DISSERTATION PROPOSAL

Macarena Navarro

"Augmenting Generative AI with Constraint Reasoning for Combinatorial Optimization"

Friday, December 5, 2025
9:00am
Tepper 4219

This dissertation investigates how generative artificial intelligence can be used to solve combinatorial optimization problems. In principle, generative models have the capacity to produce fast instance-specific heuristics by learning rich structural patterns. In practice, this promise is met by the limitation that generative models struggle to reason about complex combinatorial constraints and therefore cannot guarantee feasibility on their own. A second challenge is applicability: when can such data-driven methods outperform traditional combinatorial optimization approaches which are backed by a century of research and development? We identify that these models excel when components of the model such as the objective or constraints are unknown, or when fast heuristic decisions are needed at scale. This dissertation proposes a unified methodology that integrates constraint reasoning modules, in the form of masking mechanisms, and autoregressive generative modeling for combinatorial optimization.

Chapter 1 focuses on combinatorial optimization problems with unknown components, where we have access to a dataset of past solutions of desirable quality. Instead of recovering missing objectives or constraints through inverse optimization, this work proposes learning an instance-to-solution mapping using a transformer-based structured prediction model. Known constraints are represented using a deterministic finite automaton (DFA), which is incorporated into the autoregressive decoding process as a dynamic mask that identifies the feasible decisions at each step, both during training and testing. For the monotone constraints considered in this work, such as permutation structures, this mechanism guarantees feasibility while keeping the DFA polynomial in size. This chapter contributes a novel application of generative models to problems traditionally approached through inverse optimization, formally defines DFA-based masks for enforcing feasibility, and evaluates the proposed method on three classical combinatorial optimization problems with unknown components: the knapsack problem with an unknown objective, the bipartite matching problem with an unknown objective, and the single-machine scheduling problem with release times and unknown precedence constraints. The approach consistently outperforms inverse optimization methods and LSTM-based sequence models, particularly when the underlying functional form of unknown components is complex or misspecified, or when the dataset contains corrupted observations. This work is in collaboration with Karan Singh and Willem-Jan van Hoeve.

Chapter 2 focuses on combinatorial optimization problems with fully specified objectives and constraints, but where feasibility depends on non-monotone constraints that require global reasoning across the entire decision space. A canonical example is the time window structure in the travelling salesperson problem with time windows, where arriving within the time window at one customer does not guarantee that future windows remain reachable, making feasibility inherently non-local. Prior work in neural combinatorial optimization has shown that deep reinforcement learning can train policies that generate near-optimal solutions, but they struggle to enforce feasibility under non-monotone constraints. Unlike the constraints considered in Chapter 1, non-monotone constraints lead to exponentially large DFAs that are computationally expensive to construct. Therefore, this work investigates non-exact DFAs as constraint reasoning modules that provide approximate yet computationally tractable guidance to neural policies. Two examples of such tools are relaxed multi-valued decision diagrams, which offer structured lookahead capabilities to detect downstream infeasibility, or data structures that encode valid constraints. This chapter studies how these mechanisms can be used to reduce the search space without hindering learnability, analyzing their effect on feasibility preservation, computational cost, and solution quality. As application problems, this work applies this methodology to routing problems with time window constraints. This work is in collaboration with Karan Singh, Willem-Jan van Hoeve, and Segev Wasserkrug.

Chapter 3 considers general combinatorial optimization problems with non-monotone constraints and provides feasibility guarantees that Chapter 2 lacks. It introduces an iterative refinement procedure that is exact with respect to finding a feasible solution, whenever one exists, and integrates generative models with non-exact DFAs to predict solutions. The approach follows a two-stage pipeline. First, the method trains a generative model to predict solutions to the optimization problem. Second, during testing decoding, it iteratively constructs candidate solutions using the generative model together with any non-exact DFA as the constraint-reasoning module, and uses an oracle to detect infeasible ones. Whenever an infeasible solution is identified, the method records a no-good and triggers a refinement of the underlying non-exact DFA structure, thereby eliminating that infeasible solution from subsequent generations. This work is in collaboration with Karan Singh and Willem-Jan van Hoeve.

Proposed Committee: Karan Singh (Co-Chair), Willem-Jan van Hoeve (Co-Chair), Andrew Li, Quentin Cappart (École Polytechnique de Montréal)