Big picture so far:

1. Lossless data compression: Given a discrete ergodic source S^k, we know how to encode to pure bits $W \in [2^k]$.
2. Binary HT: Given two distribution P and Q, we know how to distinguish them optimally.
3. Channel coding: How to send bits over a channel $[2^k] \ni W \to X \to Y$.
4. JSCC: how to send discrete data optimally over a noisy channel.

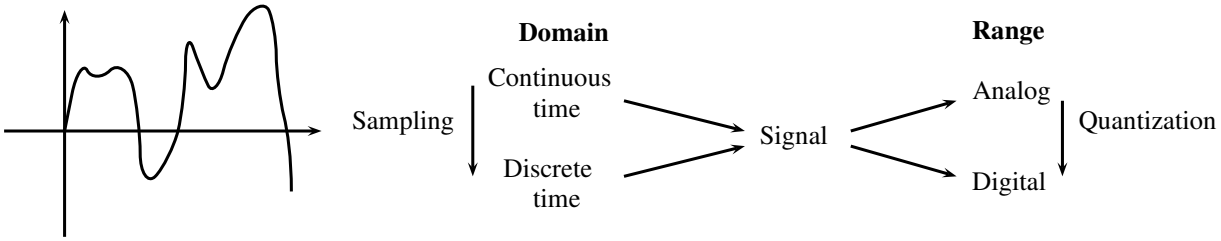
Next topic, lossy data compression: Given X, find a k-bit representation W, $X \to W \to \hat{X}$, such that $\hat{X}$ is a good reconstruction of X.

Real-world examples: codecs consist of a compressor and a decompressor

- Image: JPEG...
- Audio: MP3, CD...
- Video: MPEG...

23.1 Scalar quantization

Problem: Data isn't discrete! Often, a signal (function) comes from voltage levels or other continuous quantities. The question of how to map (naturally occurring) continuous time/analog signals into (electronics friendly) discrete/digital signals is known as *quantization*, or in information theory, as *rate distortion theory*.

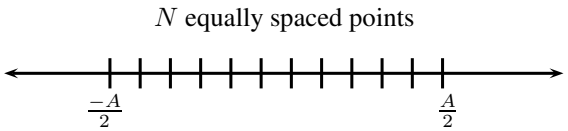


We will look at several ways to do quantization in the next few sections.

23.1.1 Scalar Uniform Quantization

The idea of qunatizing an inherently continuous-valued signal was most explicitly expounded in the patenting of Pulse-Coded Modulation (PCM) by A. Reeves, cf. [Ree65] for some interesting historical notes. His argument was that unlike AM and FM modulation, quantized (digital) signals could be sent over long routes without the detrimental accumulation of noise. Some initial theoretical analysis of the PCM was undertaken in 1947 by Oliver, Pierce, and Shannon (same Shannon), cf. [OPS48].

For a random variable $X \in [-A/2, A/2] \subset \mathbb{R}$, the scalar uniform quantizer $q_U(X)$ with N quantization points partitions the interval $[-A/2, A/2]$ uniformly

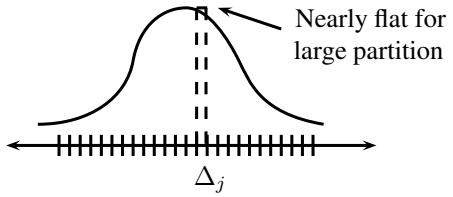


where the points are in $\{\frac{-A}{2} + \frac{kA}{N}, k = 0, \ldots, N-1\}$.

What is the *quality* (or fidelity) of this quantization? Most of the time, mean squared error is used as the quality criterion:

$$D(N) = \mathbb{E}|X - q_U(X)|^2$$

where D denotes the average *distortion*. Often $R = \log_2 N$ is used instead of N, so that we think about the number of bits we can use for quantization instead of the number of points. To analyze this scalar uniform quantizer, we'll look at the high-rate regime ($R \gg 1$). The key idea in the high rate regime is that (assuming a smooth density P_X), each quantization interval Δ_j looks nearly flat, so conditioned on Δ_j, the distribution is accurately approximately by a uniform distribution.



Let c_j be the j-th quantization point, and Δ_j be the j-th quantization interval. Here we have

$$\begin{aligned}
\mathbb{E}|X - q_U(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X - c_j|^2 | X \in \Delta_j] \mathbb{P}[X \in \Delta_j] \\
\text{(high rate approximation)} \quad &\approx \sum_{j=1}^{N} \frac{|\Delta_j|^2}{12} \mathbb{P}[X \in \Delta_j] \\
&= \frac{(\frac{A}{N})^2}{12} = \frac{A^2}{12} 2^{-2R}
\end{aligned}$$

How much do we gain per bit?

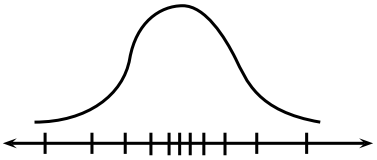
$$\begin{aligned}
10 \log_{10} SNR &= 10 \log_{10} \frac{Var(X)}{\mathbb{E}|X - q_U(X)|^2} \\
&= 10 \log_{10} \frac{12 Var(X)}{A^2} + (20 \log_{10} 2) R \\
&= \text{constant} + (6.02dB) R
\end{aligned}$$

For example, when X is uniform on $[-\frac{A}{2}, \frac{A}{2}]$, the constant is 0. Every engineer knows the rule of thumb "6dB per bit"; adding one more quantization bit gets you 6 dB improvement in SNR. However, here we can see that this rule of thumb is valid only in the high rate regime. (Consequently, widely articulated claims such as "16-bit PCM (CD-quality) provides 96 dB of SNR" should be taken with a grain of salt.)

Note: The above deals with X with a bounded support. When X is unbounded, a wise thing to do is to allocate the quantization points to the range of values that are more likely and saturate the large values at the dynamic range of the quantizer. Then there are two contributions, known as the granular distortion and overload distortion. This leads us to the question: Perhaps instead of uniform quantization optimal?

23.1.2 Scalar Non-uniform Quantization

Since our source has density p_X, a good idea might be to use more quantization points where p_X is larger, and less where p_X is smaller.

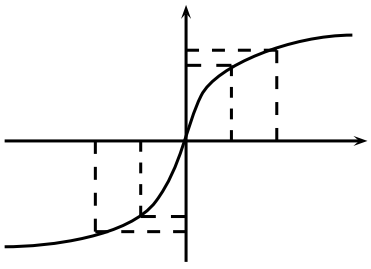


Often the way such quantizers are implemented is to take a monotone transformation of the source $f(X)$, perform uniform quantization, then take the inverse function:

$$\begin{array}{ccc} X & \xrightarrow{f} & U \\ \downarrow{\scriptstyle q} & & \downarrow{\scriptstyle q_U} \\ \hat{X} & \xleftarrow[f^{-1}]{} & q_U(U) \end{array} \tag{23.1}$$

i.e., $q(X) = f^{-1}(q_U(f(X)))$. The function f is usually called the *compander* (compressor+expander). One of the choice of f is the CDF of X, which maps X into uniform on $[0, 1]$. In fact, this compander architecture is optimal in the high-rate regime (fine quantization) but the optimal f is not the CDF (!). We defer this discussion till Section 23.1.4.

In terms of practical considerations, for example, the human ear can detect sounds with volume as small as 0 dB, and a painful, ear-damaging sound occurs around 140 dB. Achieveing this is possible because the human ear inherently uses logarithmic companding function. Furthermore, many natural signals (such as *differences* of consecutive samples in speech or music (but not samples themselves!)) have an approximately Laplace distribution. Due to these two factors, a very popular and sensible choice for f is the μ-companding function



$$f(X) = \text{sign}(X)\frac{\ln(1+\mu|X|)}{\ln(1+\mu)}$$

which compresses the dynamic range, uses more bits for smaller $|X|$'s, e.g. $|X|$'s in the range of human hearing, and less quantization bits outside this region. This results in the so-called μ-law which is used in the digital telecommunication systems in the US, while in Europe they use a slightly different compander called the A-law.

23.1.3 Optimal Scalar Quantizers

Now we look for the optimal scalar quantizer given R bits for reconstruction. Formally, this is

$$D_{scalar}(R) = \min_{q:|\text{Im } q|\leq 2^R} \mathbb{E}|X - q(X)|^2$$

Intuitively, we would think that the optimal quantization regions should be contiguous; otherwise, given a point c_j, our reconstruction error will be larger. Therefore quantizers are piecewise constant:

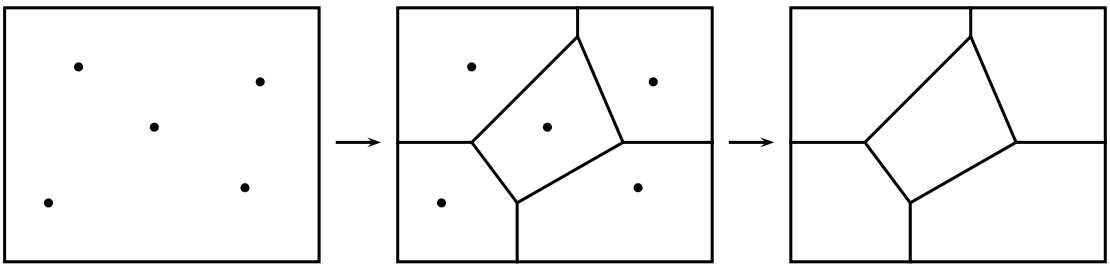
$$q(x) = c_j \mathbf{1}_{T_j \leq x \leq T_{j+1}}$$

for some $c_j \in [T_j, T_{j+1}]$.

Simple example: One-bit quantization of $X \sim \mathcal{N}(0, \sigma^2)$. Then optimal quantization points are $c_1 = \mathbb{E}[X|X \geq 0] = \sqrt{\frac{2}{\pi}}\sigma$, $c_2 = \mathbb{E}[X|X \leq 0] = -\sqrt{\frac{2}{\pi}}\sigma$.

With ideas like this, in 1982 Stuart Lloyd developed an algorithm (called *Lloyd's algorithm*) for iteratively finding optimal quantization regions and points. This works for both the scalar and vector cases, and goes as follows:

1. Pick any $N = 2^k$ points
2. Draw the Voronoi regions around the chosen quantization points (aka minimum distance tessellation, or set of points closest to c_j), which forms a partition of the space.
3. Update the quantization points by the centroids ($\mathbb{E}[X|X \in D]$) of each Voronoi region.
4. Repeat.



Steps of Lloyd's algorithm

Lloyd's clever observation is that the centroid of each Voronoi region is (in general) different than the original quantization points. Therefore, iterating through this procedure gives the *Centroidal Voronoi Tessellation* (CVT - which are very beautiful objects in their own right), which can be viewed as the fixed point of this iterative mapping. The following theorem gives the results about Lloyd's algorithm

Theorem 23.1 (Lloyd)**.**

1. *Lloyd's algorithm always converges to a Centroidal Voronoi Tessellation.*

2. *The optimal quantization strategy is always a CVT.*

3. *CVT's are non-unique, and the algorithm may converge to non-global optima.*

Remark: The third point tells us that Lloyd's algorithm isn't always guaranteed to give the optimal quantization strategy.[1] One sufficient condition for uniqueness of a CVT is the log-concavity of the density of X [Fleischer '64]. Thus, for Gaussian P_X, Lloyd's algorithm outputs the optimal quantizer, but even for Gaussian, if $N > 3$, optimal quantization points are not known in closed form! So it's hard to say too much about optimal quantizers. Because of this, we next look for an approximation in the regime of huge number of points.

23.1.4 Fine quantization

[Panter-Dite '51] Now we look at the high SNR approximation. For this, introduce the probability density function $\lambda(x)$, which represents the density of our quantization points and allows us to approximate summations by integrals[2]. Then the number of quantization points in any interval $[a, b]$ is $\approx N \int_a^b \lambda(x)dx$. For any point x, denote its distance to the closest quantization point by $\Delta(x)$. Then $N\lambda(x)\Delta(x) \approx 1 \implies \Delta(x) \approx \frac{1}{N\lambda(x)}$. With this approximation, the quality of reconstruction is

$$\begin{aligned}\mathbb{E}|X - q(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X - c_j|^2 | X \in \Delta_j]\mathbb{P}[X \in \Delta_j] \\ &\approx \sum_{j=1}^{N} \mathbb{P}[X \in \Delta_j]\frac{|\Delta_j|^2}{12} \approx \int p(x)\frac{\Delta^2(x)}{12}dx \\ &= \frac{1}{12N^2}\int p(x)\lambda^{-2}(x)dx\end{aligned}$$

To find the optimal density λ that gives the best reconstruction (minimum MSE) when X has density p, we use Hölder's inequality: $\int p^{1/3} \le (\int p\lambda^{-2})^{1/3}(\int \lambda)^{2/3}$. Therefore $\int p\lambda^{-2} \ge (\int p^{1/3})^3$, with equality iff $p\lambda^{-2} \propto \lambda$. Hence the optimizer is $\lambda^*(x) = \frac{f^{1/3}(x)}{\int f^{1/3}dx}$. Therefore when $N = 2^R$,[3]

$$D_{scalar}(R) \approx \frac{1}{12}2^{-2R}\left(\int p^{1/3}(x)dx\right)^3$$

So our optimal quantizer density in the high rate regime is proportional to the cubic root of the density of our source. This approximation is called the *Panter-Dite approximation*. For example, when $X \sim \mathcal{N}(0, \sigma^2)$, this gives

$$D_{scalar}(R) \approx \sigma^2 2^{-2R}\frac{\pi\sqrt{3}}{2}$$

Note: In fact, in *scalar* case the optimal non-uniform quantizer can be realized using the compander architecture (23.1) that we discussed in Section 23.1.2: As an exercise, use Taylor expansion to

[1] As a simple example one may consider $P_X = \frac{1}{3}\phi(x-1) + \frac{1}{3}\phi(x) + \frac{1}{3}\phi(x+1)$ where $\phi(\cdot)$ is a very narrow pdf, symmetric around 0. Here the CVT with centers $\pm\frac{2}{3}$ is not optimal among binary quantizers (just compare to any quantizer that quantizes two adjacent spikes to same value).

[2] This argument is easy to make rigorous. We only need to define reconstruction points c_j as solutions of

$$\int_{-\infty}^{c_j} \lambda(x)\,dx = \frac{j}{N}\,.$$

[3] In fact when $R \to \infty$, "$\approx$" can be replaced by "$= 1 + o(1)$" [Zador '56].

analyze the quantization error of (23.1) when $N \to \infty$. The optimal compander $f : \mathbb{R} \to [0,1]$ turns out to be $f(x) = \frac{\int_{-\infty}^{t} p^{1/3}(t)dt}{\int_{-\infty}^{\infty} p^{1/3}(t)dt}$ [Bennett '48, Smith '57].

23.1.5 Fine quantization and variable rate

So far we were considering quantization with restriction on the cardinality of the image of $q(\cdot)$. If one, however, intends to further compress the values $q(X)$ via noiseless compressor, a more natural constraint is to bound $H(q(X))$.

Koshelev [Kos63] discovered in 1963 that in the high rate regime uniform quantization is asymptotically optimal under the entropy constraint. Indeed, if q_Δ is a uniform quantizer with cell size Δ, then it is easy to see that

$$H(q_\Delta(X)) = h(X) - \log \Delta + o(1)\,, \tag{23.2}$$

where $h(X) = -\int p_X(x) \log p_X(x)\, dx$ is the differential entropy of X. So a uniform quantizer with $H(q(X)) = R$ achieves

$$D = \frac{\Delta^2}{12} \approx 2^{-2R} \frac{2^{2h(X)}}{12}\,.$$

On the other hand, any quantizer with unnormalized point density function $\Lambda(x)$ (i.e. smooth function such that $\int_{-\infty}^{c_j} \Lambda(x)dx = j$) can be shown to achieve (assuming $\Lambda \to \infty$ pointwise)

$$D \approx \frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \tag{23.3}$$

$$H(q(X)) \approx \int p_X(x) \log \frac{\Lambda(x)}{p_X(x)}\, dx \tag{23.4}$$

Now, from Jensen's inequality we have

$$\frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \geq \frac{1}{12} \exp\{-2 \int p_X(x) \log \Lambda(x)\, dx\} \approx 2^{-2H(q(X))} \frac{2^{2h(X)}}{12}\,,$$

concluding that uniform quantizer is asymptotically optimal.

Furthermore, it turns out that for any source, even the optimal vector quantizers (to be considered next) can not achieve distortion better that $2^{-2R} \frac{2^{2h(X)}}{2\pi e}$ – i.e. the maximal improvement they can gain (on any iid source!) is 1.53 dB (or 0.255 bit/sample). This is one reason why scalar uniform quantizers followed by lossless compression is an overwhelmingly popular solution in practice.

23.2 Information-theoretic vector quantization

By doing vector quantization (namely, compressing $(X_1, \ldots, X_n) \to 2^{nR}$ points), rate-distortion theory tells us that when n is large, we can achieve the per-coordinate MSE:

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

which saves 4.35 dB (or 0.72 bit/sample). This should be rather surprising, so we repeat it again: even when $X_1, \ldots, X_n$ are iid, we can get better performance by quantizing X_i jointly. One instance of this surprising effect is the following:

Hamming Game: Given 100 unbiased bits, we want to look at them and scribble something down on a piece of paper that can store 50 bits at most. Later we will be asked to guess the

original 100 bits, with the goal of maximizing the number of correctly guessed bits. What is the best strategy? Intuitvely, the optimal strategy would be to store half of the bits then guess on the rest, which gives 25% BER. However, as we will show in the next few lectures, the optimal strategy amazingly achieves a BER of 11%. Note does this happen? After all we are guessing independent bits and the utility function (BER) treats all bits equally. Some intuitive explanation:

1. Applying scalar quantization componentwise results in quantization region that are hypercubes, which might not be efficient for covering.
2. Concentration of measures removes many source realizations that are highly unlikely. For example, if we think about quantizing a single Gaussian X, then we need to cover large portion of $\mathbb{R}$ in order to cover the cases of significant deviations of X from 0. However, when we are quantizing many $(X_1, \ldots, X_n)$ together, the law of large numbers makes sure that many X_j's cannot conspire together and all produce large values. Thus, we may exclude large portions of the $\mathbb{R}^n$ from consideration.

Math Formalism: A lossy compressor is an encoder/decoder pair (f, g) where

$$X \xrightarrow{f} W \xrightarrow{g} \hat{X}$$

- $X \in \mathcal{X}$ - continuous source
- W - discrete data
- $\hat{X} \in \hat{\mathcal{X}}$ - reproduction

A *distortion metric* is a function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R} \cup \{+\infty\}$ (loss function). There are various formulations of the lossy compression problem:

1. Fixed length (fixed rate), average distortion: $W \in [M]$, minimize $\mathbb{E}[d(X, \hat{X})]$.
2. Fixed length, excess distortion: $W \in [M]$, minimize $\mathbb{P}[d(X, \hat{X}) > D]$.
3. Variable length, max distortion: $W \in \{0,1\}^*$, $d(X, \hat{X}) \le D$ a.s., minimize $\mathbb{E}[\text{length}(W)]$ or $H(\hat{X}) = H(W)$.

Note: In this course we focus on fixed length and average distortion loss compression. The difference between average distortion and excess distortion is analogous to average risk bound and high-probability bound in statistics/machine learning.

Definition 23.1. Rate-distortion problem is characterized by a pair of alphabets $\mathcal{A}$, $\hat{\mathcal{A}}$, a single-letter distortion function $d(\cdot, \cdot) : \mathcal{A} \times \hat{\mathcal{A}} \to \mathbb{R} \cup \{+\infty\}$ and a source – a sequence of $\mathcal{A}$-valued r.v.'s $(S_1, S_2, \ldots)$. A separable distortion metric is defined for n-letter vectors by averaging the single-letter distortions:

$$d(a^n, \hat{a}^n) \triangleq \frac{1}{n} \sum d(a_i, \hat{a}_i)$$

An (n, M, D)-code is

- Encoder $f : \mathcal{A}^n \to [M]$
- Decoder $g : [M] \to \hat{\mathcal{A}}^n$

- Average distortion: $\mathbb{E}[d(S^n, g(f(S^n)))] \le D$

Fundamental limit:

$$M^*(n, D) = \min\{M : \exists(n, M, D)\text{-code}\}$$
$$R(D) = \limsup_{n\to\infty} \frac{1}{n} \log M^*(n, D)$$

Now that we have the definition, we give the (surprisingly simple) general converse

Theorem 23.2 (General Converse)**.** *For all lossy codes $X \to W \to \hat{X}$ such that $\mathbb{E}[d(X, \hat{X})] \le D$, we have*

$$\log M \ge \varphi_X(D) \triangleq \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\le D} I(X; Y)$$

where $W \in [M]$.

Proof.

$$\log M \ge H(W) \ge I(X; W) \ge I(X; \hat{X}) \ge \varphi_X(D)$$

where the last inequality follows from the fact that $P_{\hat{X}|X}$ is a feasible solution (by assumption). □

Theorem 23.3 (Properties of φ_X)**.**

1. *φ_X is convex, non-increasing.*
2. *φ_X continuous on (D_0, ∞), where $D_0 = \inf\{D : \varphi_X(D) < \infty\}$.*
3. *If*

$$d(x, y) = \begin{cases} D_0 & x = y \\ > D_0 & x \ne y \end{cases}$$

   Then $\varphi_X(D_0) = I(X; X)$.

4. *Let*

$$D_{\max} = \inf_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}d(X, \hat{x}).$$

   Then $\varphi_X(D) = 0$ for all $D > D_{\max}$. If $D_0 > D_{max}$ then also $\varphi_X(D_{max}) = 0$.

Note: If $D_{\max} = \mathbb{E}d(X, \hat{x})$ for some $\hat{x}$, then $\hat{x}$ is the "default" reconstruction of X, i.e., the best estimate when we have no information about X. Therefore $D \ge D_{\max}$ can be achieved for free. This is the reason for the notation $D_{\max}$ despite that it is defined as an infimum.
Example: (Gaussian with MSE distortion) For $X \sim \mathcal{N}(0, \sigma^2)$ and $d(x, y) = (x - y)^2$, we have $\varphi_X(D) = \frac{1}{2}\log^+ \frac{\sigma^2}{D}$. In this case $D_0 = 0$ which is not attained; $D_{\max} = \sigma^2$ and if $D \ge \sigma^2$, we can simply output $\hat{X} = 0$ as the reconstruction which requires zero bits.

Proof.

1. Convexity follows from the convexity of $P_{Y|X} \mapsto I(P_X, P_{Y|X})$.
2. Continuity on interior of the domain follows from convexity.

3. The only way to satisfy the constraint is to take $X = Y$.

4. For any $D > D_{max}$ we can set $\hat{X} = \hat{x}$ deterministically. Thus $I(X;\hat{x}) = 0$. The second claim follows from continuity. □

In channel coding, we looked at the capacity and the information capacity. We define the *Information Rate-Distortion function* in an analogous way here, which by itself is *not* an operational quantity.

Definition 23.2. The Information Rate-Distortion function for a source is

$$R_i(D) = \limsup_{n\to\infty} \frac{1}{n}\varphi_{S^n}(D) \text{ where } \varphi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n,\hat{S}^n)]\le D} I(S^n;\hat{S}^n)$$

And $D_0 = \inf\{D : R_i(D) < \infty\}$.

The reason for defining $R_i(D)$ is because from Theorem 23.2 we immediately get:

Corollary 23.1. $\forall D,\ R(D) \ge R_i(D)$.

Naturally, the information rate-distortion function inherit the properties of φ:

Theorem 23.4 (Properties of R_i)**.**

1. $R_i(D)$ *is convex, non-increasing*

2. $R_i(D)$ *is continuous on* (D_0, ∞), *where* $D_0 \triangleq \inf\{D : R_i(D) < \infty\}$.

3. *If*

$$d(x,y) = \begin{cases} D_0 & x = y \\ > D_0 & x \ne y \end{cases}$$

   Then for stationary ergodic $\{S^n\}$, $R_i(D) = \mathcal{H}$ *(entropy rate) or* $+\infty$ *if* S_k *is not discrete.*

4. $R_i(D) = 0$ *for all* $D > D_{\max}$, *where*

$$D_{\max} \triangleq \limsup_{n\to\infty} \inf_{\hat{x^n}\in\hat{\mathcal{X}}} \mathbb{E}d(X^n, \hat{x^n}) .$$

   If $D_0 < D_{\max}$, *then* $R_i(D_{\max}) = 0$ *too.*

5. *(Single letterization) If the source* $\{S_i\}$ *is i.i.d., then*

$$R_i(D) = \phi_{S_1}(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S})$$

Proof. Properties 1-4 follow directly from corresponding properties of ϕ_{S^n} and property 5 will be established in the next section. □

23.3* Converting excess distortion to average

Finally, we discuss how to build a compressor for average distortion if we have a compressor for excess distortion, which we will not discuss in details in class.

Assumption D_p. Assume that for (S,d), there exists $p>1$ such that $D_p<\infty$, where

$$D_p \triangleq \sup_n \inf_{\hat{x}} (\mathbb{E}|d(S^n,\hat{x})|^p)^{1/p} < +\infty$$

i.e. that our separable distortion metric d doesn't grow too fast. Note that (by Minkowski's inequality) for stationary memoryless sources we have a single-letter bound:

$$D_p \le \inf_{\hat{x}} (\mathbb{E}|d(S,\hat{x})|^p)^{1/p} \tag{23.5}$$

Theorem 23.5 (Excess-to-Average)**.** *Suppose there exists $X \to W \to \hat{X}$ such that $W \in [M]$ and $\mathbb{P}[d(X,\hat{X})>D] \le \epsilon$. Suppose for some $p \ge 1$ and $\hat{x}_0 \in \hat{\mathcal{X}}$, $(\mathbb{E}[d(X,\hat{x}_0)]^p)^{1/p} = D_p < \infty$. Then there exists $X \to W' \to \hat{X}'$ code such that $W' \in [M+1]$ and*

$$\mathbb{E}[d(X,\hat{X}')] \le D(1-\epsilon) + D_p \epsilon^{1-1/p} \tag{23.6}$$

Remark 23.1. Theorem is only useful for $p>1$, since for $p=1$ the right-hand side of (23.6) does not converge to 0 as $\epsilon \to 0$.

Proof. We transform the first code into the second by adding one codeword:

$$f'(x) = \begin{cases} f(x) & d(x,g(f(x))) \le D \\ M+1 & \text{o/w} \end{cases}$$

$$g'(j) = \begin{cases} g(j) & j \le M \\ \hat{x}_0 & j = M+1 \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}[d(X,g'\circ f'(X)) &\le \mathbb{E}[d(X,\hat{X})|\hat{W} \ne M+1](1-\epsilon) + \mathbb{E}[d(X,x_0)\mathbf{1}\{\hat{W}=M+1\}] \\ \text{(Hölders Inequality)} \quad &\le D(1-\epsilon) + D_p\epsilon^{1-1/p} \end{aligned}$$

□

6.441 Information Theory

Spring 2016