Information and Communication Theory

Lecture 6

Rate Distortion Theory

Mário A. T. Figueiredo

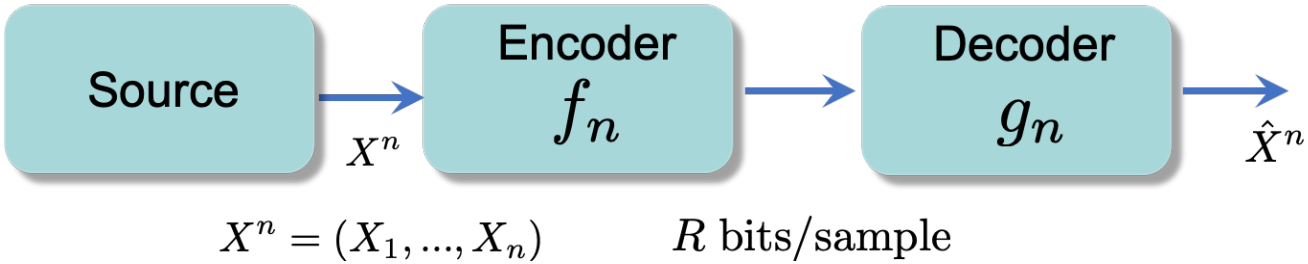
DEEC, Instituto Superior Técnico, University of Lisbon, **Portugal**

2023

Rate-Distortion Theory

- Question 1: if you want to encode a discrete source X using less than $H(X)$ bits/symbol, what is the best that can be done?
- Question 2: if you want to encode a continuous (*e.g.*, Gaussian) source with a finite number of bits, what is the best that can be done?
- Rate-distortion (R-D) theory addresses these questions.
- R-D theory studies lossy coding.
- R-D theory requires a definition of "best", a distortion/loss measure.
- Assumption on the channel: rate limitation, but no errors.

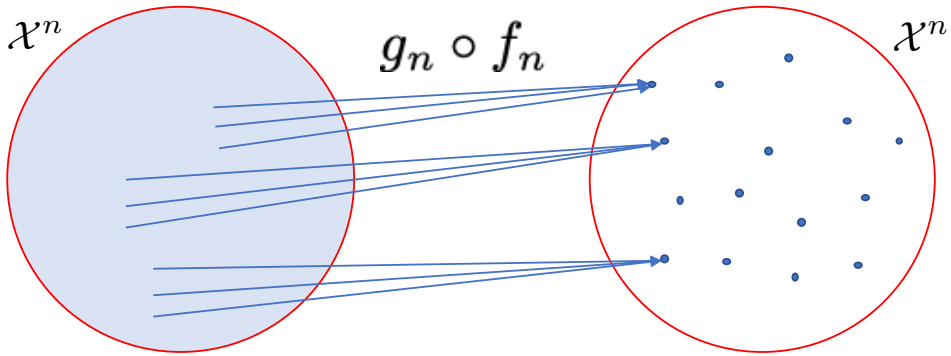
Rate-Distortion Theory: Setting



- Memoryless source $X \in \mathcal{X}$, discrete or continuous (*e.g.*, $\mathcal{X} = \{0, 1\}$ or $\mathcal{X} = \mathbb{R}$)
- Order n extension: $X^n = (X_1, ..., X_n) \in \mathcal{X}^n$.
- Encoder: $f_n : \mathcal{X}^n \to \{0, 1, ..., 2^{nR} - 1\}$, using R bits/sample.
- Decoder: $g_n : \{0, 1, ..., 2^{nR} - 1\} \to \mathcal{X}^n$.
- Assuming g_n is deterministic, this is called a quantizer, with codebook

$$\mathcal{C} = \{g_n(0), ..., g_n(2^{nR} - 1)\} \subset \mathcal{X}^n$$

Lossy Coding



- In lossy coding, $f_n \circ \gamma_n$ is not injective.
- If $\mathcal{X} = \mathbb{R}$, coding is necessarily lossy.
- For $\mathcal{X} = \mathbb{R}$, this is called vector quantization (VQ).
- VQ is characterized by a codebook $\mathcal{C} = \{y_0, ..., y_{M-1}\} \subset \mathbb{R}^n$ ($M = 2^{nR}$) and the quantization regions/cells:

$$R_i = \{x^n \in \mathbb{R}^n : g_n(f_n(x^n)) = y_i\} = \{x^n \in \mathbb{R}^n : f_n(x^n) = i\}$$

- Clearly, the regions are a partition of $\mathbb{R}^n$.

Definitions: Distortion

- The pair (f_n, g_n) (as defined above) is called a $(2^{nR}, n)$ code.
- Distortion measure: $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Standard distortion measures:
  - ✓ Hamming ($\mathcal{X} = \{0, 1\}$): $d(x, \hat{x}) = d_H(x, \hat{x}) = \begin{cases} 0, & \text{if } x = \hat{x} \\ 1, & \text{if } x \neq \hat{x}. \end{cases}$
  - ✓ Squared-error ($\mathcal{X} = \mathbb{R}$): $d(x, \hat{x}) = (x - \hat{x})^2$.
- Distortion between sequences: $d : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}_+$:

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

- Expected distortion for $(2^{nR}, n)$ code (f_n, g_n):

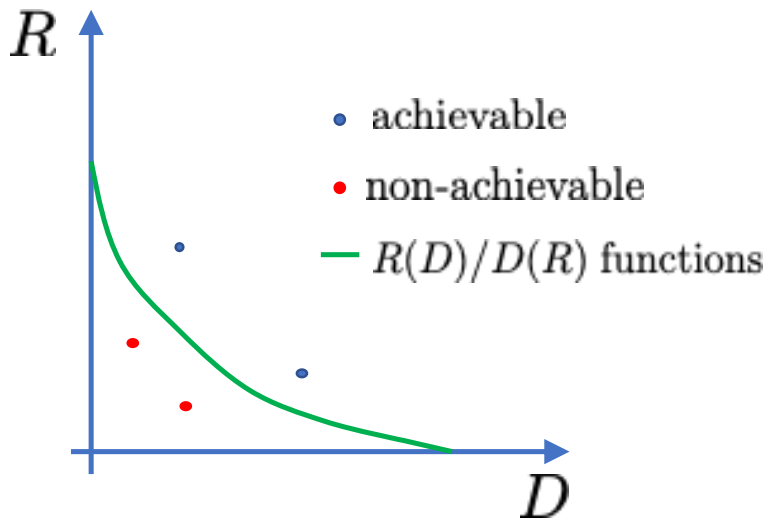
$$D = \mathbb{E}_{X^n} \Big[d\big(X^n, g_n(f_n(X^n))\big) \Big]$$

Achievability and Rate-Distortion Functions

- A rate-distortion pair (R, D) is achievable if there exists a sequence of $(2^{nR}, n)$ codes (f_n, g_n) such that

$$\lim_{n\to\infty} \mathbb{E}_{X^n}\left[d\big(X^n, \underbrace{g_n(f_n(X^n))}_{\hat{X}^n}\big)\right] \leq D.$$

- The rate-distortion region is the set of achievable rate-distortion pairs.
- The rate-distortion function: $R(D) = \inf\{R : (R, D) \text{ is achievable}\}$.
- The distortion-rate function: $D(R) = \inf\{D : (R, D) \text{ is achievable}\}$.



Information Rate-Distortion Function

- Source X, with probability density of mass function f_X.
- The information rate-distortion function $R^{(I)}(D)$, for source X, with distortion measure d is defined as

$$R^{(I)}(D) = \min_{f_{\hat{X}|X}} I(X;\hat{X})$$
$$\text{subject to } \mathbb{E}_{X,\hat{X}}[d(X,\hat{X})] \leq D$$

- Discrete source:

$$\mathbb{E}_{X,\hat{X}}[d(X,\hat{X})] = \sum_x \sum_{\hat{x}} f_X(x) f_{\hat{X}|X}(\hat{x}|x)\, d(x,\hat{x}).$$

- Continuous source:

$$\mathbb{E}_{X,\hat{X}}[d(X,\hat{X})] = \iint f_X(x) f_{\hat{X}|X}(\hat{x}|x)\, d(x,\hat{x})\, d\hat{x}\, dx.$$

- Main R-D theorem: $R(D) = R^{(I)}(D)$.

Information R-D Function: Binary Source

- Source $X \in \{0, 1\}$, with $\mathbb{P}(X = 1) = p$; Hamming distortion.
- Expected Hamming distortion: $\mathbb{E}[d(X, \hat{X})] = \mathbb{P}(X \neq \hat{X})$.

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= H(p) - H(X + \hat{X}|\hat{X}) \quad (+ \text{ is binary addition}) \\ &\geq H(p) - H(X + \hat{X}) \quad (\text{conditioning reduces entropy}) \\ &\geq H(p) - H(D) \end{aligned}$$

...under the constraint $\mathbb{P}[X + \hat{X} = 1] \leq D$, since $H(D, 1 - D)$ decreases with D, if $D < 1/2$.

- Clearly, for $D \geq p$, fix $\hat{X} = 0$, leading to

$$\mathbb{P}[X + \hat{X} = 1] = \mathbb{P}[X = 1] = p \quad \text{and} \quad I(X; \hat{X}) = 0.$$

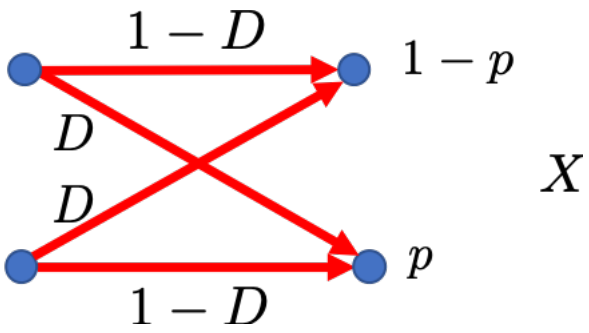
- Next step: show we can have $I(X; \hat{X}) = H(p) - H(D)$, for $D < p$.

Information R-D Function: Binary Source

- Show we can have $I(X;\hat{X}) = H(p) - H(D)$, for $D < p$.
- Use a binary symmetric channel.

$$\mathbb{P}[\hat{X} = 0] = \frac{1-p-D}{1-2D}$$

$$\mathbb{P}[\hat{X} = 1] = \frac{p-D}{1-2D}$$



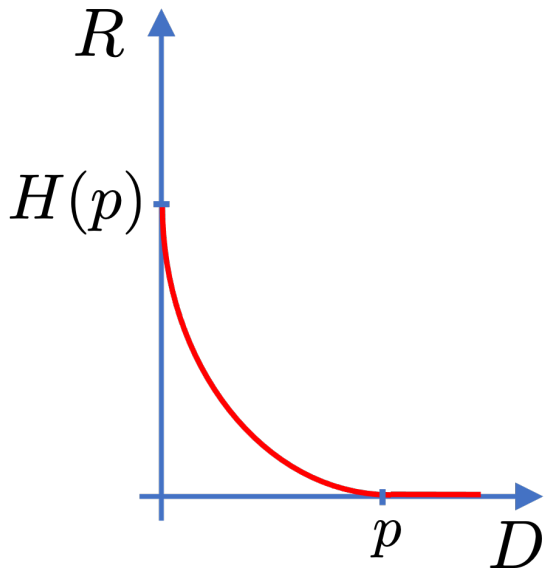
- Exercise: check that these probabilities for $\hat{X}$ lead to $\mathbb{P}(X = 1) = p$.
- With this choice: $I(X;\hat{X}) = H(X) - H(X|\hat{X}) = H(p) - H(D)$.
- Conclusion: for $D < p$,

$$R^{(I)}(D) = H(p) - H(D)$$

Information R-D Function: Binary Source

- Information rate-distortion function for the binary source:

$$R^{(I)}(D) = \begin{cases} H(p) - H(D), & \text{if } D < p \\ 0 & \text{if } D \geq p. \end{cases}$$



- For $D = 0$, lossless coding; entropy $H(p)$ is the rate lower bound.

Binary Source: Example 1

- Binary source with $p = 1/2$, thus $H(p) = 1$.
- Example of $(2^{nR}, n) = (2^1, 2)$ code, *i.e.*, with $R = 1/2$, $n = 2$.

$$f_2(00) = 0, f_2(01) = 0, f_2(11) = 1, f_2(10) = 1$$
$$g_2(0) = 00, g_2(1) = 11$$

- Hamming distortion:

$$D = \mathbb{E}\Big[\frac{1}{2} d_H\big(X^2, g_2(f_2(X^2))\big)\Big]$$
$$= \Big(\frac{1}{2}\Big) \frac{d_H(00,00) + d_H(01,00) + d_H(11,11) + d_H(10,11)}{4} = \frac{1}{4}$$

- For $D = \frac{1}{4}$, we have $R^{(I)}(\frac{1}{4}) = 1 - H(\frac{1}{4}) = \frac{3}{4}\log_2 3 - 1 \simeq 0.19$.
- $R = \frac{1}{2}$ is ≈ 2.6 times higher than the R-D bound $R^{(I)}(\frac{1}{4}) \simeq 0.19$.

Can we do better?

Binary Source: Example 2

- Binary source with $p = 1/2$, thus $H(p) = 1$.
- Example of $(2^{nR}, n) = (2^2, 4)$ code, *i.e.*, with $R = 1/2$, $n = 4$.

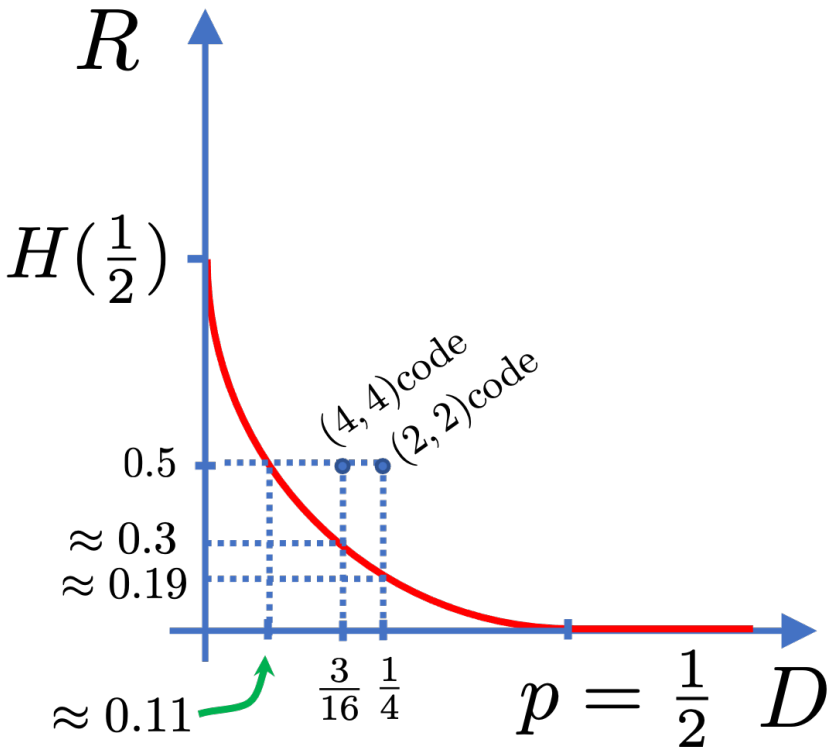
$$
\begin{aligned}
&f_4(0000) = f_4(0100) = f_4(0010) = f_4(0001) = 00\\
&f_4(0111) = f_4(0011) = f_4(0101) = f_4(0110) = 01\\
&f_4(1000) = f_4(1100) = f_4(1010) = f_4(1001) = 10\\
&f_4(1111) = f_4(1011) = f_4(1101) = f_4(1110) = 11\\
&g_n(00) = 0000, \quad g_n(01) = 0111, \quad g_n(10) = 1000, \quad g_n(11) = 1111.
\end{aligned}
$$

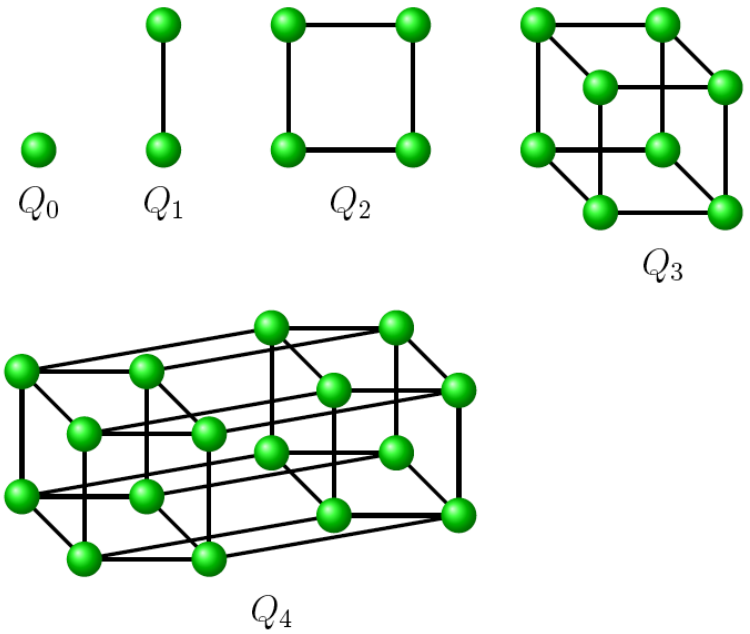
- Hamming distortion (notice that, $d_H(x^n, g_n(f_n(x^n))) \in \{0, 1\}$),

$$
D = \mathbb{E}\Big[\frac{1}{4} d_H\big(X^n, g_n(f_n(X^n))\big)\Big] = \frac{1}{4}\,\frac{4 \times 0 + 12 \times 1}{16} = \frac{3}{16}
$$

- For $D = \frac{3}{16}$, we have $R^{(I)}(\frac{3}{16}) = 1 - H(\frac{3}{16}) = \simeq 0.30$.
- $R = \frac{1}{2}$ is $\approx 1.64\times$ higher the R-D bound $R^{(I)}(\frac{2}{16}) \simeq 0.30$.

Binary Source: Summary of the Examples





- All thanks to the geometry of high-dimensional hypercubes.

Information R-D Function: Gaussian Source

- Source $X \in \mathbb{R}$, with $f_X = \mathcal{N}(0, \sigma^2)$; squared-error distortion.
- Noticing that $\mathbb{E}[(X - \hat{X})^2] \geq \mathsf{var}[X - \hat{X}]$,

$$
\begin{aligned}
I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\
&= \frac{1}{2}\log(2\pi e\sigma^2) - h(X - \hat{X}|\hat{X}) \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - h(X - \hat{X}) \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\big(2\pi e\,\mathsf{var}[X - \hat{X}]\big) \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\big(2\pi e\,\mathbb{E}[(X - \hat{X})^2]\big) \\
&\geq \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\big(2\pi e\,D\big) = \frac{1}{2}\log\Big(\frac{\sigma^2}{D}\Big)
\end{aligned}
$$

...since $\mathbb{E}[(X - \hat{X})^2] \leq D$.

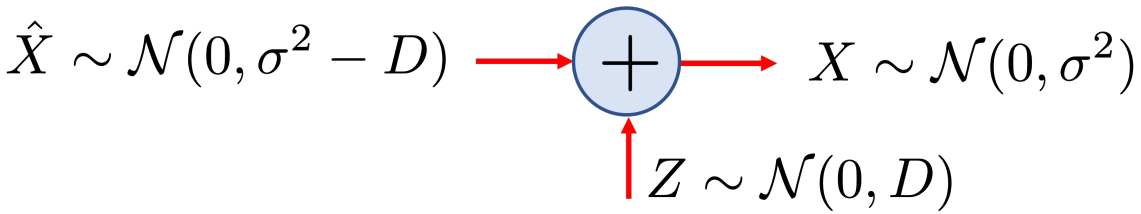
- Next step: show that we can have $I(X;\hat{X}) = \frac{1}{2}\log\Big(\frac{\sigma^2}{D}\Big)$, for $D \leq \sigma^2$

Information R-D Function: Gaussian Source

- For $D \geq \sigma^2$, fix $\hat{X} = 0$, leading to $I(X;\hat{X}) = 0$ and

$$\mathbb{E}\big[(X - \hat{X})^2\big] = \mathbb{E}\big[X^2\big] = \sigma^2 \leq D.$$

- For $D < \sigma^2$, use



...leading to

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

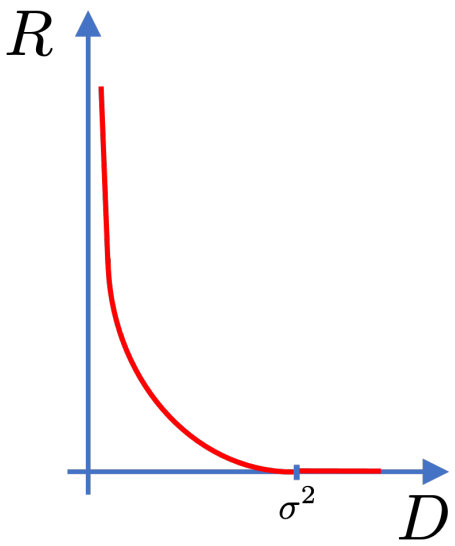
and

$$\mathbb{E}[(X - \hat{X})^2] = D.$$

Information R-D Function: Gaussian Source

- In summary, for a Gaussian source of variance σ^2

$$R^{(I)}(D) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right), & \text{if } D < \sigma^2 \\ 0 & \text{if } D \geq \sigma^2. \end{cases}$$



- Notice: $\lim_{D\to 0} \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right) = +\infty$

- Information distortion-rate function: $D^{(I)}(R) = 2^{-2R}\sigma^2$

Gaussian Source: Example

- Gaussian source: $X \sim \mathcal{N}(0, \sigma^2)$. Squared-error distortion.
- Optimal $(2^1, 1)$ code, *i.e.*, $R = 1$ and $n = 1$.
- Encoder: $f(x) = 0$, if $x \leq 0$; $f(x) = 1$, if $x > 0$.
- Decoder (optimal, *i.e.*, minimum expected distortion):

$$g_1(0) = -\sqrt{2\sigma^2/\pi}, \quad g_1(1) = \sqrt{2\sigma^2/\pi}$$

- Expected distortion (squared error):

$$D = \mathbb{E}[(X - g_1(f_1(X)))^2] = 2\int_0^\infty \left(x - \sqrt{2\sigma^2/\pi}\right)^2 \mathcal{N}(x; 0, \sigma^2)\, dx = \frac{\pi - 2}{\pi}\sigma^2$$

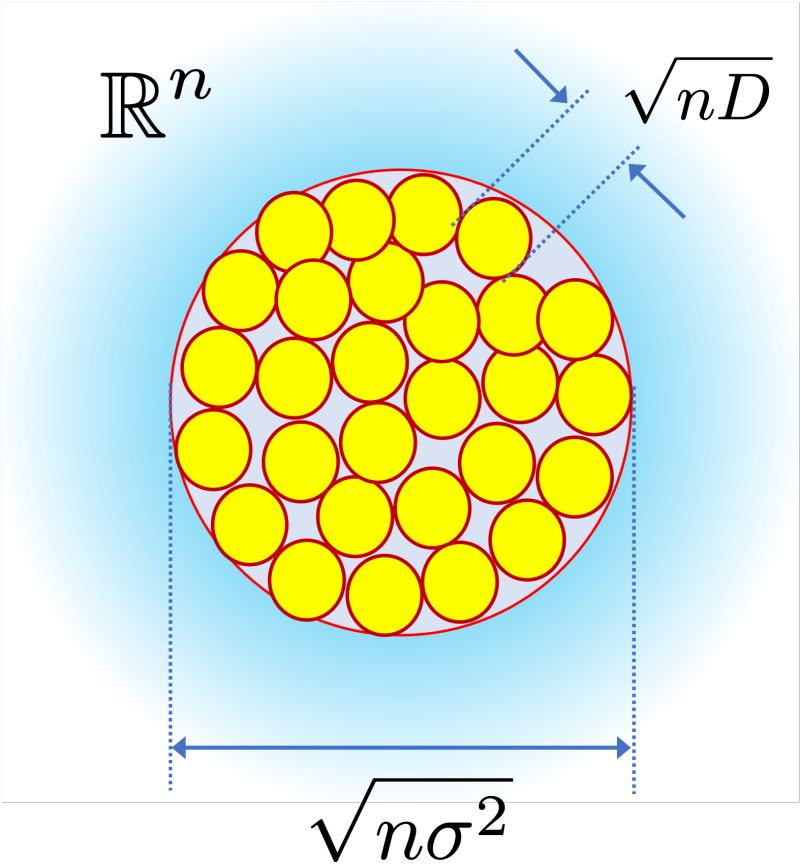
- For $R = 1$, $D^{(I)}(R) = 2^{-2}\sigma^2 = 0.25\ \sigma^2$ (compare with $\frac{\pi-2}{\pi} \simeq 0.36$)

Sphere Covering

- Geometric insight for the Gaussian source.
- Since $X \sim \mathcal{N}(0, \sigma^2)$, for large n, X^n is in a sphere of radius $\sqrt{n\sigma^2}$.
- The distortion bound means that any x^n in this sphere must be within a distance $\sqrt{nD}$ of an element of $\mathcal{C}$.
- Equivalently, any x^n must be within a sphere of radius $\sqrt{nD}$ with center at one of the elements of $\mathcal{C}$.
- The volume of a radius-r sphere is $V(r) = C_n r^n$.
- Minimum number of spheres required:

$$2^{nR} \geq \frac{(\sqrt{n\sigma^2})^n}{(\sqrt{nD})^n} = \left(\frac{\sigma^2}{D}\right)^{n/2} \quad \Leftrightarrow \quad R \geq \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right)$$

Sphere Covering



- Thus picture becomes accurate for large n, since "in high dimensions, Gaussian distributions are soap bubbles."[1]

[1] www.inference.vc/high-dimensional-gaussian-distributions-are-soap-bubble/

Recommended Reading

- T. Cover and J. Thomas, "Elements of Information Theory", John Wiley & Sons, 2006 (Sections 10.1 to 10.5, except the proofs).