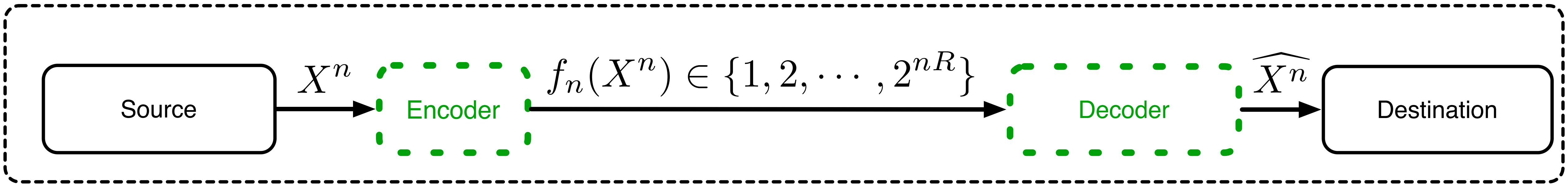
Chapter 10: Rate distortion theory

Chapter 10 outline

- Quantization
- Definitions
- Calculation of the rate-distortion function
- Converse of rate distortion theorem
- Strongly typical sequences
- Achievability of rate distortion theorem
- Characterization of the rate-distortion function
- Computation and channel capacity and rate-distortion function

Rate-distortion



Source $\longrightarrow$ minimum E[# bits] for error free representation

Source $\longrightarrow$ < minimum E[# bits] for error free representation

There will be errors and ***distortion*** in reconstructing the source!

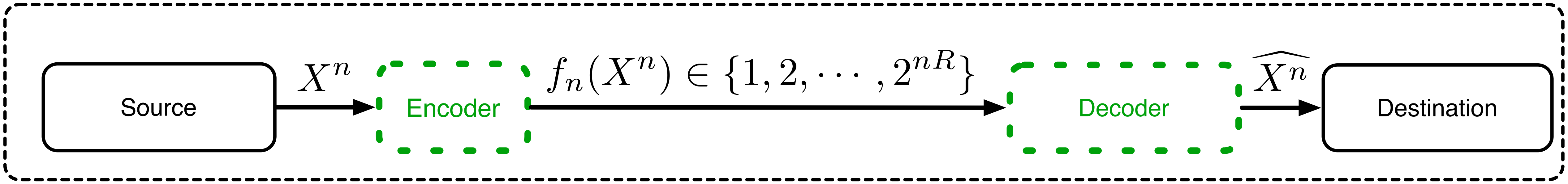
Rate-distortion theory describes the ***trade-off*** between lossy compression rate and the resulting distortion.

Quantization

- Consider representing a continuous valued random source - need infinite precision to represent it exactly!

- **Q:** what is the **best** possible representation of X for a given data rate?

- X: random variable to be represented
- $\hat{X}(X)$: representation of X
- R bits for the representation $\rightarrow$ $$|\hat{X}| = 2^{nR}$$
- Want to find the optimum set of values for $\hat{X}$ (reproduction points / code points) and associated regions



Quantization example: 1 bit Gaussian

Let $X \sim \mathcal{N}(0, \sigma^2)$ and assume a squared-error distortion measure. We wish to find the function $\hat{X}(X)$ such that $\hat{X}$ takes on 2^{nR} values and minimizes $E(X - \hat{X}(X))^2$.

Optimal 1 bit strategy?

Optimal 2 bit strategy?

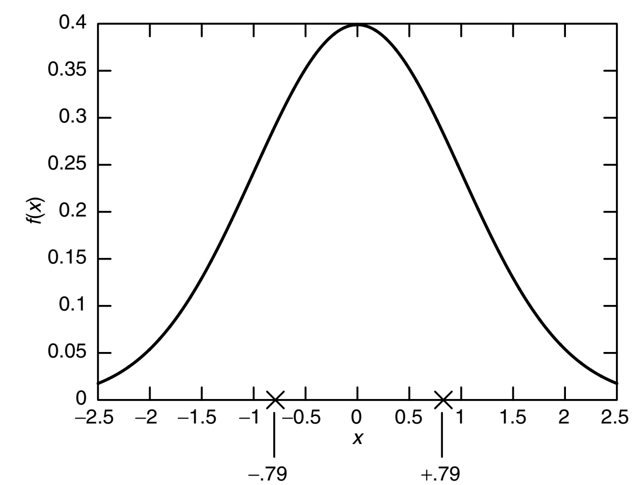


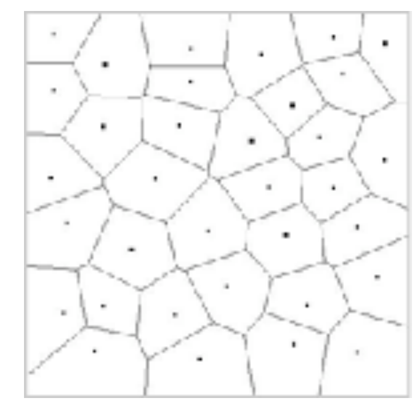
FIGURE 10.1. One-bit quantization of Gaussian random variable.

General observations:

- Given a set $\{\hat{X}(w)\}$ of reconstruction points, the distortion is minimized by mapping a source random variable X to the point closest to it, forming a set of regions called a *Voronoi* or Dirichlet partition.
- The reconstruction points should minimize the conditional expected distortion over their respective assignment regions.

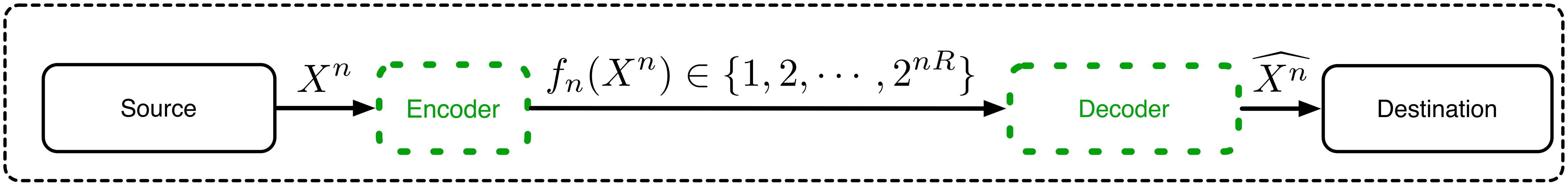
Quantization example: 1 bit Gaussian

- Lloyd algorithm - iterative way of finding a “good” quantizer
  - Find set of reconstruction points (centroids if MSE)
  - Find optimal reconstruction regions



- Benefits to quantizing many RVs at once?
  - Yes! *n* iid RVs represented using nR bits
  - Surprisingly, better to represent whole sequence than each RV independently, even though chosen iid!!!

Definitions



- $X_1, X_2, \cdots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$
- A **distortion function** or **distortion measure** is a mapping

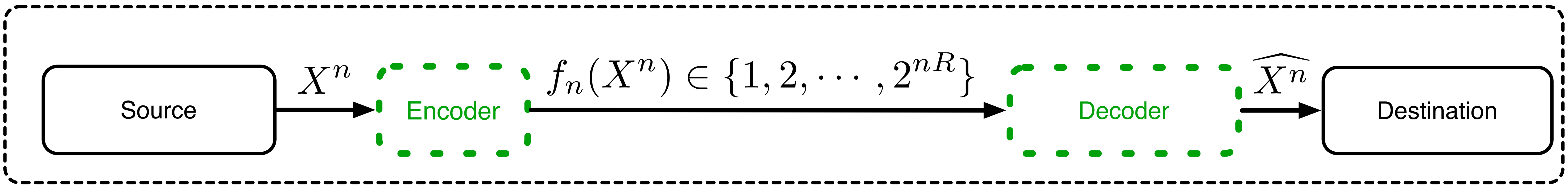
$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$$

from the set of source alphabet-reproduction alphabet pairs into the set of non-negative real numbers. Measures the "cost" of representing symbol x by $\hat{x}$.

- A distortion measure is said to be **bounded** if the maximum value of the distortion is finite,

$$d_{max} := \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty$$

Definitions



- Two most common distortion functions:
  - Hamming distortion:

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

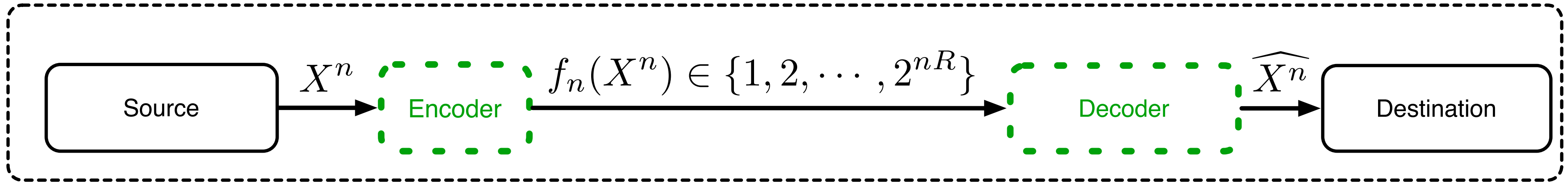
  - Squared-error distortion:

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

- We *define* the *distortion between sequences* x^n and $\hat{x}^n$ as

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

Definitions



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

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

  and a decoding (reproduction) function,

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

- The distortion associated with the $(2^{nR}, n)$ code is defined as

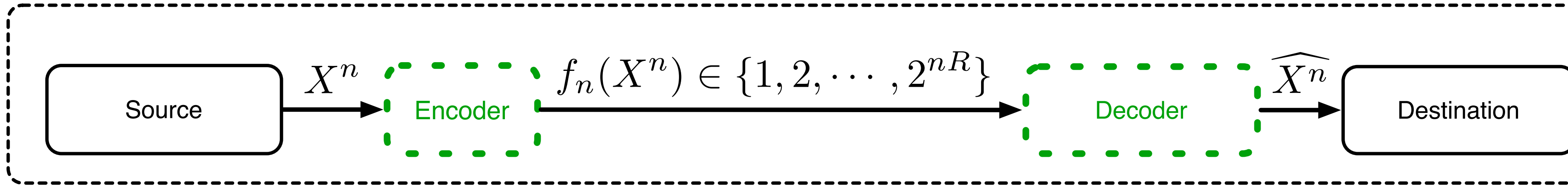
$$D = E[d(X^n, g_n(f_n(X^n))),$$

  where the expectation is with respect to the probability distribution on $\mathcal{X}$,

$$D = \sum p(x^n) d(x^n, g_n(f_n(x^n))).$$

- The set of n-tuples $g_n(1), g_n(2), \cdots, g_n(2^{nR})$, denoted by $\hat{X}^n(1), \hat{X}^n(2), \cdots, \hat{X}^n(2^{nR})$ constitutes the *codebook* and $f_n^{-1}(1), \cdots f_n^{-1}(2^{nR})$ are the associated *assignment regions.*

Definitions

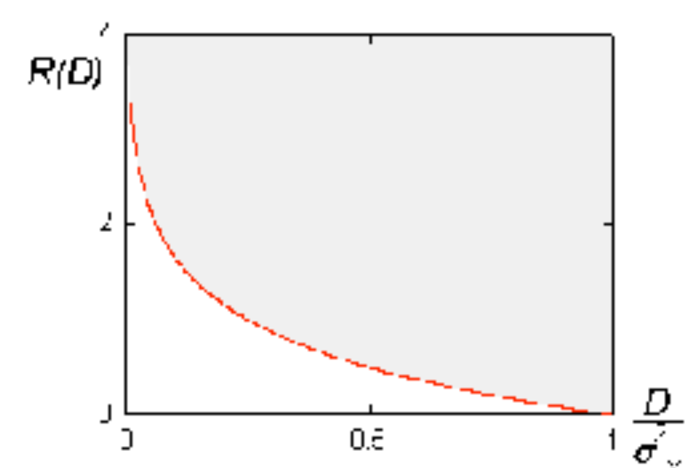


- A rate-distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) with $\lim_{n\to\infty} E[d(X^n, g_n(f_n(X^n))] \leq D$.

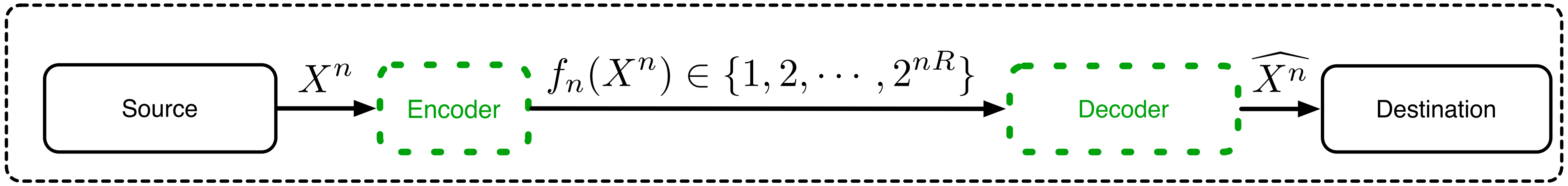
- The *rate-distortion region* for a source is the closure of the set of achievable rate distortion pairs (R, D).

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

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



Definitions

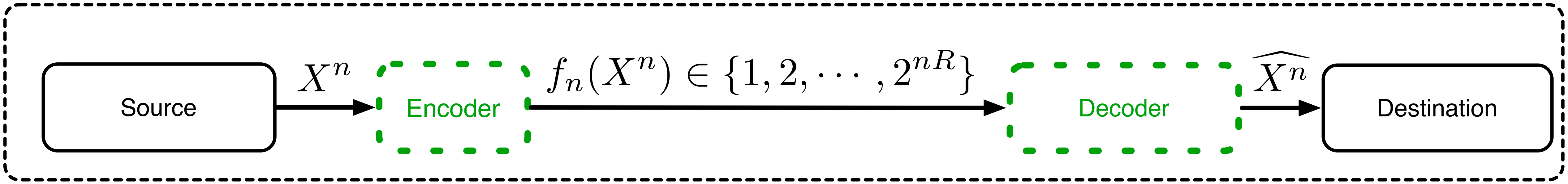


- The *rate-distortion function* $R(D)$ is the **infimum** of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.
- The *information rate distortion function* $R^{(I)}(D)$ for a source X with distortion measure $d(x, \hat{x})$ is defined as

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

  where the minimization is over all conditional distributions $p(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ satisfies the expected distortion constraint.

Main Theorem

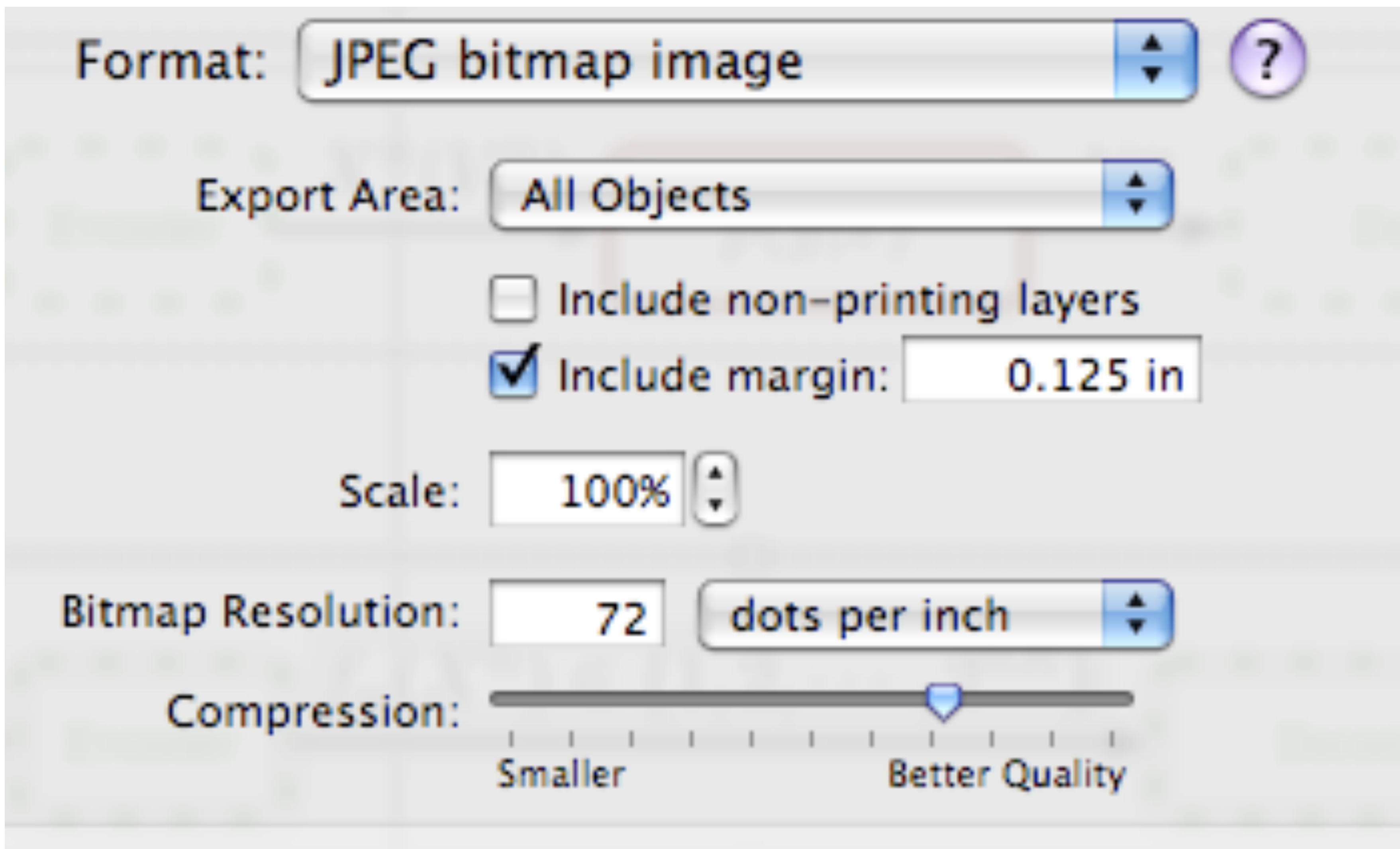


Theorem: The rate distortion function for an i.i.d. source X with distributed $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. Thus,

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

is the minimum achievable rate at distortion D.

A few examples



Calculating R(D) - binary source

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

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

Key proof ideas:

- Hamming distance, modulo 2 sum, $X \oplus \hat{X} = 1$ whenever $X \neq \hat{X}$.
- Find a *lower* bound on $I(X; \hat{X})$
- Show that this lower bound is achievable by finding a lower-bound achieving distribution for $\hat{X}$.

Calculating R(D) - binary source

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

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

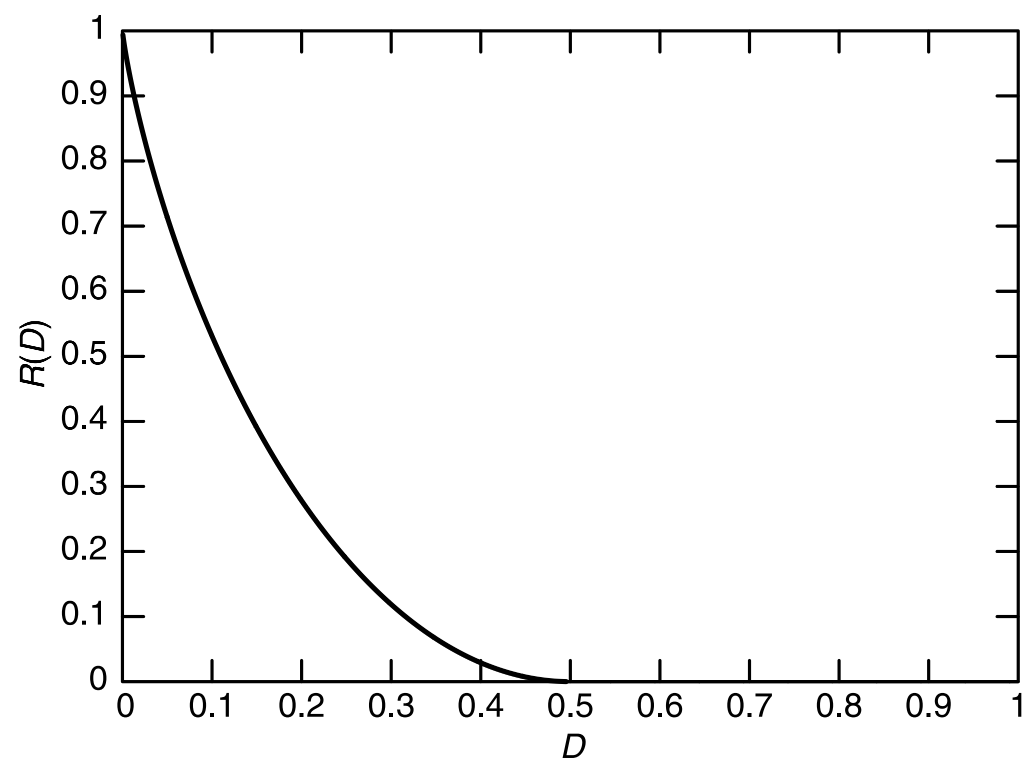
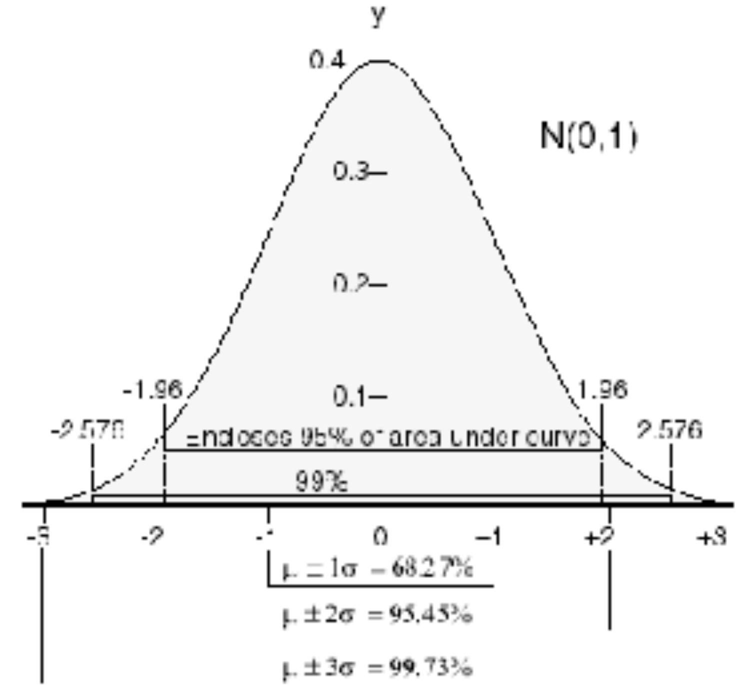


FIGURE 10.4. Rate distortion function for a Bernoulli $(\frac{1}{2})$ source.

Calculating R(D) - Gaussian source



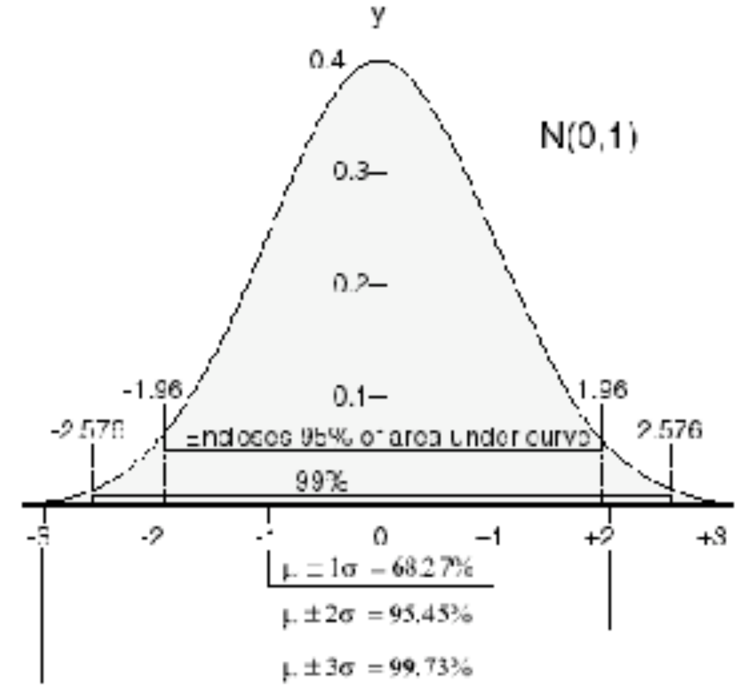
Theorem: The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared-error distortion is given by

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

Key proof ideas:

- Find a *lower* bound on $I(X; \hat{X})$
- Show that this lower bound is achievable by finding a lower-bound achieving distribution for $\hat{X}$.
- Exploit entropy maximizing (subject to 2nd moment constraint) property of Gaussian distribution

Calculating R(D) - Gaussian source



Theorem: The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared-error distortion is given by

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

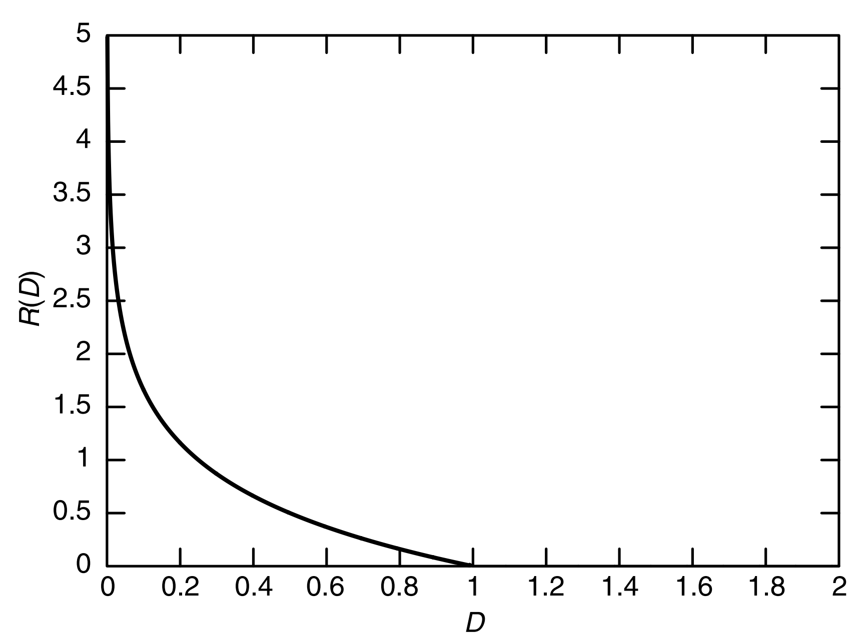
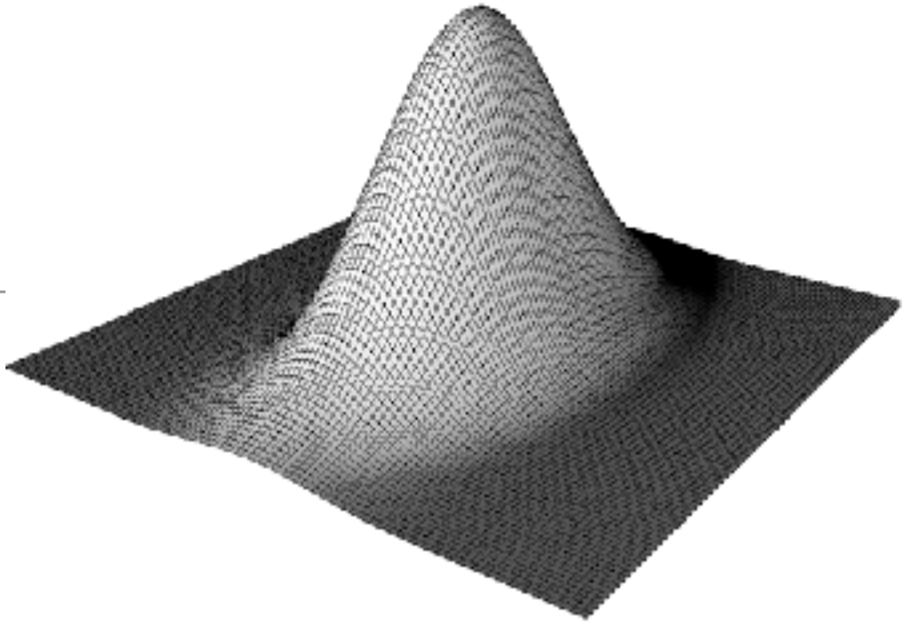


FIGURE 10.6. Rate distortion function for a Gaussian source.

Calculating R(D) - Gaussian source



Theorem: Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, 2, \cdots, m$ be independent Gaussian random variables, and let the distortion measure be $d(x^m, \hat{x}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$. Then rate distortion function is given by

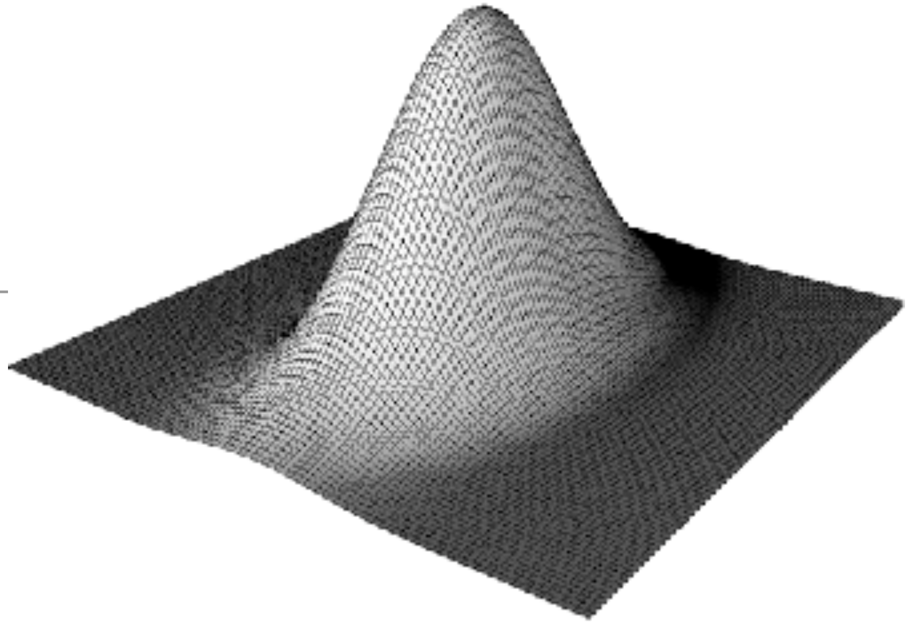
$$R(D) = \sum_{i=1}^m \frac{1}{2} \log \frac{\sigma_i^2}{D_i},$$

where

$$D_i = \begin{cases} \lambda, & \text{if } \lambda < \sigma_i^2 \\ \sigma^2, & \text{if } \lambda \geq \sigma_i^2, \end{cases}$$

where λ is chosen so that $\sum_{i=1}^m D_i = D$.

Calculating R(D) - Gaussian source



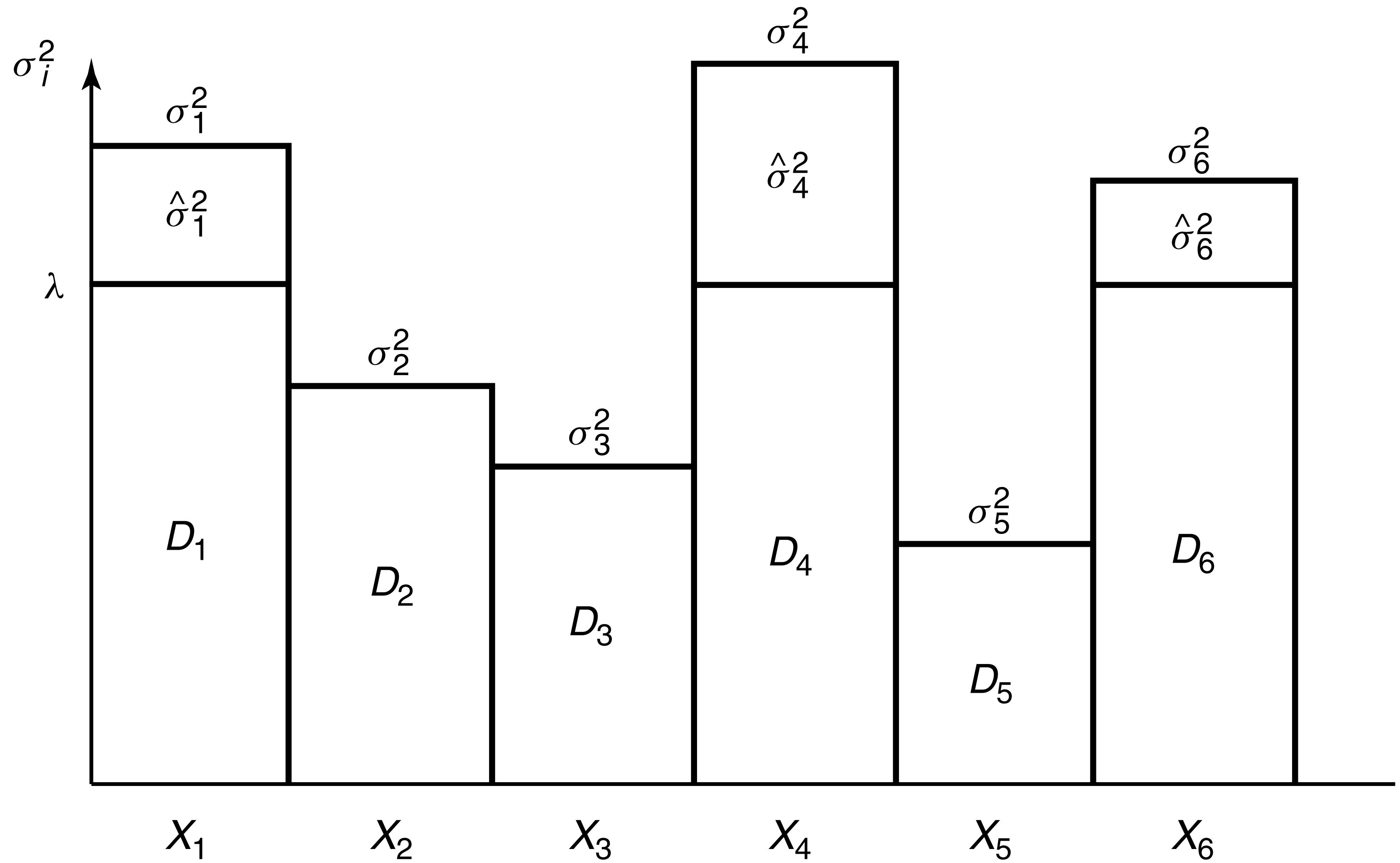
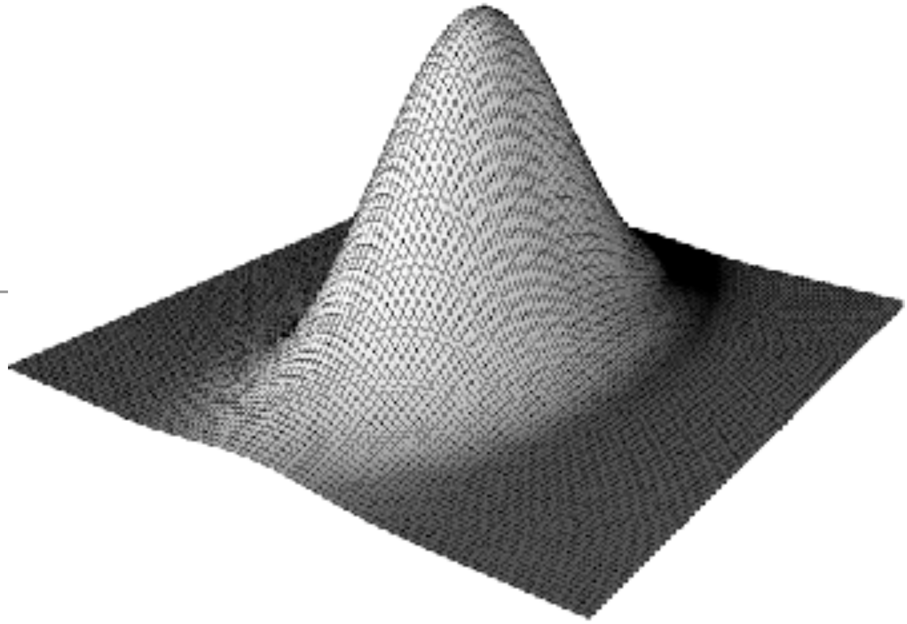


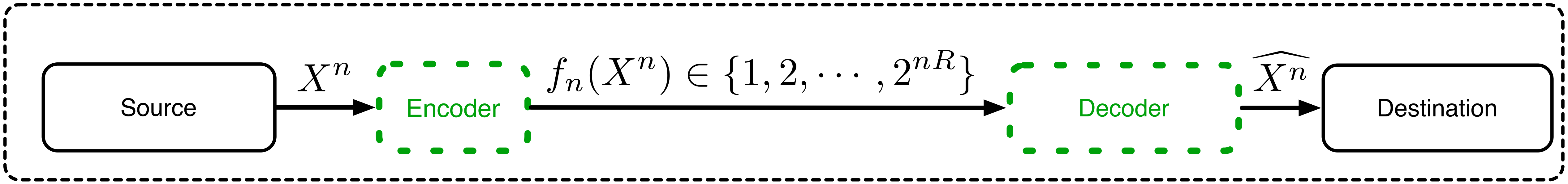
FIGURE 10.7. Reverse water-filling for independent Gaussian random variables.

Calculating R(D) - Gaussian source



- Reverse water-filling on independent Gaussian RVs
- Reverse water-filling on general multi-variate Gaussian RVs
- Reverse water-filling on Gaussian stochastic process

Main Theorem



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

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

is the minimum achievable rate at distortion D.

CONVERSE

Rate-distortion theorem: **CONVERSE**

We show that we cannot achieve a distortion of less than D if we describe X at a rate less than $R(D)$ given as $\min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$. We first need a lemma.

Lemma: The rate-distortion function $R(D) = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$ is a nonincreasing convex function of D.

Converse: Consider an $(2^{nR}, n)$ rate distortion code defined by functions f_n and g_n. Let $\hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n))$ be the reproduced sequence corresponding to X^n. Assume that $E[d(X^n, \hat{X}^n)] \leq D$. We thus need to show that $R \geq R(D)$. This follows as:

Achievability of R(D)

- We will skip 10.5 and go directly for an achievability proof based on **strong typicality**
- **Strong typicality** holds only for discrete alphabets and sequences.
- Why do we need it?
- To find an upper bound on the probability that a given source sequence is NOT well represented by a randomly chosen codeword. Analogous to probability of error calculations in channel coding / capacity theorems.

Two types of typicality

- Strong typicality:

• *Definition:* A sequence $x^n \in \mathcal{X}^n$ is said to be ϵ-strongly typical with respect to a distribution $p(x)$ on $\mathcal{X}$ if:
1. For all $a \in \mathcal{X}$ with $p(a) > 0$ we have

$$\left|\frac{1}{n}N(a|x^n) - p(a)\right| < \frac{\epsilon}{|\mathcal{X}|}.$$

2. For all $a \in \mathcal{X}$ with $p(a) = 0$, $N(a|x^n) = 0$.

Here $N(a|x^n)$ is the number of occurrences of the symbol a in the sequence x^n. The set of sequences $x^n \in \mathcal{X}^n$ such that x^n is strongly typical is called the *strongly typical set* and is denoted as $A_\epsilon^{*(n)}$.

- Weak typicality:

• *Definition:* The *typical set* $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}.$$

Examples of typicality

Let $\mathcal{X} = \{A, B, C\}$, $p_{\mathbf{x}} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}), n = 4, \epsilon = 0.8$.

- Is $x^n = BAAC \in A_\epsilon^{(n)}$?
- Is $x^n = BAAC \in A_\epsilon^{*(n)}$?
- Is $x^n = BBBB \in A_\epsilon^{(n)}$?
- Is $x^n = BBBB \in A_\epsilon^{*(n)}$?

Which do you think is true (intuitively for now)?

$$A_\epsilon^{(n)} \subset A_\epsilon^{*(n)} \text{ OR } A_\epsilon^{*(n)} \subset A_\epsilon^{(n)}?$$

Prove that $x^n \in A_\epsilon^{*(n)} \Rightarrow x^n \in A_\epsilon^{(n)}$.

Strong joint typicality

• *Definition:* A pair of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is said to be ϵ-strongly jointly typical with respect to a distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ if:
1. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) > 0$ we have

$$\left| \frac{1}{n} N(a, b | x^n, y^n) - p(a, b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}.$$

2. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) = 0$, $N(a, b | x^n, y^n) = 0$.

Here $N(a, b | x^n, y^n)$ is the number of occurrences of the symbol (a, b) in the sequence (x^n, y^n). The set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that (x^n, y^n) is strongly jointly typical is called the *strongly jointly typical set* and is denoted as $A_\epsilon^{*(n)}(X, Y)$.

Examples of joint typicality

Let $\mathcal{X} = \{A, B, C\}$ and $\mathcal{Y} = \{D, E\}$, with joint distribution $p(x, y)$ given as in the table

	D	E
A	$\frac{1}{3}$	$\frac{1}{12}$
B	$\frac{1}{3}$	$\frac{1}{12}$
C	$\frac{2}{12}$	$\frac{1}{12}$

- What do elements of $A_\epsilon^{*(n)}$ look like?

- Is $(A, D)(B, D), (B, E) \equiv (ABB, DDE) \in A_\epsilon^{(n)}$?
- Is $(A, D)(B, D), (B, E) \equiv (ABB, DDE) \in A_\epsilon^{*(n)}$?

Some useful Lemmas

- Strong typicality is a very powerful technique more thoroughly explored in Chapters 11 and 12. Related to the Method of Types, and useful in proving stronger results than can be obtained using weak typicality - universal source coding, rate distortion theory, large deviation theory.

Lemma: Let (X_i, Y_i) be drawn i.i.d. $\sim p(x,y)$. Then $\Pr(A_\epsilon^{*(n)}) \to 1$ as $n \to \infty$.

Lemma: Let $Y_1, Y_2, \cdots, Y_n$ be drawn i.i.d.$\sim p(y)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_1)} \leq \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)},$$

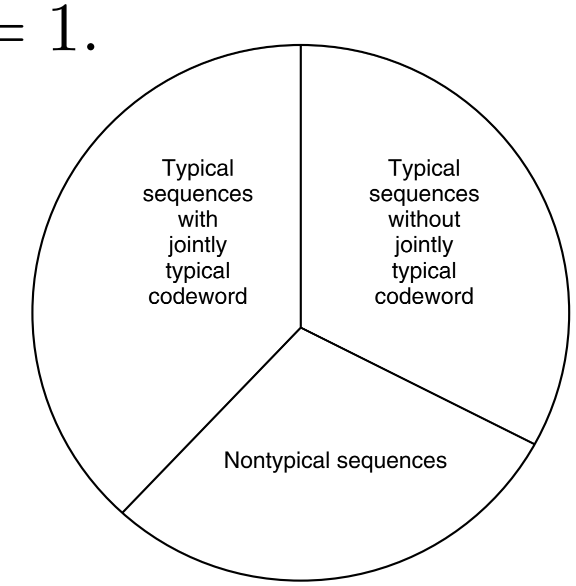
where $\epsilon_1 \to 0$ as $\epsilon \to 0$ and $n \to \infty$.

Proof of achievability

Let $X_1, X_2, \cdots, X_n$ be i.i.d.$\sim p(x)$ and let $d(x,\hat{x})$ be a bounded distortion measure for this source with rate distortion function $R(D)$. Then for any rate distortion pair (R,D) we will prove the existence of a sequence of rate distortion codes with rate R and asymptotic distortion D.

Key steps:

- Fix $p(\hat{x}|x)$ and find $p(\hat{x})$. Fix $\epsilon > 0$.
- Describe codebook generation: 2^{nR} sequences (indexed by w) $\hat{\mathcal{X}}^n$ drawn i.i.d. $\sim p(\hat{x})$.
- Describe encoding of a given sequence X^n: index X^n by w if there exists a w: $(X^n, \hat{X}^n(w)) \in A_\epsilon^{*(n)}$. If > 1, send first, else send $w = 1$.
- Decoding: reproduce $\hat{X}^n(w)$
- Calculate the distortion (see figure)
- Calculate the probability of error



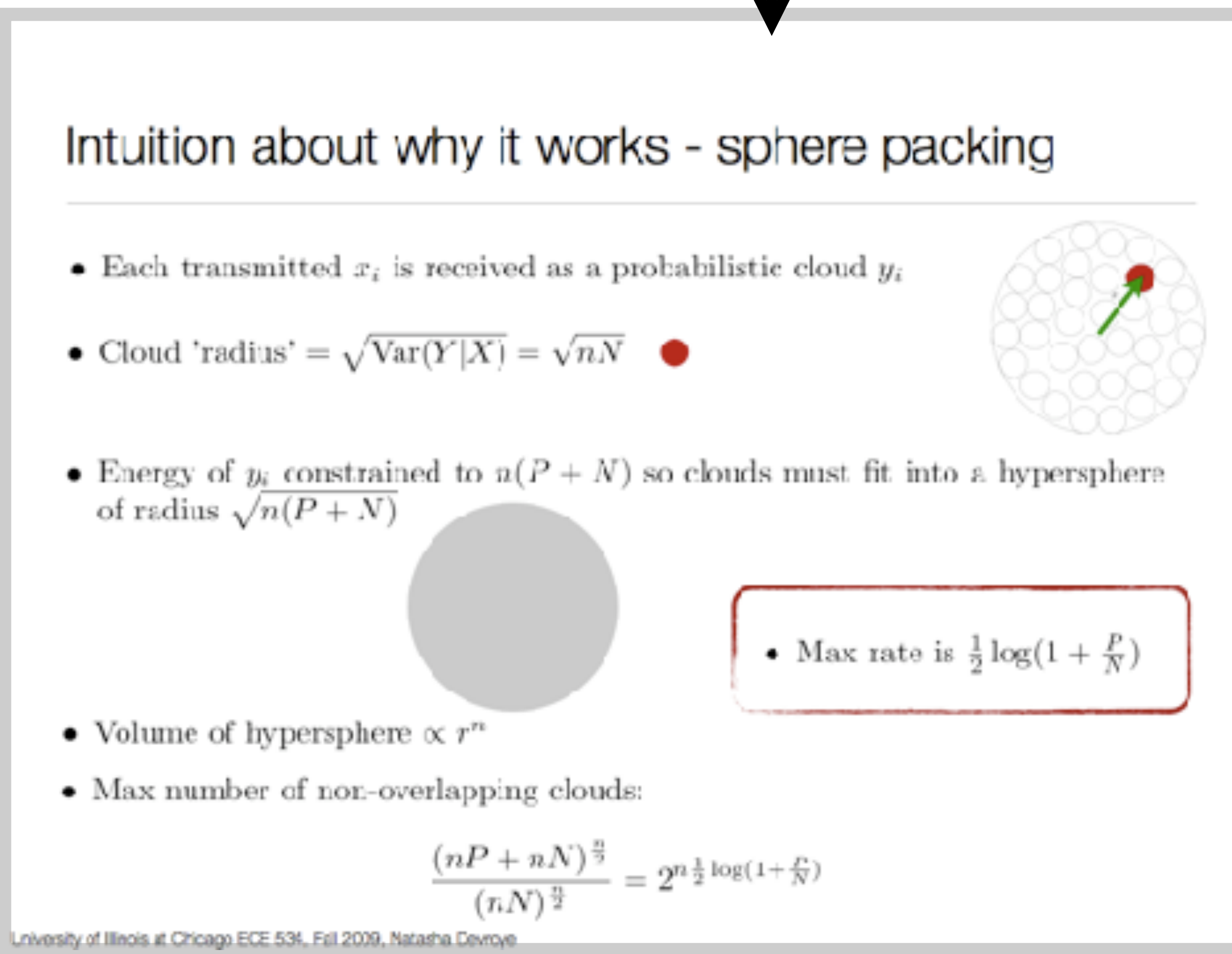
Some interesting parallels

- Channel coding
  - Random codebook generation
  - Encoding is simply lookup
  - Joint typicality decoders
  - Probability of error - decoder side
- Rate distortion
  - Random codebook generation
  - Encoding is jointly typical
  - Decoding is a lookup
  - Probability of error - encoder side

Some more interesting parallels

- Channel coding for Gaussian channel

Sphere packing

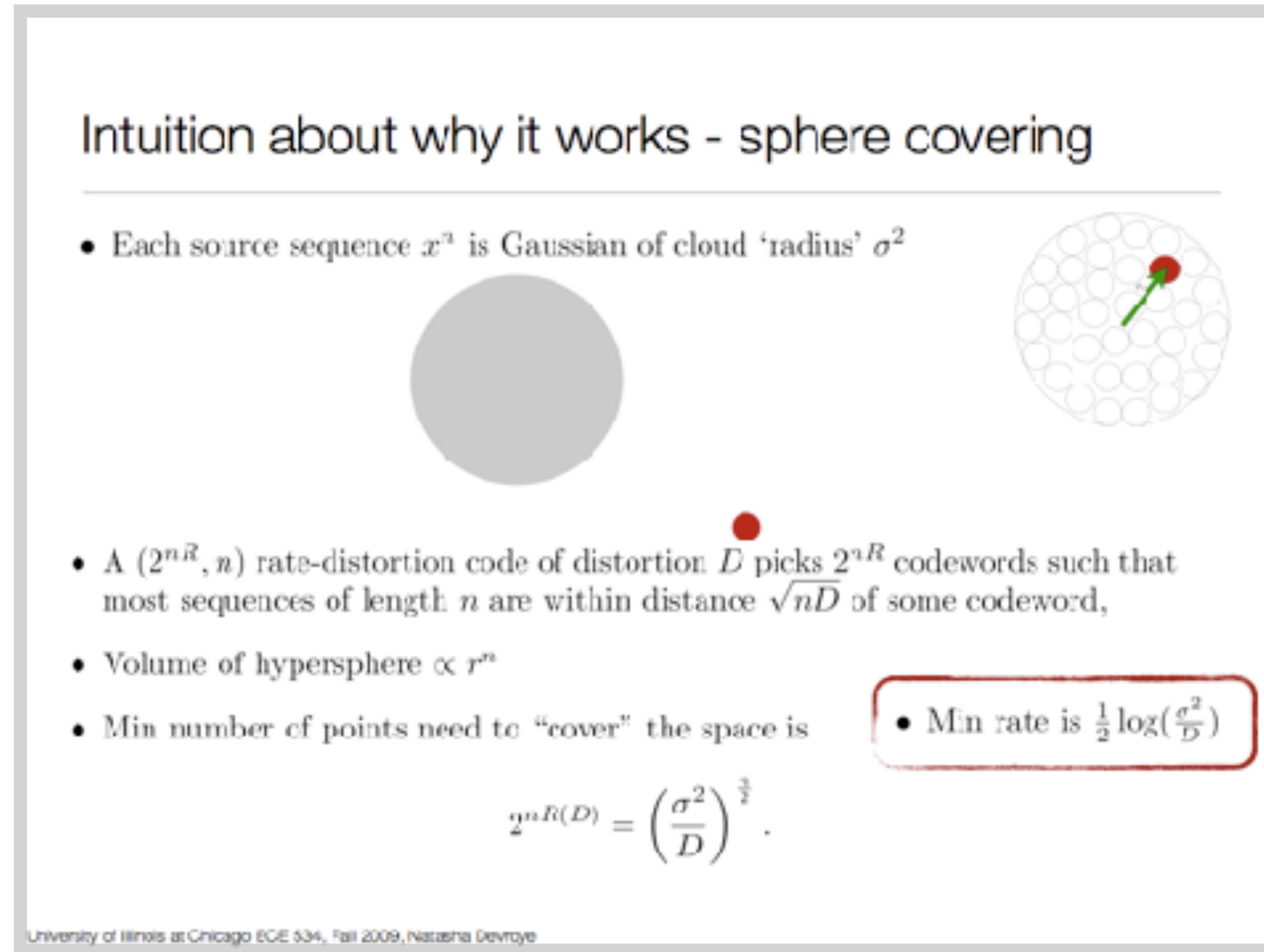


Intuition about why it works - sphere packing

- Each transmitted x_i is received as a probabilistic cloud y_i
- Cloud 'radius' $= \sqrt{\mathrm{Var}(Y|X)} = \sqrt{nN}$
- Energy of y_i constrained to $n(P+N)$ so clouds must fit into a hypersphere of radius $\sqrt{n(P+N)}$
- Max rate is $\frac{1}{2}\log(1+\frac{P}{N})$
- Volume of hypersphere $\propto r^n$
- Max number of non-overlapping clouds:

$$\frac{(nP+nN)^{\frac{n}{2}}}{(nN)^{\frac{n}{2}}} = 2^{n\frac{1}{2}\log(1+\frac{P}{N})}$$

University of Illinois at Chicago ECE 534, Fall 2009, Natasha Devroye



Intuition about why it works - sphere covering

- Each source sequence x^n is Gaussian of cloud 'radius' σ^2
- A $(2^{nR}, n)$ rate-distortion code of distortion D picks 2^{nR} codewords such that most sequences of length n are within distance $\sqrt{nD}$ of some codeword,
- Volume of hypersphere $\propto r^n$
- Min number of points need to "cover" the space is

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

- Min rate is $\frac{1}{2}\log(\frac{\sigma^2}{D})$

University of Illinois at Chicago ECE 534, Fall 2009, Natasha Devroye

- Rate-distortion for Gaussian channel

Sphere covering