (minimize increase of distortion per decrease in bits, or maximize decrease of distortion per increase in bits)

Partition subcodes and supercodes

A quantizer q determined by \mathcal{S} is a *partition subcode* of the quantizer q' determined by \mathcal{S}' if $\mathcal{S} \subset \mathcal{S}'$. (The Lloyd conditions then imply the corresponding decoder and weighting)

Similarly, q' is a *partition supercode* of q .

Lemma 4. *If $\mathcal{S} \subset \mathcal{S}'$, then*

$$D(\mathcal{S}) \geq D(\mathcal{S}')$$

$$H(\mathcal{S}) \leq H(\mathcal{S}')$$

$$N(\mathcal{S}) \leq N(\mathcal{S}')$$

$$R_\eta(\mathcal{S}) \leq R_\eta(\mathcal{S}')$$

where they appropriate optimal weighted codebook for each partition is assumed.

A partition subcode (supercode) will have larger (smaller) average distortion and larger (smaller) average rate, but the Lagrangian distortion might increase or decrease depending on the ratio of the change and λ .

Weighted codebook subcodes and supercodes

weighted codebooks (\mathcal{D}, w) (\mathcal{D}', w') with index sets $\mathcal{I} \subset \mathcal{I}'$

(\mathcal{D}, w) is a *codebook subcode* of (\mathcal{D}', w') ((\mathcal{D}', w') is a *codebook supercode* of (\mathcal{D}, w)) and write $(\mathcal{D}, w) \subset (\mathcal{D}', w')$ if

- the reproduction codebook of the subcode is a subset of the reproduction codebook of the larger code: $\{\mathcal{D}(i), i : w(i) > 0\} \subset \{\mathcal{D}'(i), i : w'(i) > 0\}$ and the codewords are ordered so that that $\mathcal{D}(i) = \mathcal{D}'(i)$ for all $i : w(i) > 0$.
- Define $J = \{i : w'(i) > 0, w(i) = 0\}$, the set of all indices corresponding to reproduction codewords removed from the larger code. Then $w'(i) = \alpha w(i)$ for all $i \notin J$ for some $\alpha \in (0, 1]$.

Note: Relative weights for common codewords unchanged.

Can use $\alpha = 1$ for pruning existing codebook, but could violate subpmf condition if want to grow codebook with w already a pmf.

Can find subcodebook by using list encoding.

Large literature exists for growing and pruning codes based on partitions – tree-structured vector quantization.

Relatively little done for growing and pruning codes based on weighted codebooks. Similar ideas arise in agglomerative and conglomerative clustering algorithms.

Weighted codebook sub and supercodes have similar behavior to partition sub and supercodes in terms of distortion and rate, but not quite the same:

$(\mathcal{D}, w), (\mathcal{D}', w'), \mathcal{S} = \{S_i\}$ and $\mathcal{S}' = \{S'_i\}$ corresponding Lloyd optimal encoder partitions.

$$\mathcal{I}' = \{i : w'(i) > 0\}, J = \{i : i \in \mathcal{I}', w(i) = 0\}$$

If $x \in S'_i$ then

$$d(x, \mathcal{D}'(i)) - \lambda(1 - \eta) \ln w'(i) \leq d(x, \mathcal{D}'(j)) - \lambda(1 - \eta) \ln w'(j), \text{ all } j \neq i$$

If $i, j \notin J$, then also

$$d(x, \mathcal{D}(i)) - \lambda(1 - \eta) \ln w(i) \leq d(x, \mathcal{D}(j)) - \lambda(1 - \eta) \ln w(j) \Rightarrow x \in S_i \Rightarrow S'_i \subset S_i$$

For each $j \in J$, define $S_{i,j} = S'_j \cap S_i$ (part of removed atom S'_j put into S_i)

$$S_i = S'_i \cup \bigcup_{j \in J} S_{i,j}; i \notin J$$

$$S'_j = \bigcup_{i \notin J} S_{i,j}; j \in J$$

$$S'_i = S_i - \bigcup_{j \in J} S_{i,j}; i \notin J$$

Let $p'(i) = P(S'_i)$, $p(i) = P(S_i)$ and observe that for $i \notin J$, $p(i) \geq p'(i)$.

Rate Obviously $N(w) \leq N(w')$. It also follows easily that

$$\begin{aligned} - \sum_i p'(i) \ln w'(i) &= - \sum_{i \notin J} p'(i) \ln w'(i) - \sum_{i \in J} p'(i) \ln w'(i) \\ &\geq - \sum_{i \notin J} p'(i) \ln w'(i) \\ &\geq - \sum_i p(i) \ln w(i). \end{aligned} \tag{28}$$

Thus,

$$R_\eta(\mathcal{E}', w') \geq R_\eta(\mathcal{E}, w), \tag{29}$$

as was the case with partition sub and supercodes.

Average distortion

$$\begin{aligned} D(\mathcal{E}, \mathcal{D}, w) &= \sum_{i \in \mathcal{I}' - J} \int_{S_i} d(x, \text{cent}(S_i)) dP(x) \\ &= \sum_{i \in \mathcal{I}' - J} \int_{S'_i \cup \bigcup_{j \in J} S_{i,j}} d(x, \text{cent}(S_i)) dP(x) \\ &= \sum_{i \in \mathcal{I}' - J} \int_{S'_i} d(x, \text{cent}(S_i)) dP(x) + \\ &\quad \sum_{i \in \mathcal{I}' - J} \int_{\bigcup_{j \in J} S_{i,j}} d(x, \text{cent}(S_i)) dP(x). \end{aligned}$$

Unfortunately this does not imply that $D(q) \geq D(q')$ in general. If, however $\alpha = 1$, then **the implication does follow:**

$$\min_i d(x, \mathcal{D}(i)) - \lambda(1 - \eta) \ln w(i) \geq \min_i d(x, \mathcal{D}'(i)) - \lambda(1 - \eta) \ln w'(i)$$

and hence

$$D(q) - \lambda(1 - \eta) \sum_i p(i) \ln w(i) \geq D(q') - \lambda(1 - \eta) \sum_i p'(i) \ln w'(i)$$

which with (28) implies that $D(q) \geq D(q')$. Thus

Lemma 5. *If q is a subcode of q' with $\alpha = 1$, then*

$$\begin{aligned} D(q) &\geq D(q') \\ R_\eta(q) &\leq R_\eta(q') \end{aligned}$$

that is, subcodes have smaller distortion and larger rate.

Codebook size If a quantizer q is optimal there can be no subcode or supercode q' for which

$$D(q') + \lambda R(q') < D(q) + \lambda R(q).$$

pruning/growing

Growing and pruning can be used either for improving a code at a fixed λ , or finding codes for smaller or larger λ that optimize the distortion/rate tradeoff.

Step 0: Initialization Given initial (\mathcal{D}_0, w_0) .

Compute $\rho_0 = \rho(\mathcal{S}(\mathcal{D}_0, w_0), \mathcal{D}_0, w_0)$. Set $m = 1$

Step 1: Partition improvement Given $(\mathcal{D}_{m-1}, w_{m-1})$, form an optimum partition $\mathcal{S}_m = \mathcal{S}(\mathcal{D}_{m-1}, w_{m-1})$.

Step 2: Weighted codebook improvement Given the partition \mathcal{S}_m , form an optimum weighted codebook $(\mathcal{D}_m, w_m) = (\mathcal{D}(\mathcal{S}_m), w(\mathcal{S}_m))$. Compute $\rho_m = \rho(\mathcal{S}_m, \mathcal{D}_m, w_m)$.

Step 3: Test Test $\rho_{m-1} - \rho_m$. If small enough, go to Step 4. Else set $m = m + 1$, go to Step 1.

Step 4: Grow/Prune Test sub/super codes for possible improvement for fixed λ or for changed λ . Quit or go to step 1.

Shannon rate-distortion theory

Shannon (1949, 1959)

Branch of information theory: source coding subject to a fidelity criterion

Information measures

So far:

X is a random vector with distribution P

α is a quantizer or a quantizer encoder

Then

$$H(\alpha(X)) \triangleq \sum_{i \in \mathcal{I}} P(\alpha(X) = i) \ln \frac{1}{P(\alpha(X) = i)}.$$

General definition of entropy for a random vector (Komogorov-Sinai):

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

supremum over all quantizers

If X continuous, $H(X) = \infty$ (except for trivial cases)

Given two distributions P and P' describing a random variable X and a quantizer α , the relative-entropy or Kullback Leibler divergence is

$$H_{P||P'}(\alpha(X)) \triangleq \sum_{i \in \mathcal{I}} P(\alpha(X) = i) \ln \frac{P(\alpha(X) = i)}{P'(\alpha(X) = i)}$$

Relative entropy of the random vector X

$$H(P||P') \triangleq \sup_{\alpha} H_{P||P'}(\alpha(X)).$$

If P and P' are discrete with pmf's f and g , then

$$H(P||P') = H(f||g) = \sum_i f_i \ln \frac{f_i}{g_i}.$$

If P and P' are determined by densities f and g , then

$$H(P||P') = H(f||g) = \int dx f(x) \ln \frac{f(x)}{g(x)}$$

Average mutual information between quantized random vectors X and Y :

$$I(\alpha(X), \beta(Y)) \triangleq H(\alpha(X)) + H(\beta(Y)) - H(\alpha(X), \beta(Y))$$

mutual information between the random vectors by

$$I(X, Y) \triangleq \sup_{\alpha, \beta} I(\alpha(X), \beta(Y)).$$

Alternatively,

$$I(\alpha(X), \beta(Y)) = H_{P_{\alpha(X), \beta(Y)} \| P_{\alpha(X)} \times P_{\beta(Y)}}(\alpha(X), \beta(Y)),$$

and

$$I(X, Y) = H_{P_{X, Y} \| P_X \times P_Y}.$$

It is straightforward to prove the following useful inequalities:

$$I(\alpha(X), \beta(Y)) \geq 0$$

$$H(\alpha(X)) \leq H(\alpha(X), \beta(X))$$

$$I(\alpha(X), \beta(Y)) \leq H(\beta(Y))$$

$$I(X, Y) \leq H(Y)$$

For example, the first two follow from the divergence inequality and the third follows from the second and the definition.

A lower bound to average distortion

X k -dimensional random vector of samples from a stationary source.

Quantize by a quantizer $q = (\mathcal{E}, \mathcal{D}, \ell)$.

Focus on distortion-rate formulation variation on Shannon's rate-distortion formulation

Shannon considered fixed-rate codes of vectors, or *block codes*

But lower bound works for all cases (all $\eta \in [0, 1]$)

Fix $\eta \in [0, 1]$, $R \geq 0$ and assume $R_\eta(\mathcal{E}, w) \leq R$

From the basic information measure inequalities and Corollary 1,

$$\begin{aligned}
 R_\eta(\mathcal{E}, w) &= (1 - \eta) \sum_i p(i) \ln \frac{1}{w_i} + \eta \ln N(w) \\
 &\geq (1 - \eta) H(\mathcal{E}(X)) + \eta \ln N(w) \\
 &\geq H(\mathcal{E}(X)) \geq H(\mathcal{D}(\mathcal{E}(X))) \\
 &\geq I(X; \mathcal{D}(\mathcal{E}(X))).
 \end{aligned}$$

Thus for any code $q = (\mathcal{E}, \mathcal{D}, p)$ with $R_\eta(\mathcal{E}, w) \leq R$,

$$\begin{aligned}
 D(q) &= E[d(X, \hat{X})] \\
 &\geq \inf_{f_{Y|X}: I(X; Y) \leq R} E[d(X, Y)] \triangleq D_k(R),
 \end{aligned}$$

Shannon's *distortion-rate function (DRF)* for $X = X^k$.

Since true for *all* quantizers, $\delta_\eta(R) \geq D_k(R)$ all $\eta \in [0, 1]$

Shannon DRF's are not in general easy to compute, but

- there is a further lower bound, the *Shannon lower bound*, which is easy to compute and provides a general and useful, if conservative, lower bound to average distortion.
- In some cases, such as Gaussian processes, the Shannon DRF can be explicitly evaluated
- Arimoto-Blahut algorithm for numerical evaluation

Lower bound is a “negative” or “converse” result – can do no better, there is also a positive theorem, but requires limits of large dimension k

Fix R and normalize rate and distortion by dimension. Define

$$\delta_{\eta}^{(k)}(R) = \frac{\delta_{\eta}(f_{X^k}, kR)}{k}$$

$$\bar{\delta}_{\eta}(R) = \inf_k \delta_{\eta}^{(k)}(R),$$

Then **Shannon distortion rate function** for the process $\{X_n\}$ is

$$\bar{\delta}_{\eta}(R) \geq \inf_k \frac{1}{k} D_k(kR) \triangleq \bar{D}(R).$$

For stationary processes, infima are limits as $k \rightarrow \infty$

Positive coding theorem

$$\bar{\delta}_{\eta}(R) = \bar{D}(R).$$

Much harder to prove. Existence proof, need asymptotically large dimension.

Result holds regardless of η and Shannon DRF has no η in its formulation. For large dimension, optimizing for fixed or variable rate codes makes no difference — both have the same limiting performance!

The Shannon lower bound

Shannon (1959)

For any conditional pdf $f_{Y|X}$ (“test channel”) define the Lagrangian average distortion

$$\begin{aligned}\rho(\lambda, f_{Y|X}) &= E[d(X, Y)] + \lambda I(X; Y) \\ &= \int dx \int dy f_{Y|X}(y|x) f_X(x) \times \\ &\quad \left(\|x - y\|^2 + \lambda \ln \frac{f_{Y|X}(y|x)}{\int du f_{Y|X}(y|u) f_X(u)} \right)\end{aligned}$$

Optimization problem is: given f_X , find

$$\rho(\lambda) = \inf_{f_{Y|X}} \rho(\lambda, f_{Y|X}),$$

Rewrite the Lagrangian

$$\begin{aligned}\rho(\lambda, f_{Y|X}) &= \lambda \int dy f_Y(y) \int dx f_{X|Y}(x|y) \left(\frac{\|x - y\|^2}{\lambda} + \ln \frac{f_{X|Y}(x|y)}{f_X(x)} \right) \\ &= \lambda h(X) + \lambda \int dy f_Y(y) \int dx f_{X|Y}(x|y) \times \\ &\quad \left[-\ln e^{-\frac{\|x-y\|^2}{\lambda}} + \ln f_{X|Y}(x|y) \right]\end{aligned}$$

where

$$h(X) = \int f_X(x) \ln \frac{1}{f_X(x)} dx \quad (30)$$

Shannon differential entropy.

Define

$$g_\lambda(x) \triangleq \frac{e^{-\frac{\|x\|^2}{2\sigma_g^2}}}{(2\pi\sigma_g^2)^{k/2}}$$

where $2\sigma_g^2 = \lambda$. Using the divergence inequality,

$$\begin{aligned} \rho(\lambda, f_{Y|X}) &= \lambda h(X) + \lambda \int dy f_Y(y) \int dx f_{X|Y}(x|y) \ln \frac{f_{X|Y}(x|y)}{g_\lambda(x-y)} - \frac{k\lambda}{2} \ln(2\pi\sigma_g^2) \\ &\geq \lambda h(X) - \frac{\lambda}{2} \ln(\pi\lambda)^k, \end{aligned}$$

achieved if $f_{X|Y}(x|y) = g_\lambda(x-y)$. The conditional density $f_{X|Y}$ is called the “backward channel distribution.”

Suppose $f_{Y|X}$ is an arbitrary pdf yielding $I \leq R$ and distortion D .
Then

$$D + \lambda I = \rho(\lambda, f_{Y|X}) \geq \lambda h(X) - \frac{\lambda}{2} \ln(\pi \lambda)^k \quad (31)$$

and hence

$$\begin{aligned} D &\geq \lambda h(X) - \frac{\lambda}{2} \ln(\pi \lambda)^k - \lambda I \\ &\geq \lambda h(X) - \frac{\lambda}{2} \ln(\pi \lambda)^k - \lambda R \\ &= \lambda \left[h(X) - R - \frac{1}{2} \ln(\pi \lambda)^k \right] \end{aligned}$$

Bound holds for any value of λ , so maximize over λ .

Globally optimal value follows from the the $\ln r \leq r - 1$ inequality:

$$\begin{aligned} \lambda \frac{k}{2} \ln \frac{e^{\frac{2}{k}(h(X)-R)}}{\pi \lambda} &= \lambda \frac{k}{2} \left(\ln \frac{e^{\frac{2}{k}(h(X)-R)}}{\pi \lambda e} + 1 \right) \\ &\leq \lambda \frac{k e^{\frac{2}{k}(h(X)-R)}}{2 \pi \lambda e} = \frac{k e^{\frac{2}{k}(h(X)-R)}}{2 \pi e} \end{aligned}$$

with equality iff

$$\lambda = \frac{e^{\frac{2}{k}(h(X)-R)}}{\pi e} \quad \text{erroneous } \lambda \text{ removed}$$

which yields the lower bound

$$D(R) \geq \frac{k}{2} \lambda = \frac{k}{2 \pi e} e^{-\frac{2}{k}(R-h(X))} \triangleq D_{\text{SLB}}(R) \quad (32)$$

Dual argument yields the Shannon lower bound to the RDF (what Shannon derived)

$$R(D) \geq h(X) - \frac{k}{2} \ln\left(\frac{2\pi e D}{k}\right) \triangleq R_{\text{SLB}}(D)$$

Recall the lower bound holds with equality iff $f_{X|Y}(x|y) = g_\lambda(x-y)$.

This will be true if one can find an f_Y for which

$$f_X(x) = \int g_\lambda(x-y) f_Y(y) dy,$$

which is not always possible.

Example where it is possible: 1 dimensional case with $f_X = \mathcal{N}(0, \sigma^2)$. Choosing $f_Y = \mathcal{N}(0, \sigma^2 - D)$ yields the Shannon lower bound with equality provided $R > 0$. $f_{X|Y}(x|y) = g_\lambda(x - y)$. Recall $\sigma_g^2 = \lambda/2$ and for equality in bound $\lambda = e^{2(h(X)-R)}/\pi e = 2D \Rightarrow \sigma_g^2 = D$. So X is $N(0, \sigma^2 - D + D = \sigma^2)$.

$$h(X) = \frac{1}{2} \ln(2\pi e\sigma^2)$$

$$R(D) = \frac{1}{2} \ln(2\pi e\sigma^2) - \frac{1}{2} \ln(2\pi eD) = \frac{1}{2} \ln \frac{\sigma^2}{D}; \quad D \in [0, \sigma^2]$$

and

$$D(R) = \sigma^2 2^{-2R}; \quad R \geq 0$$

It is known that the Shannon lower bound is tight as $R \rightarrow \infty$, that is, as $R \rightarrow \infty$, $D(R) - D_{\text{SLB}}(R) \rightarrow 0$.

High-rate quantization theory

Bennett's approximations

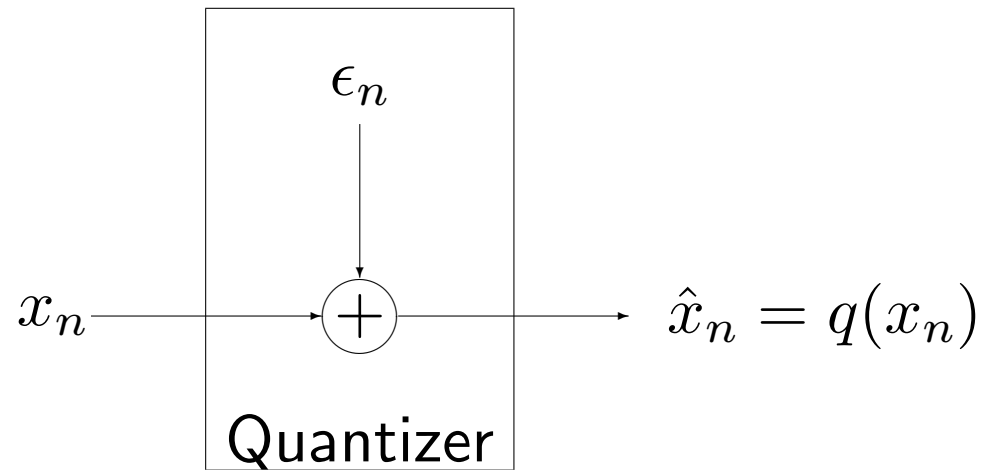
High rate, large N , large H , small distortion, fixed dimension

Suppose q is a simple scalar uniform quantizer with bin width Δ and N levels as in (10).

Define quantizer error $\epsilon = \epsilon(x) = q(x) - x$

Apply q to sequence X_n , $\epsilon_n = q(X_n) - X_n$

Sign chosen for “additive noise model” representation



“Model” because common in the literature to make assumptions on ϵ_n as behaving like signal-independent additive noise.

Consider system when Δ becomes small and M becomes large under the assumption that the input probability density function f_X is smooth.

First consider marginal cdf $F_{\epsilon_n}(\alpha) = \Pr(\epsilon_n \leq \alpha)$ and pdf $f_{\epsilon_n}(\alpha) = dF_{\epsilon_n}(\alpha)/d\alpha$, $\alpha \in (-\Delta/2, \Delta/2)$

$$\begin{aligned} \Pr(\epsilon_n \leq \alpha) &= \sum_{k=0}^{N-1} \Pr(\epsilon_n \leq \alpha \text{ and } X_n \in S_k) \\ &= \sum_{k=0}^{N-1} \Pr(a + (k + \frac{1}{2})\Delta - X \leq \alpha \text{ and } X \in [a + k\Delta, a + (k + 1)\Delta]) \end{aligned}$$

Since the pdf smooth, mean value theorem \Rightarrow

$$\begin{aligned} \Pr(\epsilon_n \leq \alpha \text{ and } X_n \in S_k) &= \int_{a+(k+\frac{1}{2})\Delta-\alpha}^{a+(k+1)\Delta} f_{X_n}(\beta) d\beta \\ &\approx f_{X_n}(\mathcal{D}(k))(\alpha + \frac{\Delta}{2}) \end{aligned}$$

so that

$$\begin{aligned}\Pr(\epsilon_n \leq \alpha) &\approx \left(\frac{\alpha}{\Delta} + \frac{1}{2}\right) \sum_{k=0}^{N-1} f_{X_n}(\mathcal{D}(k)) \Delta \\ &\approx \left(\frac{\alpha}{\Delta} + \frac{1}{2}\right) \int f_{X_n}(x) dx \approx \left(\frac{\alpha}{\Delta} + \frac{1}{2}\right), \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)\end{aligned}$$

Riemann sum approximation to integral. \Rightarrow

$$f_{\epsilon_n}(\alpha) \approx \frac{1}{\Delta} \text{ for } \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right),$$

consistent with the assumed behavior of the “additive noise” model.

In the high rate regime, average distortion of a uniform quantizer $\approx \Delta^2/12$ that predicted by adding uniform noise. Origin of “6dB per bit” improvement in SNR of a quantizer with a high bit rate.

Look at vectors $(\epsilon_n, \dots, \epsilon_{n+k-1})$:

$$\Pr(\epsilon_{n+m} \leq \alpha_m; m = 0, \dots, k-1) = \sum_{i_0, \dots, i_{k-1}} \Pr(\epsilon_{n+m} \leq \alpha_m \text{ and } X_{n+m} \in S_{i_m}; m = 0, \dots, k-1),$$

For $\alpha_m \in (-\Delta/2, \Delta/2)$, $m = 0, \dots, k-1$

$$\begin{aligned} \Pr(\epsilon_{n+m} \leq \alpha_m \text{ and } X_{n+m} \in S_{i_m}; m = 0, \dots, k-1) &= \\ \int_{a+(k+\frac{1}{2})\Delta-\alpha_0}^{a+(k+1)\Delta} \cdots \int_{a+(k+\frac{1}{2})\Delta-\alpha_{k-1}}^{a+(k+1)\Delta} f_{X_n, \dots, X_{n+k-1}}(\beta_0, \dots, \beta_{k-1}) d\beta_0 \cdots d\beta_{k-1} &= \\ \approx f_{X_n, \dots, X_{n+k-1}}(\mathcal{D}(i_0), \dots, \mathcal{D}(i_{k-1}))(\alpha_0 + \frac{\Delta}{2}) \cdots (\alpha_{k-1} + \frac{\Delta}{2}) & \quad (33) \end{aligned}$$

whence

$$f_{\epsilon_n, \dots, \epsilon_{n+k-1}}(\alpha_0, \dots, \alpha_{k-1}) \approx \frac{1}{\Delta^k}$$

Thus in the high rate regime, the quantizer errors \approx independent, hence white, and uniform!

But, requires very large N and very small cell size.

Eq (33) has an even stronger implication:

$$\begin{aligned}
\Pr(\epsilon_m \leq \alpha_m; m = 0, \dots, k-1 | X_m \in S_{i_m}; m = 0, \dots, k-1) &= \\
&= \frac{\Pr(\epsilon_m \leq \alpha_m \text{ and } X_m \in S_{i_m}; m = 0, \dots, k-1)}{\Pr(X_m \in S_{i_m}; m = 0, \dots, k-1)} \approx \\
&= \frac{f_{X_n, \dots, X_{n+k-1}}(\mathcal{D}(i_0), \dots, \mathcal{D}(i_{k-1})) (\alpha_0 + \Delta/2) \cdots (\alpha_{k-1} + \Delta/2)}{f_{X_n, \dots, X_{n+k-1}}(\mathcal{D}(i_0), \dots, \mathcal{D}(i_{k-1})) \Delta^k} \\
&= \frac{(\alpha_0 + \Delta/2) \cdots (\alpha_{k-1} + \Delta/2)}{\Delta^k}
\end{aligned}$$

$$\Rightarrow f_{\epsilon_n, \dots, \epsilon_{n+k-1} | X_n \in S_{i_0}, \dots, X_{n+k-1} \in S_{i_{k-1}}}(\alpha_0 \cdots \alpha_{k-1}) = 1/\Delta^k$$

so that the errors are uniform and independent conditioned on each specific k -dimensional cell!.

These approximations have several implications:

- $E[\epsilon_n] \approx 0, \sigma_{\epsilon_n}^2 \approx \Delta^2/12.$
- $R_\epsilon(n, m) \approx E[\epsilon_n \epsilon_m] \approx \sigma_\epsilon^2 \delta_{n-m}$
- For $i = n, \dots, n + k - 1$

$$f_{\epsilon_i | X_m \in S_{i_m}; m=0, \dots, k-1}(\alpha) \approx \frac{1}{\Delta} \text{ for } \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$$

so that the marginal error is uniform conditioned on k quantizer outputs including the same index. This leads to

- $f_{\epsilon_m | X_n \in S_\ell}(\alpha) \approx \frac{1}{\Delta} \text{ for } \alpha \in \left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right) \Rightarrow E[\epsilon_m | q(X_n) = \mathcal{D}(\ell)] \approx 0$

- Since $q(X_n) = \mathcal{D}(l)$ iff $X_n \in S_l$,

$$\begin{aligned} R_{q,\epsilon}(n, m) &= E[q(X_n)\epsilon_m] \\ &= \sum_{l=0}^{M-1} \mathcal{D}(l) E[\epsilon_m | q(X_n) = \mathcal{D}(l)] \Pr(q(X_n) = \mathcal{D}(l)) = 0 \end{aligned}$$

so that the quantizer output is uncorrelated with the input.

- The previous formula implies that

$$\begin{aligned} R_{X,\epsilon}(n, m) &= E[X_n\epsilon_m] \\ &= E[(q(X_n) - \epsilon_n)\epsilon_m] \\ &= E[q(X_n)\epsilon_m] - R_\epsilon(n, m) \\ &= -R_\epsilon(n, m), \end{aligned}$$

i.e., that

$$R_{X,\epsilon}(n, m) = -\sigma_\epsilon^2 \delta_{n-m}.$$

input and quantizer error are *not uncorrelated!*

Hence common assumption of independence between the quantizer error and the input will yield incorrect results.

Bennett extended his approximations to nonuniform scalar quantizers and used his approximations to quantify the optimal performance in fixed-rate systems. His methods, however, do not extend to the multidimensional case so we shall proceed to more general results for vectors.

High-rate theory for vector quantizers

Zador (1963, 1966, 1982)

Bucklew, Wise, Gersho

Assume density f absolutely continuous and satisfies moment condition

$$E_f(\|X\|^{2+\delta}) < \infty \quad (34)$$

for some $\delta > 0$

Fixed rate:

$$\lim_{N \rightarrow \infty} N^{\frac{2}{k}} \delta_1(f, \ln N) = a_k \|f\|_{\frac{k}{k+2}}, \quad (35)$$

where

$$a_k \triangleq \inf_{N \geq 1} N^{\frac{2}{k}} \delta_1(u, \ln N),$$

u is the uniform pdf on the unit cube $[0, 1]^k$, and $\|f\|_p = (\int f^p)^{1/p}$.

Variable rate:

$$\lim_{R \rightarrow \infty} e^{\frac{2}{k}R} \delta_0(f, R) = b_k e^{\frac{2}{k}h(f)} \quad (36)$$

where $h(f) = -\int f \ln f$ is the differential entropy and the positive constant b_k is given by

$$b_k = \inf_{R > 0} R^{\frac{2}{k}} \delta_0(u, R).$$

Intuitively, for very large R

$$\delta_1(f, R) \approx a_k N^{-2/k} \|f\|_{\frac{k}{k+2}} \text{ fixed-rate}$$

$$\delta_0(f, R) \approx b_k e^{-\frac{2}{k}(R-h(X))} \text{ variable-rate}$$

Since fixed rate case has a stronger constraint,

$$\delta_1(f, R) \geq \delta_0(f, R).$$

The lower bound to $\delta_0(f, R)$ strongly resembles the Shannon lower bound of (32), difference is constant b_k in the Zador approximation and the constant $k/2\pi e$ in the Shannon bound. Can show

$$a_k \geq b_k \geq \frac{k}{2\pi e}$$

and

$$\lim_{k \rightarrow \infty} \frac{a_k}{k} = \lim_{k \rightarrow \infty} \frac{b_k}{k} = \frac{1}{2\pi e}$$

so the Shannon and Zador results are consistent and agree in the limit of large dimension.

The first rigorous proof of (36) used the Lagrangian approach and results are stated in terms of the Lagrangian distortion:

$$\rho(f, \lambda, \eta) = \inf_q \rho(f, \lambda, \eta, q) \quad (37)$$

$$\rho(f, \lambda, \eta, q) = D(q) + \lambda R_\eta(q) \quad (38)$$

where for the moment $\eta = 0$.

Required finite $h(f)$ and required that a uniform scalar quantized version of X with cubic cell volume 1 have finite entropy

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda, 0)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \theta_k + h(f) \quad (39)$$

where

$$\theta_k \triangleq \inf_{\lambda > 0} \left(\frac{\rho(u, \lambda, 0)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \frac{k}{2} \ln \frac{2eb_k}{k}. \quad (40)$$

Two forms (traditional and Lagrangian) known to be equivalent, i.e., each holds iff the other does.

Similar arguments show that in the fixed-rate case, (35) holds for f iff

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda, 1)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \psi_k + \ln \|f\|_{k/(k+2)}^{k/2} \quad (41)$$

where

$$\psi_k = (k/2) \ln(2ea_k/k). \quad (42)$$

Same form as the variable-rate Zador result with $\ln \|f\|_{k/(k+2)}^{k/2}$ replacing $h(f)$ and ψ_k replacing θ_k .

More generally, recently shown that Lagrangian and traditional formulations equivalent for general combined constraint case:

$$\lim_{R \rightarrow 0} e^{\frac{2}{k}R} \delta_\eta(f, R) = \delta(f, \eta) \quad (43)$$

exists for a positive finite $\delta(f, \eta)$, then

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda, \eta)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \frac{k}{2} \ln \left(\frac{2e}{k} \delta(f, \eta) \right).$$

Conversely, if

$$\lim_{\lambda \rightarrow 0} \left(\frac{\rho(f, \lambda, \eta)}{\lambda} + \frac{k}{2} \ln \lambda \right) = \theta(f, \eta) \quad (44)$$

for a finite $\theta(f, \eta)$, then

$$\lim_{R \rightarrow 0} e^{\frac{2}{k}R} \delta_\eta(f, R) = \frac{k}{2} e^{(2/k)\theta(f, \eta) - 1}.$$

$$\theta(f, \eta) = \frac{k}{2} \ln \left(\frac{2e}{k} \delta(f, \eta) \right)$$

includes (40) and (42) as special cases. The Lagrangian form has the the advantage of yielding a Lloyd algorithm for design and results in simpler proofs of some results.

The details of the proofs of the high-rate results for the traditional cases differ significantly, but most proofs of these results follow the original Zador approach.

1. Prove the result for u , the uniform pdf on the unit cube.
2. Extend the result to pdfs that are piecewise constant on disjoint cubes of equal side a .
3. Prove the result for a general pdf on a cube.
4. Prove the result for general pdfs.

The first step is a key one both because it provides the primary building block for the subsequent results, and because it suffices to study Zador's constants.

The high rate results have been proved under general assumptions only for the traditional cases $\eta = 0, 1$. The general form of the results for the combined constraint case of general $\eta \in [0, 1]$ has been conjectured based on the arguments described next, and proved for the special, but important, case of the uniform density u .

Gersho's conjectures and approximations

Gersho popularized Zador's results, used conjectures and heuristic arguments to derive basic results.

Gersho's conjecture involves two assumptions regarding asymptotically optimal sequences of fixed-rate and variable-rate quantizers q_n :

1. There exists a quantizer point density function $\Lambda(x)$ (mathematically, a pdf) such that a sequence of optimal codes with N codewords, $N = 1, 2, \dots$ will satisfy for all "reasonable" $S \subset \mathbb{R}^k$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \times (\# \text{ of reproduction vectors in a set } S) = \int_S \Lambda(x) dx.$$

2. All encoder partition cells asymptotically have same shape, that of tessellating convex polytope with minimum normalized moment of inertia (can be stretched, rotated, shifted):

$$M(S) = \frac{1}{kV(S)^{1+2/k}} \int_S \|x - y(S)\|^2 dx$$

where $y(S)$ = centroid of S with respect to the uniform distribution on S ,

$$c_k = \min_{\text{tessellating convex polytopes } S} M(S)$$

Under these assumptions, Gersho argued that for large N and small cells

$$D_f(q) \approx c_k E_f \left(\left(\frac{1}{N(q)\Lambda(X)} \right)^{2/k} \right) \quad (45)$$

$$\begin{aligned} H_f(q(X)) &\approx h(X) - E_f \left(\ln \left(\frac{1}{N(q)\Lambda(X)} \right) \right) \\ &= \ln N(q) - H(f||\Lambda). \end{aligned} \quad (46)$$

Relates asymptotic average distortion, entropy, and codebook size.

Sketch Gersho's development of these approximations: assume that the input random vector X is described by a "smooth" pdf f

Define volume $V(S)$ of a set S in \mathfrak{R}^k by

$$V(S) = \int_S dx$$

and the diameter of a set by

$$\text{diam}(S) = \sup_{a,b \in S} \|a - b\|^2.$$

Let q be a quantizer with a large finite number N of partition cells S_i .

“high rate” or “high resolution” assumption requires that the average distortion \approx determined by cells with small diameter and volume, that large cells all together contribute a negligible amount

Thus

$$D(q) \approx \sum_{i: S_i \text{ small}} \int_{S_i} \|x - \mathcal{D}(i)\|^2 f(x) dx.$$

Practical interpretation: “no-overload” region

$$D(q) \approx \sum_{i=1}^N P_X(S_i) \int_{S_i} \|x - \mathcal{D}(i)\|^2 f(x) dx,$$

let $\mathcal{D}(i)$ be the Lloyd centroid of the cell S_i . Since f is assumed smooth and the cells are small,

$$f(x) \approx f(\mathcal{D}(i)); x \in S_i$$

⇒ from the mean value theorem of calculus,

$$P_X(S_i) = \int_{S_i} f(x) dx \approx V(S_i) f(\mathcal{D}(i))$$

hence

$$f(\mathcal{D}(i)) \approx \frac{P_X(S_i)}{V(S_i)}$$

⇒

$$D(q) \approx \sum_{i=1}^N P_X(S_i) \int_{S_i} \frac{\|x - \mathcal{D}(i)\|^2}{V(S_i)} dx$$

Again assuming that the cells are small and the pdf smooth, the centroid of S_i with respect to original density \approx the centroid with respect to a uniform pdf.

Since $\mathcal{D}(i)$ is the centroid of S_i and hence

$$\int_{S_i} \frac{\|x - \mathcal{D}(i)\|^2}{V(S_i)} dx$$

yields the minimum possible squared error over y of $\int_{S_i} \frac{\|x-y\|^2}{V(S_i)} dx$, moment of inertia of the region S_i about its centroid if the total mass is 1 and the mass density is uniform.

Convenient to use *normalized* moments of inertia (invariant to scale)

$$M(S) = \frac{1}{V(S)^{2/k}} \int_S \frac{\|x - y(S)\|^2}{V(S)} dx$$

where $y(S)$ denotes the Euclidean centroid of S .

Normalization makes $M(S)$ invariant to scaling: $c > 0$ and $cS = \{cx : s \in S\}$, then

$$M(S) = M(cS)$$

so M depends only on shape and not upon scale.

Proof:

$$\begin{aligned} M(cS) &= \frac{\int_{cS} \|x\|^2 \frac{dx}{V(cS)}}{V(cS)^{2/k}} \\ &= \frac{\int_S \|cx\|^2 \frac{dcx}{c^2 V(S)}}{[c^k V(S)]^{2/k}} \end{aligned}$$

Now have

$$\begin{aligned} D(q) &\approx \sum_{i=1}^N P_X(S_i) M(S_i) V(S_i)^{2/k} \\ &= \sum_{i=1}^N f(\mathcal{D}(i)) M(S_i) V(S_i)^{1+2/k} \end{aligned}$$

Moment of inertia examples

S is a cube in \mathbb{R}^k with side length Δ , then

$$M(S) = \frac{k \frac{\Delta^2}{12}}{k (\Delta^k)^{2/k}} = \frac{1}{12}$$

Partition cell shape if a vector quantizer were formed as a combination of k identical uniform scalar quantizers on each component with a common bin width Δ .

If S a regular hexagon in \mathbb{R}^2 , then

$$M(S) = \frac{5}{36\sqrt{3}}.$$

Now invoke Gersho's conjecture on the existence of an asymptotically optimal quantizer point density function $\Lambda(x)$.

Assume N is sufficiently large and the S_i sufficiently small to ensure that each S_i contains only a single quantizer reproduction codeword

Mean value theorem \Rightarrow

$$\frac{1}{N} \approx \int_{S_i} \Lambda(x) dx \approx V(S_i) \Lambda(\mathcal{D}(i))$$

or

$$V(S_i) \approx \frac{1}{N \Lambda(\mathcal{D}(i))}.$$

Assume also Gersho's conjecture regarding optimal cell shapes: all quantization cells are convex polytopes which are scaled or rotated versions of a single polytope S^* and hence have equal normalized moments $M(S^*) \Rightarrow$

$$\begin{aligned}
D(q) &\approx \sum_{i=1}^N f(\mathcal{D}(i)) M(S_i) V(S_i)^{1+2/k} \\
&= \sum_{i=1}^N f(\mathcal{D}(i)) M(S_i) V(S_i)^{1+2/k} \\
&\approx M(S^*) N^{-2/k} \sum_{i=1}^N f(\mathcal{D}(i)) \frac{V(S_i)}{\Lambda(\mathcal{D}(i))^{2/k}} \\
&\approx N^{-2/k} M(S^*) \int f(x) \frac{1}{\Lambda(x)^{2/k}} dx \\
&= M(S^*) N^{-2/k} E\left[\frac{1}{\Lambda(X)^{2/k}}\right] = c_k N^{-2/k} E\left[\frac{1}{\Lambda(X)^{2/k}}\right]
\end{aligned}$$

Tessellating convex polytopes

Intervals tessellate \mathfrak{R}^1 and

$$c_1 = \frac{1}{12} = 0.08333 \dots$$

It is known that $a_1 = b_1 = c_1$.

For $k = 2$ Fejes Toth (1959) showed that the regular hexagon is optimal and

$$c_2 = \frac{5}{36\sqrt{3}} = 0.08019 \dots$$

It is known that $a_2 = c_2$, but b_2 is not known.

For three dimensions the optimal tessellating polytope is not known, but candidates are the hexagonal prism, the rhombic dodecahedron, the elongated dodecahedron, and the regular truncated octohedron. It is generally thought that that c_3 is $M(S)$ for the regular truncated octohedron, $\frac{19}{192 \times 2^{1/3}} = 0.07855 \dots$

For general k , the following lower bound is known:

$$c_k \geq M(\text{ sphere }) = \frac{\int_S \|x\|^2 dx}{V(S)^{1+2/k}} = \frac{kV_k^{-2/k}}{k+2}$$

where V_k is the volume of a sphere in k dimensions with unit radius:

$$V_k = \frac{\pi^{k/2}}{\Gamma(\frac{k}{2} + 1)} = \frac{2\pi^{k/2}}{k\Gamma(\frac{k}{2})}$$

where

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$

e.g., $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(n+1) = n\Gamma(n)$. For example, if $k = 3$,

$$C_k \geq \frac{kV_k^{-2/k}}{k+2} = 0.07697 \dots$$

Can show that $M(S)$ for a sphere of k dimensions decreases to $\frac{1}{2\pi e} = 0.05955 \dots$ (compare with the Shannon lower bound).

Gersho's entropy approximation

Again make the approximation that

$$P_X(S_i) \approx \frac{f(\mathcal{D}(i))}{N\Lambda(\mathcal{D}(i))} \approx f(\mathcal{D}(i))V(S_i)$$

so that

$$\begin{aligned}
H(\mathcal{E}(X)) &= - \sum_{i=1}^N P_X(S_i) \ln P_X(S_i) \\
&= - \sum_{i=1}^N \frac{f(\mathcal{D}(i))}{N\Lambda(\mathcal{D}(i))} \log \frac{f(\mathcal{D}(i))}{N\Lambda(\mathcal{D}(i))} \\
&= - \sum_{i=1}^N V(S_i) f(\mathcal{D}(i)) \ln f(\mathcal{D}(i)) + \\
&\quad \sum_{i=1}^N V(S_i) f(\mathcal{D}(i)) \ln(N\Lambda(\mathcal{D}(i))) \\
&\approx - \int dy f(y) \ln f(y) + \int dy f(y) \ln(N\Lambda(y)) \\
&\approx h(X) - E\left(\ln \frac{1}{N\Lambda(X)}\right) = \ln N - H(f\|\Lambda) \leq \ln N,
\end{aligned}$$

Approximation relates the Shannon differential entropy to the Shannon entropy of a quantizer output.

In the special case where $\Lambda(x) = 1/V(A); x \in A$ this becomes

$$H(\mathcal{E}(X)) \approx h(X) + \ln \frac{N}{V(A)},$$

simple approximation for the entropy for uniform quantizers (uniform in one dimension or lattice quantizers in higher dimensions).

Gersho's approximations: general case

Fix $\eta \in [0, 1]$

Assume quantizer q has a quantizer point density Λ and a large number N quantization levels

Then (using the $\ln r \leq r - 1$ inequality in the final step)

$$\begin{aligned}
& \theta(f, \lambda, \eta, q) \\
& \stackrel{\Delta}{=} \frac{D_f(q)}{\lambda} + (1 - \eta)H_f(q) + \eta \ln N + \frac{k}{2} \ln \lambda \\
& \approx \frac{c_k E_f \left((N \Lambda(X))^{-2/k} \right)}{\lambda} + (1 - \eta)[\ln N - H(f \parallel \Lambda)] + \eta \ln N + \frac{k}{2} \ln \lambda \\
& = \frac{k}{2} \left[\frac{2c_k N^{-2/k} E_f \left((\Lambda(X))^{-2/k} \right)}{\lambda} - \ln \frac{2c_k N^{-2/k} E_f \left((\Lambda(X))^{-2/k} \right)}{\lambda} \right. \\
& \quad \left. - 1 \right] + \frac{k}{2} \ln \left[\frac{2ec_k}{k} E_f \left((\Lambda(X))^{-2/k} \right) \right] - (1 - \eta)H(f \parallel \Lambda) \\
& \geq \frac{k}{2} \ln \left[\frac{2ec_k}{k} E_f \left((\Lambda(X))^{-2/k} \right) \right] - (1 - \eta)H(f \parallel \Lambda)
\end{aligned}$$

with equality iff

$$N = \left[\frac{2c_k E_f ((\Lambda(X))^{-2/k})}{k \lambda} \right]^{k/2} \quad (47)$$

Since goal is to minimize $\theta(f, \lambda, \eta, q)$, this is the optimal choice of N .

Thus for small λ

$$\theta(f, \lambda, \eta, q) \approx \frac{k}{2} \ln \left(\frac{2ec_k}{k} \right) + \phi(f, \eta, \Lambda) \quad (48)$$

where

$$\begin{aligned}
 \phi(f, \eta, \Lambda) &= \frac{k}{2} \ln \left(E_f \left((\Lambda(X))^{-2/k} \right) \right) - (1 - \eta) H(f \| \Lambda) \\
 &= \frac{k}{2} \ln \int f(x) \Lambda(x)^{-2/k} dx + (1 - \eta) \int f(x) \ln \Lambda(x) dx + (1 - \eta) h(f).
 \end{aligned} \tag{49}$$

Best possible performance will be that which minimizes $\phi(f, \lambda, \eta, q)$ over all q . If

$$\phi(f, \eta) = \inf_{\Lambda} \phi(f, \eta, \Lambda) \tag{50}$$

and the infimum is over all pdfs Λ for which $\phi(f, \eta, \Lambda)$ is well-defined, then

$$\lim_{\lambda \rightarrow 0} \theta(f, \lambda, \eta) = (k/2) \ln(2ec_k/k) + \phi(f, \eta). \tag{51}$$

Functionals $\phi(f, \eta, \Lambda)$ and $\phi(f, \eta)$ of (49)–(50) called Gersho functionals.

The functional $\phi(f, \eta, \Lambda)$ can be expressed as

$$\begin{aligned}\phi(f, \eta, \Lambda) &= (1 - \eta)\phi(f, 0, \Lambda) + \eta\phi(f, 1, \Lambda) \\ &= \phi(f, 0, \Lambda) + \eta(\phi(f, 1, \Lambda) - \phi(f, 0, \Lambda)) \\ &= \phi(f, 0, \Lambda) + \eta H(f \parallel \Lambda) \\ &= \phi(f, 1, \Lambda) - (1 - \eta)H(f \parallel \Lambda).\end{aligned}\tag{52}$$

The nonnegativity of the relative entropy $H(f \parallel \Lambda)$ implies immediately that

$$\phi(f, 0, \Lambda) \leq \phi(f, \eta, \Lambda) \leq \phi(f, 1, \Lambda).\tag{53}$$

If the derived approximations are valid, then (47) implies

$$\ln N \approx \frac{k}{2} \ln \frac{2ec_k}{k} + \phi(f, 1, \Lambda) + \ln \lambda^{-k/2}. \quad (54)$$

No explicit dependence on η here! The dependence is implicit through the selection of a Λ minimizing $\phi(f, \eta, \Lambda)$.

Optimization for traditional cases

If $\eta = 1$, Holder's inequality yields the bound

$$\phi(f, 1, \Lambda) = \frac{k}{2} \ln \left(\int f(x) \Lambda(x)^{-2/k} dx \right) \geq \frac{k}{2} \ln \|f\|_{k/(k+2)} = \phi(f, 1) \quad (55)$$

with equality iff

$$\Lambda(x) = \frac{f(x)^{k/(k+2)}}{\|f\|_{k/(k+2)}^{k/(k+2)}} \quad (56)$$

the well-known solution for the fixed-rate case.

The moment condition (34) ensures that $\|f\|_{k/(k+2)}$ is finite and the Λ minimizing $\phi(f, 1, \Lambda)$ is given by (56).

If $\eta = 0$, then from Jensen's inequality

$$\begin{aligned} \phi(f, 0, \Lambda) &= \frac{k}{2} \ln \left(\int f(x) \Lambda(x)^{-2/k} dx \right) - H(f \parallel \Lambda) \\ &\geq \frac{k}{2} \int f(x) \ln \left(\Lambda(x)^{-2/k} \right) dx - H(f \parallel \Lambda) \end{aligned} \quad (57)$$

$$= h(f) = \phi(f, 0) \quad (58)$$

with equality iff $\Lambda(x)$ is constant for the support set of X , again agreeing with the classical result.

(Here equality requires that the distribution of X has bounded support.)

Inequality (58) along with (53) imply the bounds

$$h(f) \leq \phi(f, 0, \Lambda) \leq \phi(f, \eta, \Lambda) \leq \phi(f, 1, \Lambda).$$

Thus Gersho's approach gives heuristic developments of both the classical cases of fixed-rate and variable-rate asymptotically optimal quantizers.

General case

Open

The general minimization of $\phi(f, \eta, \Lambda)$ for $\eta \in (0, 1)$ does not seem to have such a nice form.

Can show

$$\phi(f, \eta, \Lambda) \geq (1-\eta)\phi(f, 0, \Lambda) + \eta\phi(f, 1, \Lambda) = (1-\eta)h(f) + \eta \ln \|f\|_{k/(k+2)}^{\frac{2}{k}}$$

but the inequality is strict except for the endpoints since in general distinct Λ yield those minima.

The bound does hold for u , in which case Λ uniform on the unit

cube yields $\phi(u, \eta) = 0$. In this case (51) implies that

$$\lim_{\lambda \rightarrow 0} \theta(u, \lambda, \eta) = \frac{k}{2} \ln \left(\frac{2ec_k}{k} \right) \quad (59)$$

and hence c_k characterizes the performance on the unit distribution on the unit cube for all $\eta \in [0, 1]$ if Gersho's assumptions are true.

Unfortunately $\phi(f, \eta, \Lambda)$ not convex in Λ , but the following lemma shows that a transformation yields an equivalent convex optimization problem.

Lemma 6.

$$\phi(f, \eta) = \inf_{\nu} \psi(f, \eta, \nu) = \psi(f, \eta)$$

where $\phi(f, \eta)$ is given by (49)–(50), where

$$\begin{aligned} \psi(f, \eta, \nu) = \frac{k}{2} \left(\int f(x) e^{\nu(x)} dx - (1 - \eta) \int f(x) \nu(x) dx - 1 \right) \\ + \eta \ln \left(\int e^{-k\nu(x)/2} dx \right) + (1 - \eta)h(f) \quad (60) \end{aligned}$$

where the integrals are over the support set of f , where the infimum over ν is over all measurable functions ν for which $\psi(f, \eta, \nu)$ is well-defined. The functional $\psi(f, \eta, \nu)$ is (strictly) convex in ν .

There are proofs based on the heuristic approach, but they require the *assumption* that the asymptotic point density and an asymptotic normalized moment of inertia exist.

The existence of point density functions has been rigorously proved only for the fixed-rate ($\eta = 1$) case. [Bucklew (1984), Graf and Luschgy (2000)]

The existence of the density for the variable-rate case has not been similarly proved, although Gersho's heuristic arguments suggest that it is uniform (and this is often assumed).

Moving on

Have surveyed fundamental results in quantization theory and algorithms for squared error distortion.

Remainder of course will go more deeply into several of the topics.

Examples of issues to be considered include the following.

- Lossless coding and rate
- General distortion measures
- Shannon distortion-rate theory
- Lloyd improvement algorithms: structured codes (tricks often included in real-world data compression and classification systems)

- Examples
- Miscellaneous topics, conjectures, open problems