Uncertainty, Information and Learning Mechanisms (Part 1)

Intelligence for Embedded Systems

Ph. D. and Master Course

Manuel Roveri

Politecnico di Milano, DEIB, Italy

Uncertaintanty

- The real world is prone to **uncertainty**
- At different level of the data analysis process
  - Acquiring data
  - Representing information
  - Processing the information
  - Learning mechanisms

Part 1 of this lecture

Part 2 of this lecture

- To formalize the concept of uncertainty we need to define an «uncertainty-free» entity and a way to evaluate the error w.r.t. this entity

From errors to perturbations

- We have **uncertainty any time we have an approximated entity** which, to some extent, estimates the ideal -possibly unknown- one.
- Such a situation can be formalized by introducing the ideal uncertainty-free entity and the real uncertainty-affected one and evaluating the error: **the discrepancy between the two according to a suitable figure of merit**.
- The error is strictly dependent on a specific pointwise instance: **we abstract the pointwise error with the concept of perturbation**

From errors to perturbations (2)

- **A generic perturbation δ*A*** intervenes on the computation by **modifying** the status assumed by an entity from its nominal configuration A to a perturbed one A_p
- **The effect induced by the perturbation** can be evaluated through a suitable figure of merit $/\!/ A , A_p /\!/$ measuring the discrepancy between the two states.
- *Example*: *a real sensor providing the constant value* $a \in R$
  - the discrepancy between the ideal nominal value and the perturbed one can be expressed as the error

    $$/\!/ A,A_p /\!/ = e = |a_p - a|$$
  - the error would assume a different value with different instances of the perturbed acquisition a_p

Modeling the uncertainty

- The mechanism inducing uncertainty can be modeled with the signal plus noise model

$$a_p = a + \delta_a$$

and

$$\| A, A_p \| = |a_p - a| = |\delta_a| = |e|$$

- δ_a can be described in many cases as a random variable with its probability density function fully characterizing the way uncertainty disrupt the information.

General signals and perturbations

The signal $\psi \in \Psi \subset \mathbb{R}^d$

The perturbation $\delta\psi$ drawn from distribution $f_\psi(M, C_{\delta_\psi})$

- Discrete or continuous
- Mean M
- Covariance C_{δ_ψ}

Perturbation operator

Perturbed signal $\psi + \delta\psi \Rightarrow \psi_p$

Examples of perturbations

Continuous perturbations

$$\Pr(\delta\psi = \delta\bar{\psi}) = 0, \forall \psi \in \Psi$$

Acute perturbations

$$\delta A = \delta A(\delta\psi) \Rightarrow \lim_{A_p \to A} rank(A_p) = rank(A)$$

Perturbations

At representational level:

- Natural numbers
- Integer numbers
- Rational and reals

During the computational flow:

- Linear function
- Nonlinear function

Perturbations at the data representation level

- Numerical data acquired by sensors and digitalised through an ADC are represented as a sequence of bits coded according to a given transformation which depends on the **numerical information we need to represent**.
- We now introduce the main transformations used in numerical representations as well as the types and **characterization of uncertainty introduced when representing data in a digital format**:
  - Projection
  - Truncation
  - Rounding

Natural Numbers:
exact representation and uncertainties

Assume we are willing to spend n bits to represent a finite value a ∈ N. It immediately comes out that we can represent only numbers belonging to a subset N(n) ⊂ N given the finiteness of n.

$$
n \text{ bits} \quad \Longrightarrow \quad
\begin{array}{l}
\mathbb{N}(n) \subset \mathbb{N} \\
\mathbb{N}(n) = \{0, 1, 2, \cdots, 2^n - 1\}
\end{array}
$$

Exact representation

- Uncertainty introduced by projection, truncation or rounding i.e.,

  removing $q \leq n$ bits

Natural Numbers: projection to a subspace

- A projection to a lower dimensional space is achieved by simply setting to zero the least significant $n - q$ bits of the n bits codeword associated with a (the least significant q bits are set to zero leading to value $a(q)$).

Original (n=4 bits)	Projected to n-q=2 bits (q=2)
0000	0000
0001	0000
0010	0000
0011	0000
0100	0100
…	…

- The projection introduces an absolute error

$$\boldsymbol{e(q) = a - a(q) < 2^q}$$

Natural Numbers: truncation

- Truncation operates as a chopping operator that removes the least significant *q* bits

Original (n=4 bits)	Truncation to n-q=2 bits (q=2)
0000	00
0001	00
0010	00
0011	00
0100	01
…	…

- The projection introduces an absolute error

$$e(q) = a - 2^q a(q) < 2^q$$

Natural Numbers: rounding

- **Rounding of a positive number truncates** the *q* least significant bits and
  - **adds** 1 to the unchopped part if and only if the most significant bit of the truncated segment is 1.
  - **otherwise**, the rounded value is the one defined over the *n* − *q* bits

Original (n=4 bits)	Rounding to n-q=2 bits (q=2)
0000	00
0001	00
0010	01
0011	01
0100	01
…	…

Perturbation at the data level: integer numbers

Use of 2cp notation $a_{2cp} = \begin{cases} a_{b,n} \text{ for } a \geq 0 \\ (2^n - |a|)_{b,n} \text{ for } a < 0 \end{cases}$

Truncation

$$f_\psi(M, C_{\delta_\psi}) \sim U\left([0, 2^q)\right)$$

Biased approximate representation

Rounding

$$f_\psi(M, C_{\delta_\psi}) \sim U\left([-2^{q-1}, 2^{q-1})\right)$$

Unbiased approximate representation

Perturbation at the data level: the fixed point representation

$$a \in \mathbb{Q} \longrightarrow a(n) \begin{cases} l & \text{bits assigned to the integer part} \\ k & \text{bits assigned to the fractional one} \end{cases}$$

$a(n)2^k$ is integer

Example: fixed point representation

$$a = 1.56 \qquad \boxed{l = 3} \qquad [\boxed{001}\boxed{10}] \quad a(n) = 1.5$$

$$\boxed{k = 2}$$

$$|e(q)| = |a - a(n)| = 0.06 < 2^{\boxed{-2}}$$

Many sources of uncertainties at the data representation level but …

- … the question is:

“What is the effect of these uncertainties within the propagation flow?”

Sensitivity Analysis

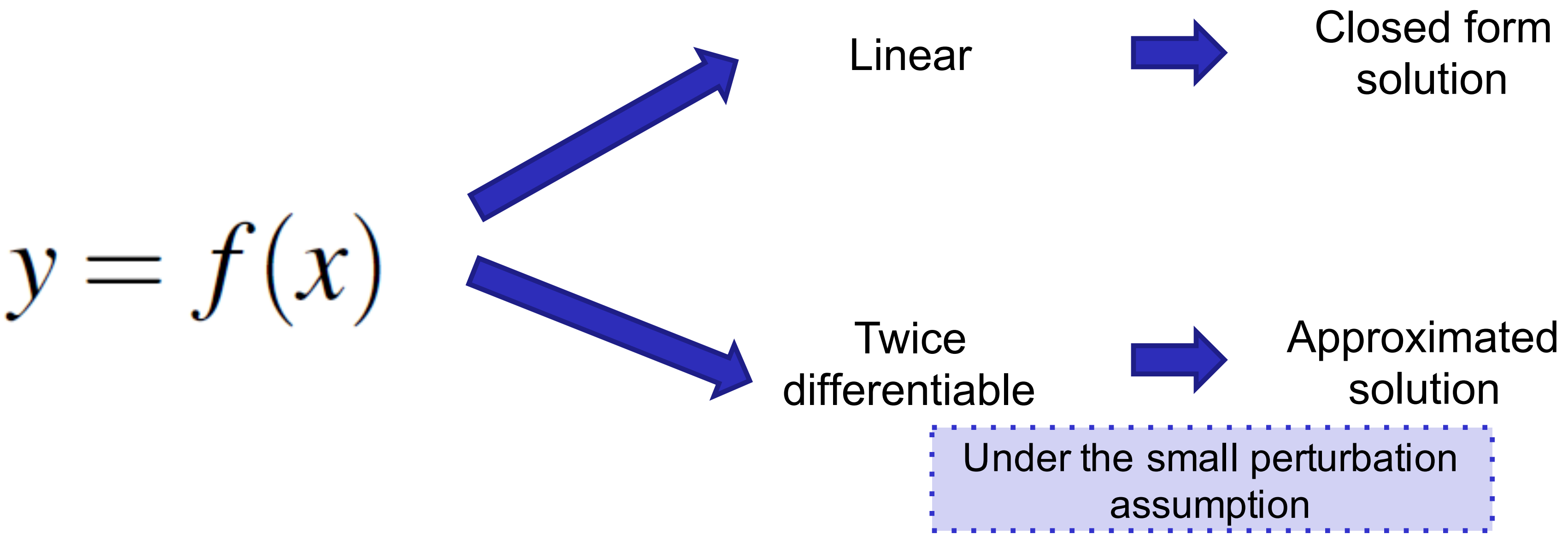
Sensitivity Analysis

- The **sensitivity analysis** provides
  - ✓ **closed form expressions for the linear function case**
  - ✓ **approximated results for the non linear one**, provided that the perturbations affecting the inputs are small in magnitude compared to the inputs (**small perturbation hypothesis**)
- The analysis of *Perturbations in the large* i.e., perturbations of arbitrary magnitude, for the nonlinear case, cannot be obtained in a closed form unless $y = f(x)$ assumes a particular structure and has properties that make the mathematics amenable.

Sensitivity Analysis: the computational flow

Measurements $x \in X \subset \mathbb{R}^d$

Output $y \in Y \subset \mathbb{R}$



$$y = f(x)$$

Linear

Closed form solution

Twice differentiable

Approximated solution

Under the small perturbation assumption

Linear Function: additive perturbation

$$y = f(x) = \theta^T x \qquad \theta \in \Theta \subset \mathbb{R}^d$$

Parameters of the linear function

$$x_p = x + \delta x \quad \longrightarrow \quad y_p = \theta^T x_p$$

Point-wise error: $$\delta y = y_p - y = \theta^T \delta x$$

Assume perturbation distribution $$f_\psi(M, C_{\delta_\psi}) = f_{\delta x}(0, C_{\delta x})$$

$$\left.\begin{array}{l} \bullet \quad E_{\delta x}[\delta y] = E_{\delta x}[\theta^T \delta x] = \theta^T E_{\delta x}[\delta x] = 0 \\ \bullet \quad Var(\delta y) = E_{\delta x}[\theta^T \delta x \delta x^T \theta] = \theta^T E_{\delta x}[\delta x \delta x^T]\theta = \\ \qquad = \theta^T C_{\delta x}\theta = trace\left(\theta^T \theta C_{\delta x}\right) \end{array}\right\} \textit{Mean and variance of the error}$$

Linear Function: additive perturbation (2)

If C_{δ_ψ} is diagonal, i.e., independence assumption on the perturbations

Squared i-th component of vectoθ

$$Var(\delta y) = \sum_{i=1}^{d} \theta_i^2 \sigma_{\delta x,i}^2$$

i-th diagonal component of covariance matrix

If all the components have the same $\sigma_{\delta x}^2$ variance

$$Var(\delta y) = \sigma_{\delta x}^2 \theta^T \theta$$

But…

“… what about the pdf of the error?”

How to get the pdf of the error?

- The pdf of the propagated error **cannot be evaluated a priori in a closed form unless we assume that the dimension *d* is large enough**.
- In such a case, we can invoke the **Central Limit Theorem (CLT) under the Lyapunov assumptions** and δ*y* can be modeled as a random variable drawn from a **Gaussian distribution**.

Central Limit Theorem under the Lyapunov Condition

Let $Y_i, i = 1 \ldots d$ a set of independent random variables characterized by finite expected value $E[Y_i]$ and variance $Var(Y_i)$. Denote $s_d^2 = \sum_{i=1}^d Var(Y_i)$ and $Y = \sum_i Y_i$. If there exists number $l > 0$ such that

$$\boxed{\lim_{d\to\infty} \left(\frac{1}{s_d^{2+l}} \sum_{i=1}^{d} E\left[|Y_i - E[Y_i]|^{2+l} \right] \right) = 0,}$$

Convergence of the moments

then $Z = \frac{(Y - E[Y])}{\sqrt{Var(Y)}}$ converges to the standard normal distribution.

W.r.t. the standard CTL here we have hypotheses on the moments but we do not require $Y_{i,}$ i=1,..,d to be identically distributed

CLT under the Lyapunov Condition (2)

- From the intuitive point of view, the central limit theorem tells us that the sum of many, not-too-large and not-too-correlated random terms, average out.
- The Lyapunov condition is one way for quantifying the not-too-large term request by inspecting the behaviour on some $2 + l$ moments.
- In most of cases, one tests the satisfaction of the condition for $l = 1$ or 2.

CLT under the Lyapunov Condition (3)

- From the theorem, with the choice $Y_i = \theta_i \delta x_i$, δ*y* can be approximated as a Gaussian random variable

  $$\delta y = \mathcal{N}(0, \sum_{i=1}^{d} \theta_i^2 \sigma_{\delta x,i}^2)$$

  and, when the variances are identical

  $$\delta y = \mathcal{N}(0, \sigma_{\delta x}^2 \theta^T \theta)$$

- It is easy to show that the Lyapunov condition holds if each component of random variable δx is uniformly distributed within a given interval, as it happens in many application cases (**think of the error distribution introduced by the rounding and truncation operators operating on binary 2cp codewords**).

Linear Function: multiplicative perturbation

$$x_p = x(1+\delta x) \longrightarrow y_p = \theta^T x_p$$

Element-wise multiplication

Point-wise error: $\delta y = y_p - y = \theta^T (x \circ \delta x)$

Assume perturbation distribution $f_\psi(M, C_{\delta_\psi}) = f_{\delta x}(0, C_{\delta x})$

Assume input distribution $f_x(0, C_x)$

- $E_{x,\delta x}[\delta y] = E_{x,\delta x}[\theta^T x \circ \delta x] = \theta^T E_x[x] \circ E_{\delta x}[\delta x] = 0$
- $Var(\delta y) = E_{x,\delta x}[\theta^T x x^T \circ \delta x \delta x^T \theta] = \theta^T \boxed{C_x} \circ C_{\delta x} \theta$

Mean and variance of the error

When the variances are identical

$$Var(\delta y) = \boxed{\sigma^2_{\delta x}} \sigma^2_x \theta^T \theta$$

Perturbations of Nonlinear function

$$y = f(x)$$

Nonlinear function modeling the computational flow

$$x_p = x + \delta x \longrightarrow y_p = f(x_p)$$

Point-wise error: $\delta y = f(x_p) - f(x)$

Small perturbation hypothesis → Second order Taylor expansion around x

$$f(x+\delta x) = f(x) + J(x)^T \delta x + \frac{1}{2}\delta x^T H(x)\delta x + o(\delta x^T \delta x)$$

Jacobian vector $J(x) = \frac{\partial f(x)}{\partial x}$

Hessian matrix $H(x) = \frac{\partial^2 f(x)}{\partial x^2}$

Nonlinear function (2)

- By discarding the terms of order larger than two, the perturbed propagated output takes the form of

$$\delta y = J(x)^T \delta x + \frac{1}{2}\delta x^T H(x)\delta x$$

- **Not much more can be said within a deterministic framework** unless we introduce strong assumptions on *f* (*x*) or δ*x*.
- However, by **moving to a stochastic framework**, which considers *x* and δ*x* mutually i.i.d random variables drawn from distributions $f_x(0, C_x)$ and $f_{\delta x}(0, C_{\delta x})$, respectively, the **first two moments of the distribution of δ*y* can be computed**

Nonlinear function (3)

- Under the above assumptions and by taking expectation w.r.t. *x* and δx, the expected value of the perturbed output

$$E[\delta y] = \frac{1}{2}E[\delta x^T H(x)\delta x] = \frac{1}{2}trace\left(E[H(x)\delta x\delta x^T]\right) = \frac{1}{2}trace\left(E[H(x)]C_{\delta x}\right)$$

Quasi-Newton approximation →

$$H(x) = \frac{\partial f(x)}{\partial x}\frac{\partial f(x)}{\partial x}^T$$

- $E[\delta y] = \frac{1}{2}trace\left(C_{fx}C_{\delta x}\right)$
- $Var(\delta y) = E\left[J(x)^T\delta x\delta x^T J(x)\right] = trace\left(E\left[J(x)J(x)^T\right]C_{\delta x}\right)$

Let’s play with MATLAB

- Download the examples related to Lecture 3
- In the ZIP file:
  - Example 3_A.m
    - About Projection and Truncation of Natural Numbers
  - Example 3_B.m
    - About the Central Limit Theorem (CLT) under the Lyapunov assumptions