Information Theory for Data Science

Assignment 2 – Part A

Prof. Giorgio Taricco

Politecnico di Torino – DET

Assignment 2 rules

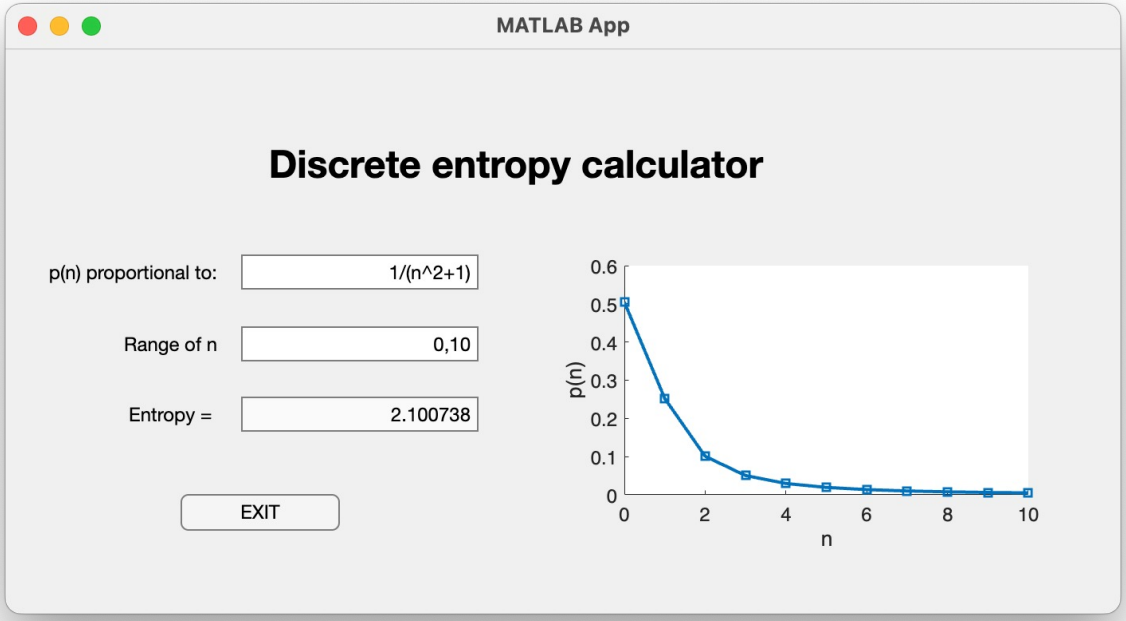
- No cooperation outside of the group is allowed
- Reports must describe the code, the numerical results, and all the analytic steps when required
- Reports must be well written and organized, provide all the implementation details
- Quality is an important element for the evaluation
- MATLAB (with AppDesigner) is the **preferred tool**
- **GUIs must start with initialized editable input data, not empty cells, to facilitate grading**

Delivery rules (Assignment 2)

- All the files relevant to the assignment (part a and b together) must be collected in a single compressed ZIP file
- The file must be uploaded through POLITO's WEB PORTAL
- **Emails deliveries are not considered in any case**
- The ZIP files must have a name "group<number>.zip"
- The group number will be published before the time you can submit the assignment
- The report must be prepared in latex and saved in PDF format with name "group<number>.pdf"
- **The report must be organized as regular text and not as slides**

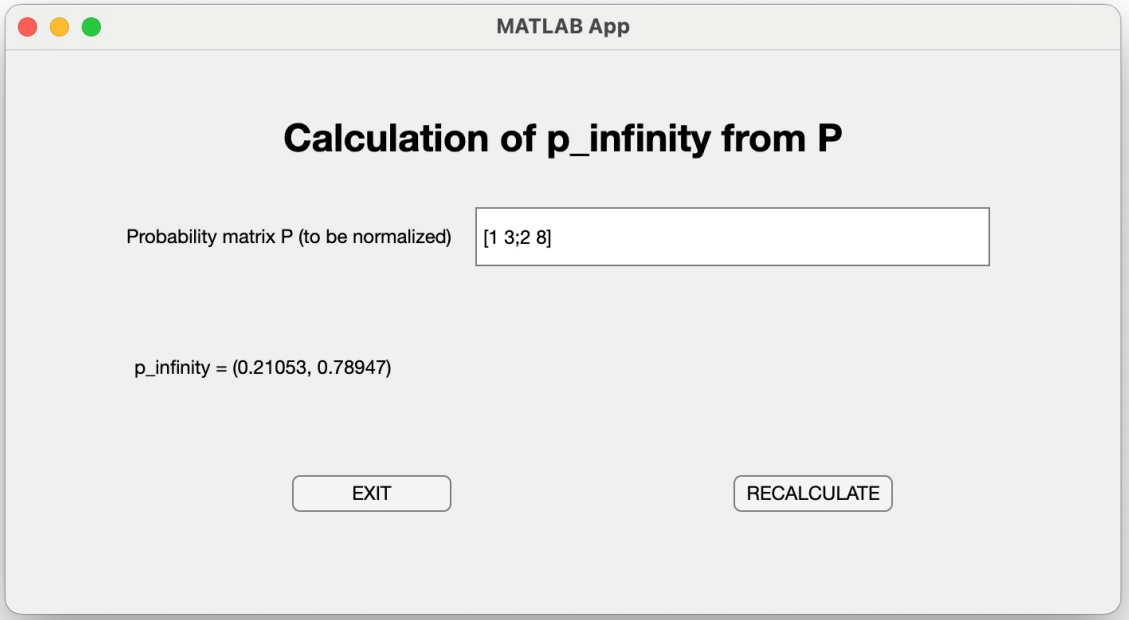
Problem 1

- Make a GUI for the evaluation of the entropy of a discrete RV



Problem 2

- Make a GUI to calculate $\boldsymbol{p}_\infty$ for a transition probability matrix $\boldsymbol{P}$



Problem 3

- For each exercise provide the analytic solution (if it is possible to find an analytic solution, and include the parameter ranges)
- Otherwise, state that the analytic solution is impossible and provide the numerical solution (the correct choice is part of the evaluation)
- Calculate the entropy of the following distributions:
  1. $p(n) \propto \exp(-\lambda n)\,, n = 0,1,2, \ldots$
  2. $p(n) \propto \exp(-n^2)\,, n = 0,1,2, \ldots$
  3. $p(n) \propto n^{-4}, n = 1,2,3, \ldots$
  4. $p(n) \propto \alpha^n, n = 1, \ldots, N$
  5. $p(n) \propto (1 + n^2)^{-k}, n = 0,1,2, \ldots$ for $k = 1,2$

Problem 4

- Design a GUI with to identify a Markov source from a text with the following input data:
  - Sample input text $x_1, x_2, \dots, x_N$ (each x_i is a a string of G characters/bytes)
  - Markov source memory M
- The transition ($x_{n:n+M-1} \rightarrow x_{n+1:n+M}$) probabilities are the frequencies:

$$p(x_{n+M}|x_{n:n+M-1}), \qquad n = 1,2, \dots, N - M$$

- For example, let $M = 3$, let $x_{n:n+2}$ =“and”. Scanning the text, we can see that $x_{n+1:n+3}$ =“nd ” occurs 33 times, “nd,” 5 times, “ndm” 2 times. Thus, the probabilities are p(“nd ” | “and”) $= 33/(33 + 5 + 2)$ etc
- Calculate in this way the probability matrix and consider only the M-symbol sequences which really appear in the text to limit the matrix size
- Generalize to symbols of G characters (bytes) (1 symbol $= G$ consecutive bytes)

Problem 4

- Use the probability matrix to calculate the entropy rate
- GUI outputs:
  - Entropy rate
  - CPU time
  - Dictionary size in bytes (approximated by the number of M-symbol sequences occurring in the text, it does not include the source code storage size)
  - Estimate of the number of bytes to describe the content, approximated by the entropy rate times $(N - M)$
  - Total size (sum of dictionary and content bytes)
  - Dominant conditional probabilities