Explainability and rationalization in decision theory: A coding theory approach

Thabang Lebese, Charles Wan, Nischal Mainali & Rongrong Xie

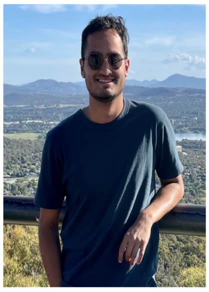
Supervisors:
Matteo Marsili (ICTP) & Isaac Pérez Castillo (UAM)

August 10, 2021

The Abdus Salam International Centre for Theoretical Physics

London Mathematical Laboratory





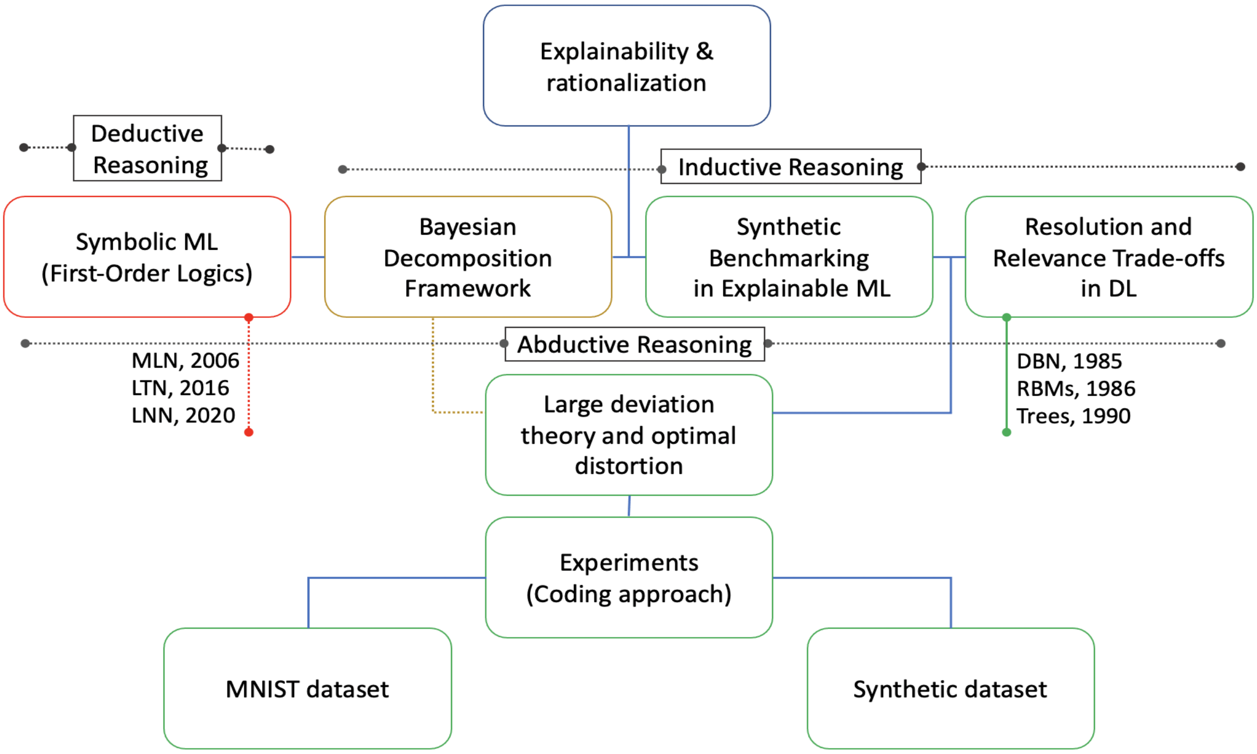

Group 8: Team members

Overview

Motivation

- The ability of clearly explaining the process that lead to a given solution is fundamental AI
- A well-known example is the 2016 (taking effect in 2018) European Union General Data Protection Regulation (EU GDPR) law
- Concrete applications includes:
  - Automated online credit or mortgage scoring,
  - E-recruiting without human intervention,
  - Automated insurance quoting, etc.
- It is fundamental to explain why a system suggests certain decisions to respect the principles of ethics and fairness
- But there seems to be a "trade-off" on **rationality** and a good **explanation**:
  - How much rationality can one retain?
  - How good enough the explanation should be?
- Hence a distortion

Approach



Work-flow diagram

Rationalization with decision theory

For a generic decision problem of outcomes s. There are S possible choices and the probability that s is an optimal choice is:

$$p_s = \frac{1}{Z} e^{\mu_s}, \qquad\qquad Z = \sum_{s=1}^{S} e^{\mu_s} \tag{1}$$

Let ℓ_s be the length of code-word that corresponds to s. For an optimal rationalisation, we have:

$$\min_{\ell} \sum_{s=1}^{S} p_s \ell_s$$

Which is known as the entropy:

$$H[p] = -\sum_{s=1}^{S} p_s \log p_s$$

But taking rational choices this way leads to choices which are hard to explain.

Large deviations theory and optimal distortion

Let q_s, p_s be probabilities of outcomes s, with rationalisation $H[q]$ and with distortion measure $D_{KL}(p||q)$

$$H[q] = -\sum_{s=1}^{S} q_s \log q_s, \qquad \min_{p:H[p]\leq H_0} D_{KL}(p||q) \tag{2}$$

From (2), we solve an optimization problem:

$$\min_p \left[D_{KL}(p||q) \pm \lambda H[p] + \nu \sum_{s=1} p_s \right], \tag{3}$$

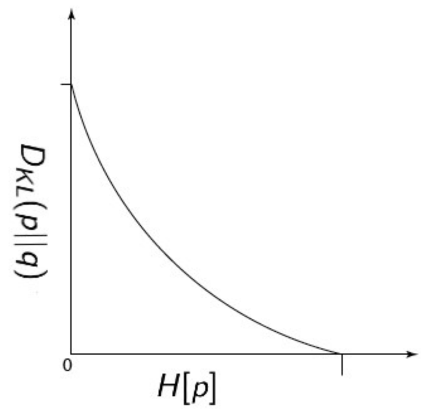
Now taking $\frac{\partial}{\partial p_s} = 0$ on (3), we get solutions:

$$p_s = \frac{q_s^{\mu}}{Z}, \qquad Z = \sum_s q_s^{\mu}, \qquad \mu = \frac{1}{1 \mp \lambda}.$$

Solutions are case-wise. For, $\lambda > 0$,$\lambda < 0$ and $\lambda \to \pm 1$.

Large deviations theory: proof of concept

- Provides a way to think about trade-offs between fidelity and compression in relaying a decision-making process.
- Given a decision-making process (or algorithm) with distribution q over outcomes
- We compress it into an explanation with distribution p.
- $D_{KL}(p||q)$ vs $H[p]$ convex with λ as the slope
- λ is the shadow price - the amount of compression that must be given up in order to achieve a certain level of fidelity.



Deep Belief Network

- A composition of Restricted Boltzmann Machines
- Learns representations of the data at decreasing scales of resolution.
- We use DBN to explore the trade-offs between accuracy and compression.
- As we go from shallow to deeper layers, original message (or decision-making process) is coarse-grained, leading to a more compressed explanation but with a distribution that is further away from the original distribution of the data set.
- According to large deviations theory, the relationship between the layers should be $p_s = \frac{q_s^{\mu}}{Z}$,
- where p is the distribution of states in the deeper layer and q is the distribution in the shallower layer.

Representations and distributions in DBN

Q: How can we compare representations between layers of a DBN and evaluate its evolution?

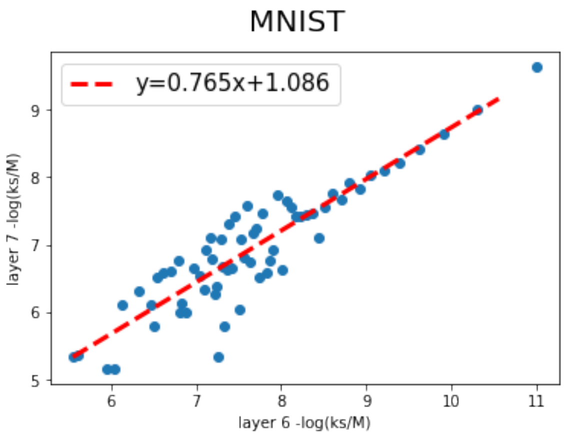
For $s \in S$, where S is the set of states over M data points, we can calculate:

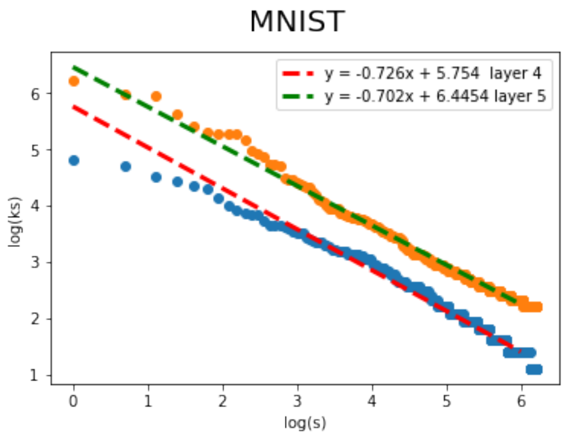
- k_s, the number of data points that take the state s.
- The statistics $\frac{k_s}{M}$ induces a distribution over states for a given layer.
- We study the evolution of this distribution across layers as predicted by the large deviations theory.

Results 1

Optimal distortion successfully predicts the behavior near the middle layer.

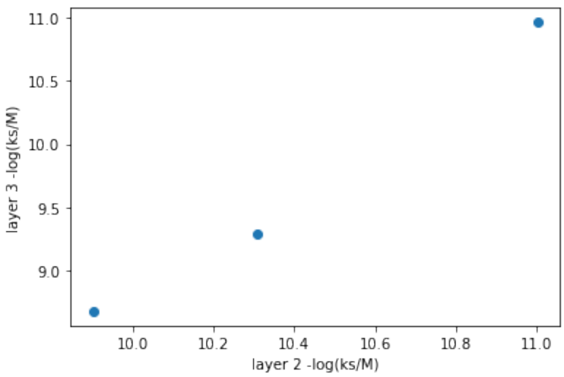
$$p_s = \frac{q_s^\mu}{Z}, \qquad Z = \sum_s q_s^\mu, \qquad \mu = \frac{1}{1 \mp \lambda}.$$

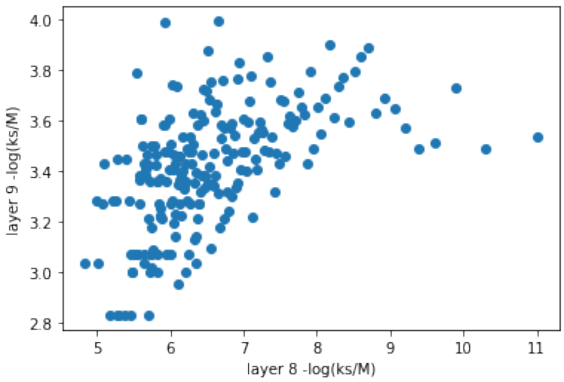




Results 2

The behavior at the shallow and deeper layer are not in the regime predicted by optimal distortion.





Conclusion

- Our hypothesis that the layers of a DBN are related via $p_s = \frac{q_s^\mu}{Z}$ obtains for the intermediate layers.
- There is a trade-off between accuracy and compression, optimally when $\lambda = 0$.
- Is there a maximum level of compression which retains the features of representation necessary for human decision-making?
- Instead of or in addition to the constraint $H[p] \leq H_0$ we specify that the compressed representation of the original decision-making process must be adequate for human decision-making.
- Can the framework be extended to supervised learning?
- Labels might enforce a distorted representation, and this might be an additional cost of compression.

References

- Song, J., Marsili, M. and Jo, J., 2018. "Resolution and relevance trade-offs in deep learning".
- Liu, Y., Khandagale, S., White, C. and Neiswanger, W., 2021. "Synthetic Benchmarks for Scientific Research in Explainable Machine Learning".
- Wojtowicz, Z. and DeDeo, S., 2020. "From Probability to Consilience: How Explanatory Values Implement Bayesian Reasoning. Trends in Cognitive Sciences".
- Marsili, M. "The peculiar statistical mechanics of optimal learning machines".
- See Cover, Thomas M. "Elements of information theory". John Wiley Sons (1999), Chapter 5.