I232 Information Theory
Chapter 10: Rate-Distortion Theory

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 May

Lossless vs. Lossy Source Coding

$$\mathbf{X} \to \boxed{\text{encoder } f} \to f(\mathbf{X}) \to \boxed{\text{decoder } g} \to \widehat{\mathbf{X}} = g\big(f(\mathbf{X})\big)$$

- $\mathbf{X}$: source sequence
- $f(\mathbf{X})$: compressed sequence
- $\widehat{\mathbf{X}} = g(f(\mathbf{X}))$: uncompressed sequence or reconstruction

For *lossless* source coding: $\widehat{\mathbf{X}} = \mathbf{X}$, as in Chapters 4–6.

For *lossy* source coding: $\widehat{\mathbf{X}} \approx \mathbf{X}$, in this chapter.

- Reconstruction $\widehat{\mathbf{X}}$ does not need to be same as $\mathbf{X}$
- But $\widehat{\mathbf{X}}$ should be similar to $\mathbf{X}$.

Lossy Source Coding: Distortion Function

We need a way to measure "$\approx$" as in $\widehat{\mathbf{X}} \approx \mathbf{X}$.

Distortion measures how similar $\widehat{\mathbf{X}}$ and $\mathbf{X}$ are.

We will consider both discrete and continuous sources.

For discrete sources:

- Some digits in the reconstruction differ from the source
- Example distortion function: Hamming metric

For continuous sources (real numbers)

- Lossless reconstruction needs an infinite number of bits
- Lossy coding uses a finite number of bits to represent a real-valued source.
- Example distortion function: Euclidean distance

PNG format, File size 736 kB

High rate R, low distortion D

JPEG compression 100%, File Size 124 kB
Medium rate *R*, medium distortion *D*

JPEG compression 1%, File Size 5 kB

Low rate *R*, high distortion *D*

Motivation

Question:

- What is the optimal tradeoff between rate and distortion (loss)? Or
- What is the minimal rate for a given distortion?

Answer:

- Rate-distortion function $\Leftarrow$ Minimization of mutual information

Outline

10.1 Rate-Distortion Codes

$$\mathbf{X} \to \boxed{\text{encoder } f} \to f(\mathbf{X}) \to \boxed{\text{decoder } g} \to \widehat{\mathbf{X}} = g\big(f(\mathbf{X})\big)$$

Definition: A $(2^{nR}, n)$ *rate distortion code* consists of an encoding function f_n:

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\},$$

and a decoding function g_n:

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n.$$

Definition: Distortion

The distortion between two sequences $\mathbf{x}$ and $\hat{\mathbf{x}}$:

$$d(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \widehat{x}_i),$$

- For *discrete* x_i, the *Hamming distortion* is:

$$d(x_i, \hat{x}_i) = \begin{cases} 0, & \text{if } x_i = \hat{x}_i \\ 1, & \text{if } x_i \neq \hat{x}_i \end{cases} .$$

- For continuous x_i, the *squared-error distortion* or *Euclidean distortion* is:

$$d(x_i, \widehat{x}_i) = (x_i - \hat{x}_i)^2.$$

Definition: Expected Distortion

Definition

The *expected distortion* is for a $(2^{nR}, n)$ code is D:

$$
\begin{aligned}
D &= E\Big[d\Big(\mathbf{X}, g(f(\mathbf{X}))\Big)\Big] \\
&= \sum_{\mathbf{x} \in \mathcal{X}^n} p_{\mathbf{X}}(\mathbf{x}) d\Big(\mathbf{x}, g(f(\mathbf{x}))\Big)
\end{aligned}
$$

10.2 Rate-Distortion Theorem

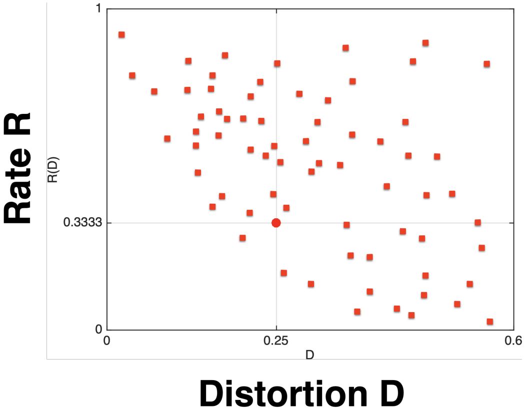
10.2.1 Rate-Distortion Region

Definition: A rate-distortion point (R, D) is *achievable* if there exists a sequence of $(2^{nR}, n)$ codes with encoder and decoder f_n, g_n with:

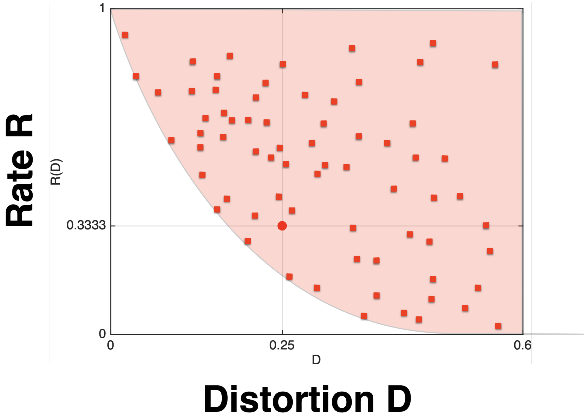
$$\lim_{n\to\infty} \mathrm{E}\Big[d\big(\mathsf{X}, g_n(f_n(\mathsf{X}))\big)\Big] \leq D.$$

Definition: A *rate-distortion region* for a source is the closure of the set of achievable rate distortion pairs (R, D).

Examples of Achievable Rate-Distortion Point



Examples of Rate-Distortion Region



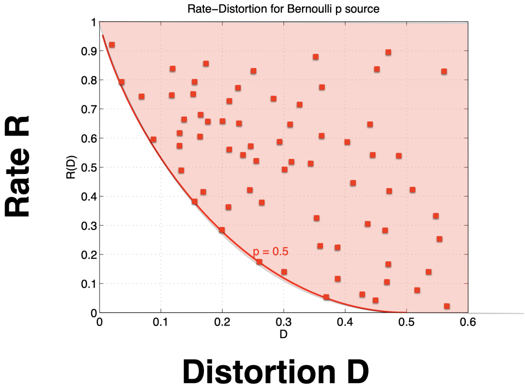
A rate-distortion region is shown in light red.

Rate-Distortion Function

Definition

The *rate-distortion function* is the infimum of rates R such that (R, D) is in the rate distortion region of the source of a given distortion D.

Example: A rate-distortion function is the red curve.



10.2.2 Rate-Distortion Theorem

Theorem: *Rate-Distortion Theorem* The rate distortion function for an iid source $\mathbf{X}$ with distribution $p(x)$ and distortion function $d(\cdot,\cdot)$ is equal to the information rate distortion function. That is:

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

where

$$\mathrm{E}\{d(X,\hat{X})\} = \sum_{x\in\mathcal{X}}\sum_{\hat{x}\in\hat{\mathcal{X}}} p(x,\hat{x})d(x,\hat{x}).$$

Notes:

- Minimization over all $p(\hat{x}|x)$ that satisfies the distortion constraint
- Minimization of mutual information!

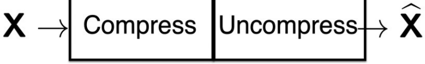
Mutual Information is the Flow of Bits

$I(X;Y)$ is the "rate of flow" of bits between X and Y.

Channel coding

Maximum transmit rate

$C = \max I(X;Y)$



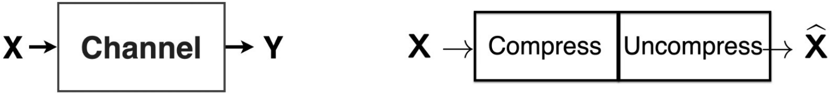
Lossy source coding

Minimum compressed rate

$R = \min I(X;\hat{X})$

Mutual Information is the Flow of Bits

$I(\mathsf{X};\mathsf{Y})$ is the "rate of flow" of bits between X and Y.



Channel coding

Maximum transmit rate

$C = \max I(X; Y)$

Lossy source coding

Minimum compressed rate

$R = \min I(X; \hat{X})$

Channel coding: *Maximization* of mutual information.

- The channel $p(y|x)$ is fixed, we optimize the input distribution $p(x)$.

Lossy source coding: *Minimization* of mutual information (subject to distortion)

- For lossy source coding we *minimize* the rate of flow,
- The source $p(x)$ is fixed, we optimize the conditional distribution $p(\hat{x}|x)$.

10.2.3 Comments on Proof

- Minimization is over $p_{\widehat{X}|X}(\widehat{x}|x)$.
- **There is a relationship between $p_{\widehat{X}|X}(\widehat{x}|x)$ and the code.**
- Recall the example:

Source $\mathbf{x}$	codeword $\widehat{\mathbf{x}} = g\big(f(\mathbf{x})\big)$	distortion $d(\mathbf{x}, \widehat{\mathbf{x}})$
0 0 0 0 0 0	0 0 0 0 0 0	0
1 0 0 0 0 0	0 0 0 0 0 0	$1/6$
0 1 0 0 0 0	0 0 0 0 0 0	$1/6$
1 0 0 1 0 0	0 0 0 0 0 0	$2/6$
0 1 0 1 0 0	0 0 0 0 0 0	$2/6$
1 1 0 1 0 0	1 1 1 0 0 0	$2/6$
0 0 1 1 0 0	0 0 0 0 0 0	$2/6$
	$\vdots$	

This code has:

$$p_{\widehat{X}|X}(\widehat{x}|x) = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}.$$

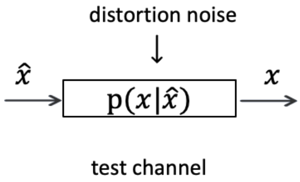
10.3 $R(D)$ for Discrete Sources

We want to find $R(D)$ for some specific cases.

1. Test Channel
2. $R(D)$ for the Binary Source

Test Channel

For a source X, and its reconstruction $\widehat{X}$, its test channel is::



- The distortion is like "noise" in the channel. Amount of noise is the distortion D.
- Easier to work with the conditional distribution $p(x|\hat{x})$ than $p(\hat{x}|x)$.
- Reconstruction $\widehat{x}$ is an a posteriori estimate of x, modeled by the test channel:

$$\text{“}x = \widehat{x} + n\text{”}.$$

General Approach for Solving $R(D)$

Given $p(x)$ and D:

1. Parameterize the test channel $p(x|\hat{x})$. Exploit any symmetries.
2. Parameterize the test channel input distribution $p(\hat{x})$.
3. Solve the unknown parameters to satisfy
   - $D = \mathrm{E}\{d(X, \hat{X})\}$
   - $p(x) = \sum_{\hat{x} \in \hat{\mathcal{X}}} p(\hat{x}) p(x|\hat{x})$
4. Compute $H(\mathsf{X}|\widehat{\mathsf{X}})$ for the test channel, and $\max H(\mathsf{X}|\widehat{\mathsf{X}})$ over parameters.
   Compute $R(D) = \min I(\mathsf{X}; \widehat{\mathsf{X}}) = H(\mathsf{X}) - \max H(\mathsf{X}|\widehat{X})$.

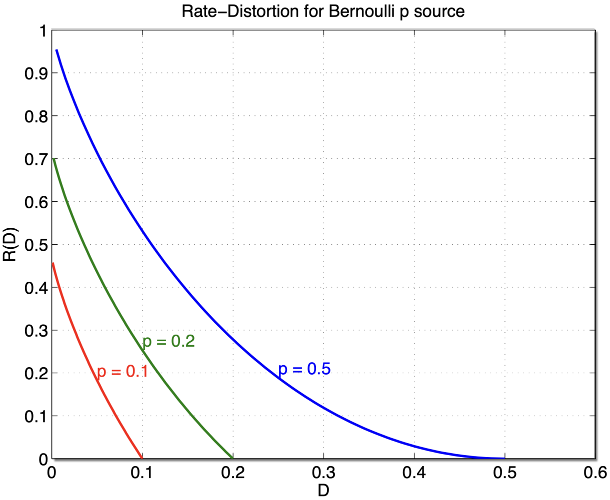
$R(D)$ for Binary Source

Proposition *Rate distortion for binary source* The $R(D)$ for $X \in \{0,1\}$ with $p(x) = [1-p, p]$ and Hamming distortion is:

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

- When $D > \min(p, 1-p)$, we let $\hat{X} = 0$ if $p \leq 0.5$, and $\hat{X} = 1$ if $p > 0.5$. Then, $D = \min(p, 1-p)$. Thus, $R(D) = 0$.
- Only need to study $D \leq \min(p, 1-p)$.

$R(D)$ for Binary Source



10.4 Quantization of Continuous-Valued Sources

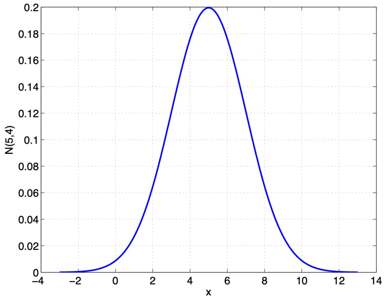
In this section, lossy source coding for continuous-valued random variables are considered.

1. Gaussian Sources
2. Rate-Distortion Function for Gaussian Sources
3. K-Means Algorithm

Gaussian Distribution

- *Gaussian probability distribution* $\mathcal{N}(m, \sigma^2)$ with mean m and variance σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$



Squared-Error Distortion

- The squared-error distortion between $\mathbf{x} = (x_1, \ldots, x_n)$ and $\widehat{\mathbf{x}} = (\widehat{x}_1, \ldots, \widehat{x}_n)$ is:

$$d(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \widehat{x}_i) \qquad = (x_i - \widehat{x}_i)^2$$

  where

$$d(x, \hat{x}) = (x - \hat{x})^2.$$

- The expected distortion for a $(2^{nR}, n)$ distortion code is:

$$D = E[(\mathsf{X} - \widehat{\mathsf{X}})^2],$$

  for random variables X and $\widehat{\mathsf{X}}$.

Codebook and Reconstruction Region

The codebook consists of K *reconstruction points*[1]:

$$\mathcal{C} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K\}$$

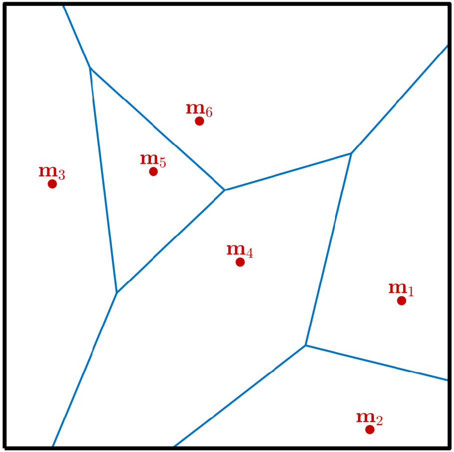
For each $\mathbf{m}_i$, the *reconstruction region* $\mathcal{R}_i \subset \mathbb{R}^n$ are the points are closest to $\mathbf{m}_i$:

$$\mathcal{R}_i = \{\mathbf{x} | d(\mathbf{x}, \mathbf{m}_i) \leq d(\mathbf{x}, \mathbf{m}_j), j \neq i\}.$$

$\mathcal{R}_i$ is sometimes called the Voronoi region of m_i.

[1] The codebook size is denoted by K instead of M

Reconstruction Regions $n = 2$



Axes are x_1, x_2 for a two-dimensional source. ★1-bit Gaussian

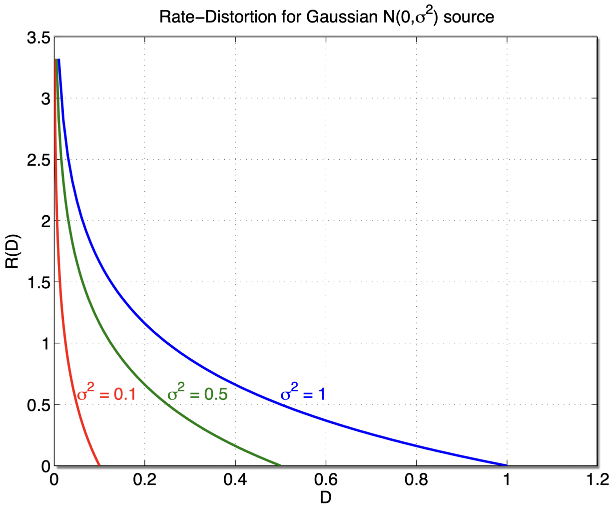
$R(D)$ for Gaussian Source

Proposition *Rate-distortion function for Gaussian source* The $R(D)$ for X $\sim \mathcal{N}(0, \sigma^2)$ with squared-error distortion is:

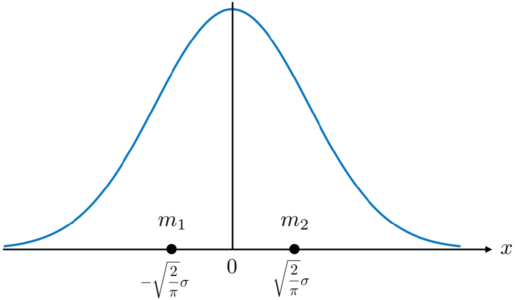
$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases}.$$

- Achieved for $n \to \infty$.
- When $D > \sigma^2$, we let $\hat{x} = 0$. Then, $D = \sigma^2$ and $R(D) = 0$.
- We only need to study $D \le \sigma^2$.

$R(D)$ for Gaussian Source



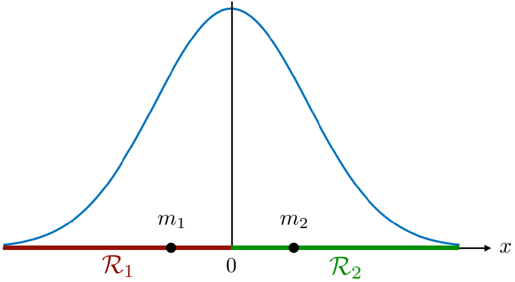
One-Bit Quantization of $n = 1$ Gaussian



Clearly symmetrical
Optimal reconstruction:

$$\widehat{x} = \begin{cases} -\sqrt{\frac{2}{\pi}}\sigma & \text{if } x < 0 \\ \sqrt{\frac{2}{\pi}}\sigma & \text{if } x \geq 0 \end{cases}$$

One-Bit Quantization of $n = 1$ Gaussian

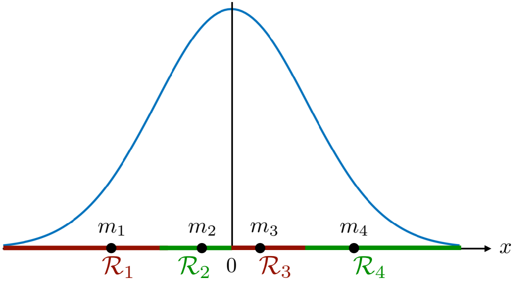


★Poll

Clearly symmetrical
Optimal reconstruction:

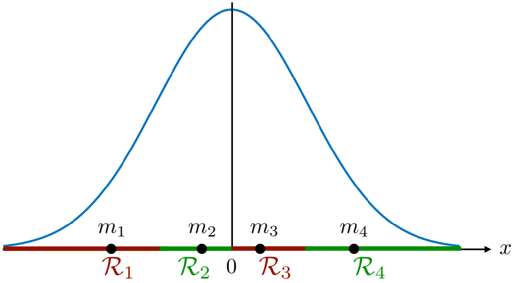
$$\widehat{x} = \begin{cases} -\sqrt{\frac{2}{\pi}}\sigma & \text{if } x < 0 \\ \sqrt{\frac{2}{\pi}}\sigma & \text{if } x \geq 0 \end{cases}$$

Two-Bit Quantization of $n = 1$ Gaussian



For two bits, no clear analytic approach to find optimal codebook.

Two Observations



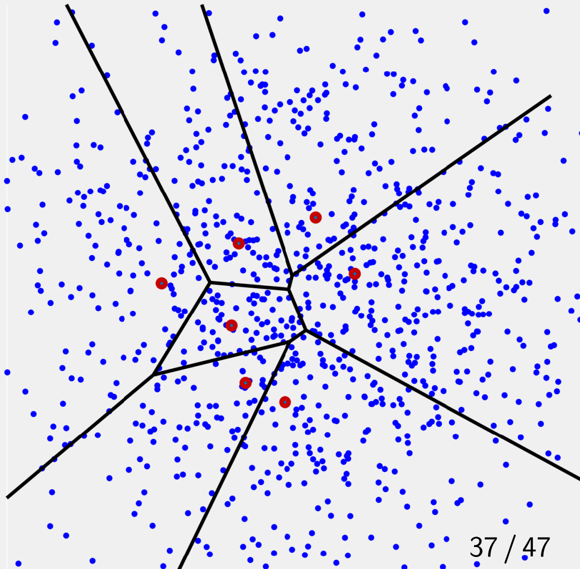
- Given the reconstruction points $\widehat{\mathcal{X}}$, the reconstruction region should minimize the average distortion.
  Any point x should be quantized to the closest point $\widehat{x}$ (with Euclidean metric).
- Within some region $\mathcal{R}_i$, the reconstruction point should be chosen to minimize the expected distortion for that region.

These Observations Lead to an Algorithm

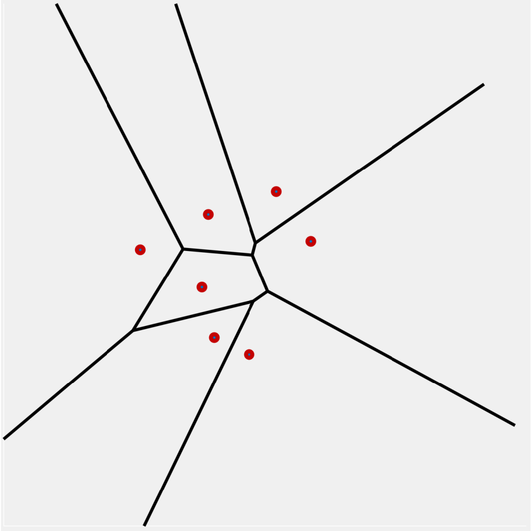
1. **Initialize** Randomly choose a code book.
2. **Assignment Step** Reconstruction region consists of points nearest each mean
3. **Update Step** For each region, move its codeword to the "center"
4. Repeat steps 2 and 3 until stable

K-Means Algorithm Initial Data

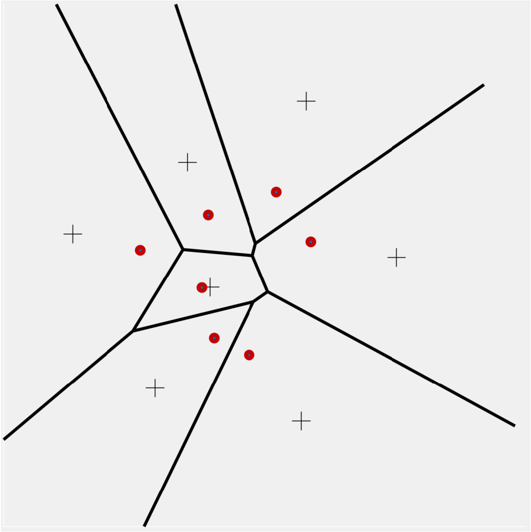
x_1, x_2 are iid Gaussian. We use sample data, not a distribution.



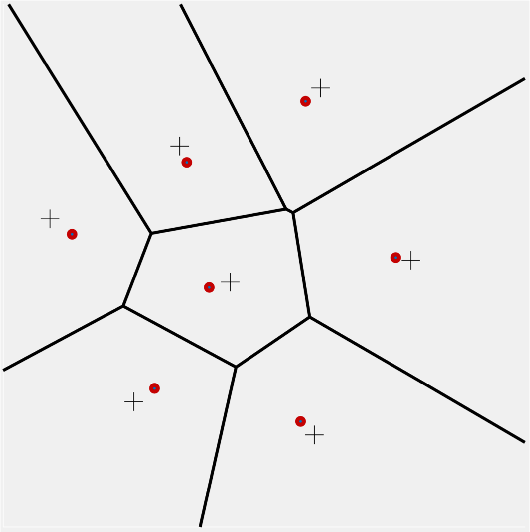
K-Means Iteration 0



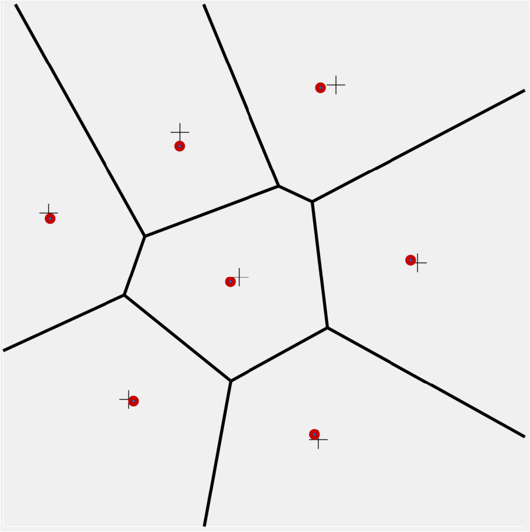
K-Means Iteration 1



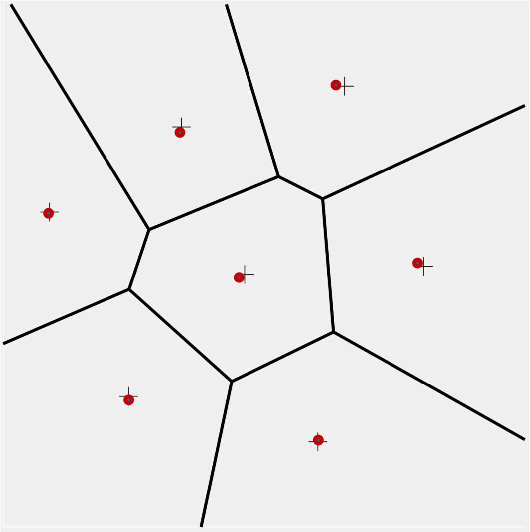
K-Means Iteration 2



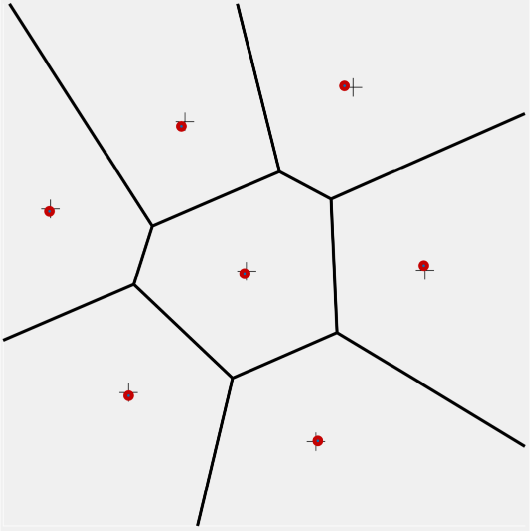
K-Means Iteration 3



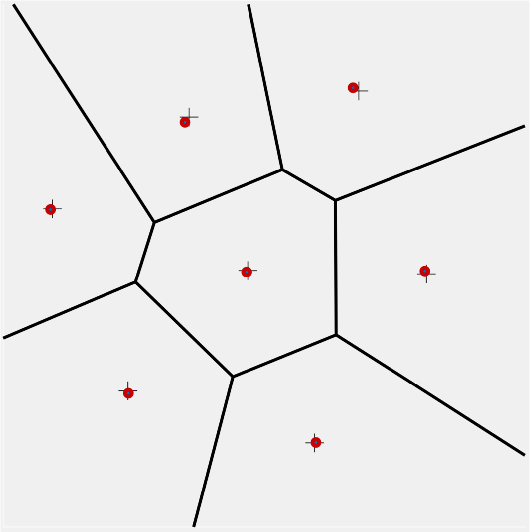
K-Means Iteration 4



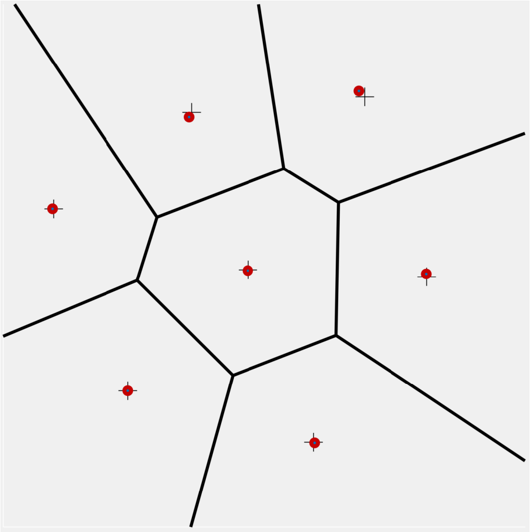
K-Means Iteration 5



K-Means Iteration 6



K-Means Iteration 7



What You Should Have Learned

- Rate-distortion theory deals with **lossy source coding**
- There is a fundamental **tradeoff** between the **compression rate** R and allowed **distortion** D
- This tradeoff is called the rate-distortion function $R(D)$
- The best possible tradeoff is the **minimization of mutual information**
- Obtain $R(D)$ analytically for independent and identically distributed binary and Gaussian sources, but not for complicated sources like digital images.

Class Info

- Tutorial Hours: Monday, May 22 at 13:30. Ask questions about homework.
- Homework 7 and 9 on LMS. Deadline: Monday, May 22 at 18:00
- Next lecture: Wednesday May 24. (I will send an LMS message if there will be a pop quiz)
- Homework 10 on LMS (soon).