Information-Theoretic System Analysis and Design

Bernhard Geiger

Graz University of Technology

Winter Term 2017/18
Lecture I: Intro and First Examples

Bernhard Geiger

Graz University of Technology

Winter Term 2017/18
Motivation – Why are we doing this?

*Information is information, not matter or energy.*

—Norbert Wiener, “Cybernetics”
Motivation – Why are we doing this?

Information [...] is that which informs. In other words, it is the answer to a question of some kind. It is also that from which data and knowledge can be derived [...].

—Wikipedia
Motivation – Why are we doing this?

Information [...] is that which informs. In other words, it is the answer to a question of some kind. It is also that from which data and knowledge can be derived [...].

—Wikipedia

The concept of information is too broad to be captured completely by a single definition. However [...] we define a quantity called entropy, which has many properties that agree with the intuitive notion of what a measure of information should be.

—Cover & Thomas, “Elements of Information Theory”
A lot of Confusion - PCA

PCA is used as a pre-processing step in dimensionality reduction.

\[
X = \begin{bmatrix}
2.33 \\
-1.13 \\
2.02
\end{bmatrix} \quad \Rightarrow \quad Y = W^T \cdot X \quad \Rightarrow \quad Y = \begin{bmatrix}
2.10 \\
-0.12 \\
0.03
\end{bmatrix}
\]

PCA preserves most of \( X \)'s variance - minimizes the mean-squared reconstruction error - BUT: the connection to information is not immediate!
PCA is used as a pre-processing step in dimensionality reduction.

\[ X = \begin{bmatrix} 2.33 \\ -1.13 \\ 2.02 \end{bmatrix} \quad \Rightarrow \quad Y = W^T \cdot X \quad \Rightarrow \quad Y = \begin{bmatrix} 2.10 \\ -0.12 \\ 0.03 \end{bmatrix} \]
A lot of Confusion - PCA

PCA is used as a pre-processing step in dimensionality reduction.

\[ X = \begin{bmatrix} 2.33 \\ -1.13 \\ 2.02 \end{bmatrix} \quad \Rightarrow \quad Y = W^T \cdot X \quad \Rightarrow \quad Y = \begin{bmatrix} 2.10 \\ -0.12 \\ 0.03 \end{bmatrix} \]

PCA
- preserves most of \( X \)'s variance
- minimizes the mean-squared reconstruction error
- BUT: the connection to information is not immediate!
A lot of Confusion - PCA

[…] PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object [the dataset] when viewed from its [...] most informative viewpoint.

and

The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information [...].

—Wikipedia
In situations where the experimenter does not know a-priori what information to keep, feature extractors can be made to incorporate unsupervised dimensionality-reduction techniques such as [PCA] to discard information while retaining most of the empirical variability.

—PhD thesis of Gustav Eje Henter
Learning Outcomes

- Work with “advanced” information measures
Learning Outcomes

- Work with “advanced” information measures
- Information measures in deterministic systems
Learning Outcomes

- Work with “advanced” information measures
- Information measures in deterministic systems
- Similarities and differences between information-theoretic and classic system design
Learning Outcomes

- Work with “advanced” information measures
- Information measures in deterministic systems
- Similarities and differences between information-theoretic and classic system design
- Design simple systems for practical problems according to information measures
Contents

- Differential entropy, information dimension, (information rate)
- Differential entropy, information dimension, (information rate)
- Information measures in (non-)linear transforms
Contents

- Differential entropy, information dimension, (information rate)
- Information measures in (non-)linear transforms
- Information loss in simple linear and non-linear transforms:
  full-wave and half-wave rectifier, quantizer, linear filter
Contents

- Differential entropy, information dimension, (information rate)
- Information measures in (non-)linear transforms
- Information loss in simple linear and non-linear transforms: full-wave and half-wave rectifier, quantizer, linear filter
- Principal components analysis: optimality regarding mean squared-error, criteria for information-theoretic optimality
Contents

- Differential entropy, information dimension, (information rate)
- Information measures in (non-)linear transforms
- Information loss in simple linear and non-linear transforms: full-wave and half-wave rectifier, quantizer, linear filter
- Principal components analysis: optimality regarding mean squared-error, criteria for information-theoretic optimality
- (Information Bottleneck Method and signal enhancement)
Contents

- Differential entropy, information dimension, (information rate)
- Information measures in (non-)linear transforms
- Information loss in simple linear and non-linear transforms: full-wave and half-wave rectifier, quantizer, linear filter
- Principal components analysis: optimality regarding mean squared-error, criteria for information-theoretic optimality
- (Information Bottleneck Method and signal enhancement)
- Resolve a lot of confusion (hopefully)
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  - Syllabus is adaptive
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  ▶ Syllabus is adaptive
  ▶ Plenty of time for discussions
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  - Syllabus is adaptive
  - Plenty of time for discussions
  - Integrated problem classes (volunteers?)
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  - Syllabus is adaptive
  - Plenty of time for discussions
  - Integrated problem classes (volunteers?)

- Homeworks
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  ▶ Syllabus is adaptive
  ▶ Plenty of time for discussions
  ▶ Integrated problem classes (volunteers?)
- Homeworks
  ▶ Two homework assignments
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  - Syllabus is adaptive
  - Plenty of time for discussions
  - Integrated problem classes (volunteers?)

- Homeworks
  - Two homework assignments
  - Work in pairs, hand in solutions
Teaching and Learning Methods

- “Lecture” (Wednesdays; schedule online)
  ▶ Syllabus is adaptive
  ▶ Plenty of time for discussions
  ▶ Integrated problem classes (volunteers?)

- Homworks
  ▶ Two homework assignments
  ▶ Work in pairs, hand in solutions

Grade = min\{1, \lfloor 0.5(\text{Homework 1}) + 0.5(\text{Homework 2}) - \text{Bonus} \rfloor \}
Information Loss (in digital systems)

$X$ and $Y$ are discrete RVs with finite supports, $y = g(X)$. Then,

$$L(X \rightarrow Y) := H(X) - H(Y).$$

“The information loss is the information at the input minus the information at the output.”
Cascade of Systems

If $Y = g(X)$ and $Z = h(Y)$, then

$$L(X \to Z) = L(X \to Y) + L(Y \to Z).$$

If $Y(\omega) = G(\omega)X(\omega)$ and $Z(\omega) = H(\omega)Y(\omega)$, then

$$\log(H \circ G)(\omega) = \log G(\omega) + \log H(\omega).$$
Quantizer Design

An $R$-level quantizer for an input with support $\mathcal{X}$ consists of two parts:

- Partitioning: $q_1: \mathcal{X} \rightarrow S$, $\text{card}(S) = R$
- Reconstruction: $q_2: S \rightarrow \mathcal{X}$

The quantizer is the cascade of both functions:

$$Q(x) = (q_2 \circ q_1)(x) = q_2(q_1(x))$$
Resolution-Constrained RD-Quantizer

- Fix \( \text{card}(\mathcal{S}) = R \)
- Find partition and reconstruction points such that

\[
\mathbb{E} ((X - Q(X))^2)
\]

is minimized.

Sub-optimal solution: Lloyd-Algorithm.
Resolution-Constrained RD-Quantizer

\[ f_X(x) \]

\[ \hat{x}_1 \hat{x}_2 \hat{x}_3 \]
Resolution-Constrained RD-Quantizer

\[ f_x(x) \]

\[ \hat{x}_1 \hat{x}_2 \hat{x}_3 \]
Resolution-Constrained RD-Quantizer

\[ f_X(x) \]

\[ \hat{x}_1 \quad \hat{x}_2 \quad \hat{x}_3 \]
Resolution-Constrained RD-Quantizer

\[ f_X(x) \]

\[ \hat{x}_1 \quad \hat{x}_2 \quad \hat{x}_3 \]

\[ b_1 \quad b_2 \]
Resolution-Constrained RD-Quantizer

\[ f_X(x) \]

\[ \hat{x}_1 \quad \hat{x}_2 \quad \hat{x}_3 \]

\[ b_1 \quad b_2 \]
Resolution-Constrained RD-Quantizer

Let $b_1$ and $b_2$ be the thresholds for quantization. The quantized values $\hat{x}_1$, $\hat{x}_2$, and $\hat{x}_3$ are determined by the input $x$ and the function $f_X(x)$. The quantization process can be represented as:

$$f_X(x)$$
Resolution-Constrained RD-Quantizer
Maximum Output Entropy Quantizer

- Fix $\text{card}(S) = R$
- Find partition and reconstruction points such that

$$H(Q(X)) = H(q_1(X))$$

is maximized.

One possible solution: Quantile quantizer (choose intervals to be the $R$-quantiles)
Maximum Output Entropy Quantizer

\[ f_X(x) \]

\[ H(Q(X)) = 1.53 \text{ vs. } H(Q(X)) = 1.58 \]