Master Project

Inverse Optimization for Decision-Making in Large Language Models

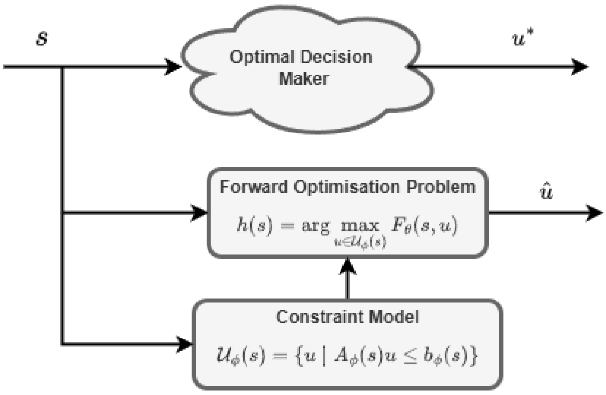
Cedric Pelsma, Delft Center for Systems and Control, TU Delft
C.E.M.Pelsma@student.tudelft.nl

Tolga Ok, Delft Center for Systems and Control, TU Delft
T.Ok@tudelft.nl

Peyman Mohajerin Esfahani, DCSC, TU Delft; Operations Research, University of Toronto
P.MohajerinEsfahani@tudelft.nl P.MohajerinEsfahani@utoronto.ca

Context

In recent years, there has been a growing interest in inverse optimization as a way to reconstruct decision-making processes from observed behavior. It has been used in fields like transportation, control, and power systems [2, 1], where we often have access to the outcomes of optimal decisions but not the underlying rules or preferences that generated them. This can be naturally modeled as a parametric optimization problem to infer the relationship between the input and the optimal decision. Most research on inverse optimization has focused on recovering the objective function, which can even be formulated as a convex training program.


s
Optimal Decision Maker
u^*
Forward Optimisation Problem
$h(s) = \arg\max_{u \in \mathcal{U}_\phi(s)} F_\theta(s, u)$
$\hat{u}$
Constraint Model
$\mathcal{U}_\phi(s) = \{u \mid A_\phi(s)u \leq b_\phi(s)\}$

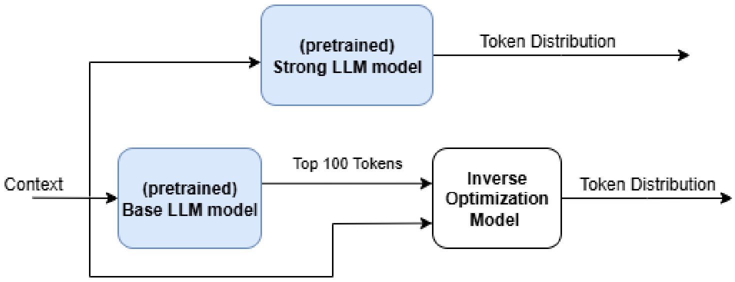
However, in many application domains, such as Large Language Models (LLMs), the constraints in decision-making are unknown and context-dependent, limiting the expressiveness of the inverse optimization framework. By learning the feasible region of our inverse model, we can use the rich expressiveness of the inverse optimization framework in domains like LLMs, which don't have known constraints [3].

LLMs present a new and challenging application domain for inverse optimization. They exhibit complex decision-making behavior, but the rules guiding their outputs are hidden. Inverse optimization offers a way to uncover both what these models optimize for and the implicit constraints shaping their decisions. Leveraging this rich, constraint-learning inverse optimization framework could also yield valuable insights for training an LLM.

Project tasks

This project will begin by exploring how the inverse optimization framework can be extended to learn constraints. This extension leads to a non-convex problem that will require suitable solution methods. The resulting constraint-learning-based inverse optimization approach will then be applied to large language models in collaboration with Cognichip, starting with efforts to recover a strong LLM model using supervised learning. The project goals are as follows:

1. Investigate data driven constrain estimation in the inverse optimization framework
2. Apply a proposed method to learn constraints in an inverse optimization benchmark
3. Integrate the method to use with Large Language Models



References

[1] Syed Adnan Akhtar, Arman Sharifi Kolarijani, and Peyman Mohajerin Esfahani. Learning for control: An inverse optimization approach. *IEEE Control Systems Letters*, 6:187–192, 2022.

[2] Timothy C. Y. Chan, Rafid Mahmood, and Ian Yihang Zhu. Inverse optimization: Theory and applications, 2022.

[3] Ke Ren, Peyman Mohajerin Esfahani, and Angelos Georghiou. Inverse optimization via learning feasible regions, 2025.