From Inverse Optimization to Feasibility to ERM

Saurabh Mishra[1] Anant Raj[2] Sharan Vaswani [1]

[1]Simon Fraser University [2]SIERRA Project Team (Inria)

ICML, 2024



Motivation

Contextual Inverse Optimization (CIO) is inferring unknown parameters of an optimization problem from known solutions.

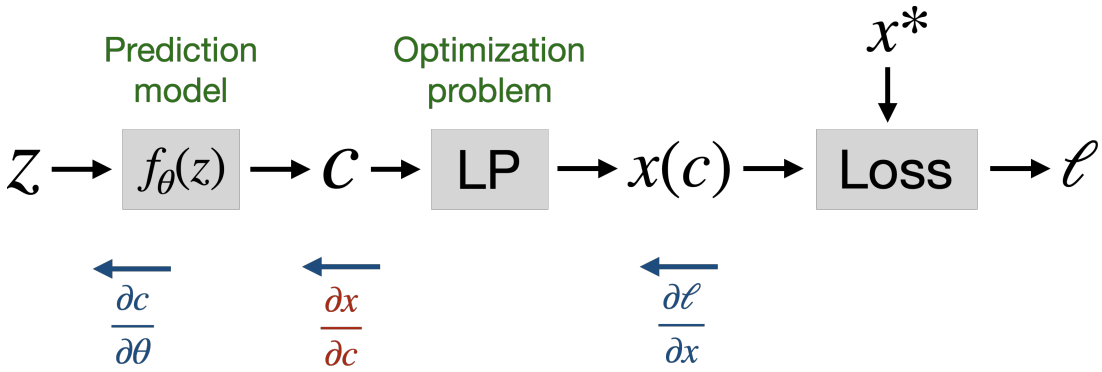
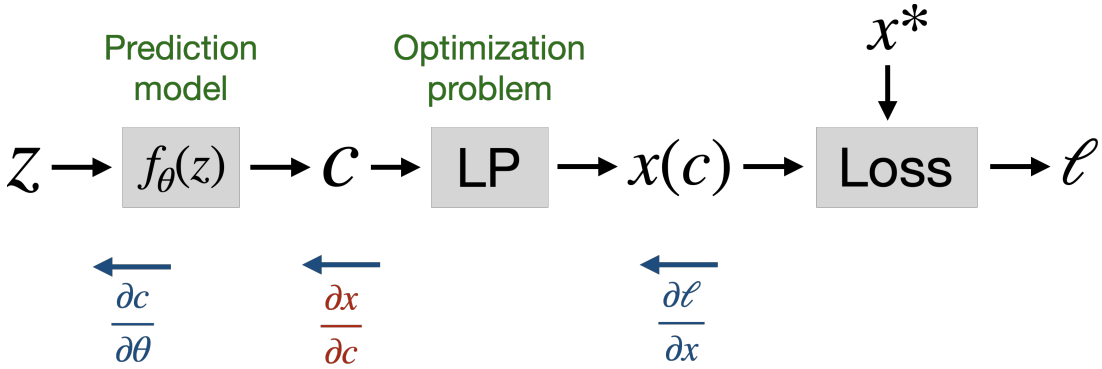


Figure: General Framework

Applications:

- [WSC23] Energy-cost aware scheduling: Use weather data to predict future energy prices to schedule jobs efficiently.
- [WSC23] Inverse reinforcement learning.
- [WDT19] Recommendation Systems.

Problem Overview



- Dataset : $\mathcal{D} = \{z_i, x_i^*\}_{i=1}^N$ where $z_i \in \mathbb{R}^d$ is the input and $x_i^* \in \mathbb{R}^m$ is the corresponding optimal decision.
- Prediction Model : $f_\theta(z) : z \to c$.
- Linear Program (LP): $x(c) := \arg\min_x \langle c, x \rangle$ s.t $Ax = b,\ x \geq 0$.
- Objective : Learn θ s.t. $\forall i \in [N];\ x(f_\theta(z_i)) = x_i^*$.
- Key challenge : Since $x(c) \mapsto c$ is not unique, $\frac{\partial x(c)}{\partial c}$ is either 0 or ∞.

TL;DR: Contributions

End-to-end training

Using the KKT conditions for LPs and make connections from CIO to feasibility and ERM

Theoretical Guarantees

Our method comes with theoretical guarantees without the extra assumption of non-degeneracy or no-noise.

Key Idea: Reduction to Feasibility

- Define C s.t. $\forall c \in C,\ x(c) = x^*$.
- Using KKT [KT51] optimality conditions:

$$C = \{c \,|\, \exists\, \lambda, \nu \ \text{s.t.}\ \nu^T A + \lambda - c = 0, x^* \cdot \lambda = 0, \lambda \geq 0\} \tag{1}$$

- For linear model $\theta \in \mathbb{R}^{d \times m}$,

$$F = \{c \,|\, \exists \theta \ \text{s.t.}\ c = z\theta\}. \tag{2}$$

- Both C and F are convex.
- Objective: Find a $c \in C \cap F$ (known as convex feasibility).

Algorithm and Challenges

Alternating Projections (POCS)

- Requires projection onto sets C and F alternatively.
- Converge to a point in $C \cap F$ if $C \cap F \neq \emptyset$. Else to a point in F (closest to C).
- Projecting on C involves solving a QP.
- Projecting on F involves solving a regression problem.

Challenges:

- $\times$ Exact projection to F is expensive for large dataset.

Reduction to Empirical Risk Minimization (ERM)

Consider the loss function:

$$h(\theta) := \frac{1}{2N} \sum_{i=1}^{N} \min_{q_i \in C_i} \|f_\theta(z_i) - q_i\|^2 \tag{3}$$

Properties of $h(\theta)$

- For a linear model f_θ, h is 1-smooth, convex function.
- Not necessarily strongly convex but satisfies PL [Pol64] condition.

Convergence

- Can use stochastic gradient descent with $O(1)$ iteration cost.
- For a linear model f_θ, when $C \cap F \neq \emptyset$, SGD converges to $C \cap F$ at linear rate.

Sub-optimality and Generalisation

Sub-optimality of decision quality (with $c_\theta = \mathcal{P}_C(f_\theta(z))$

$$\Gamma(\theta, (z, x^*)) = \left\langle \frac{c_\theta}{\|c_\theta\|_2}, x(c) - x^* \right\rangle \tag{4}$$

Relation to $h(\theta)$

$$\Gamma(\theta, (z, x^*)) \leq O\left(\sqrt{m\, h(\theta)}\right) \tag{5}$$

Sub-optimality for unseen-instances

$$\mathbb{E}_{(z,x^*)\sim\rho}\left[\mathbb{E}_{\mathcal{D}\sim\rho}\left[\mathbb{E}[\Gamma(\theta_T, (z, x^*))]\right]\right] = O\left(\sqrt{m}\left[\exp(-T)\right]^{1/2}\right)$$
$$\text{(when } C \cap F \neq \emptyset\text{)}$$

Experiment Details: Dataset

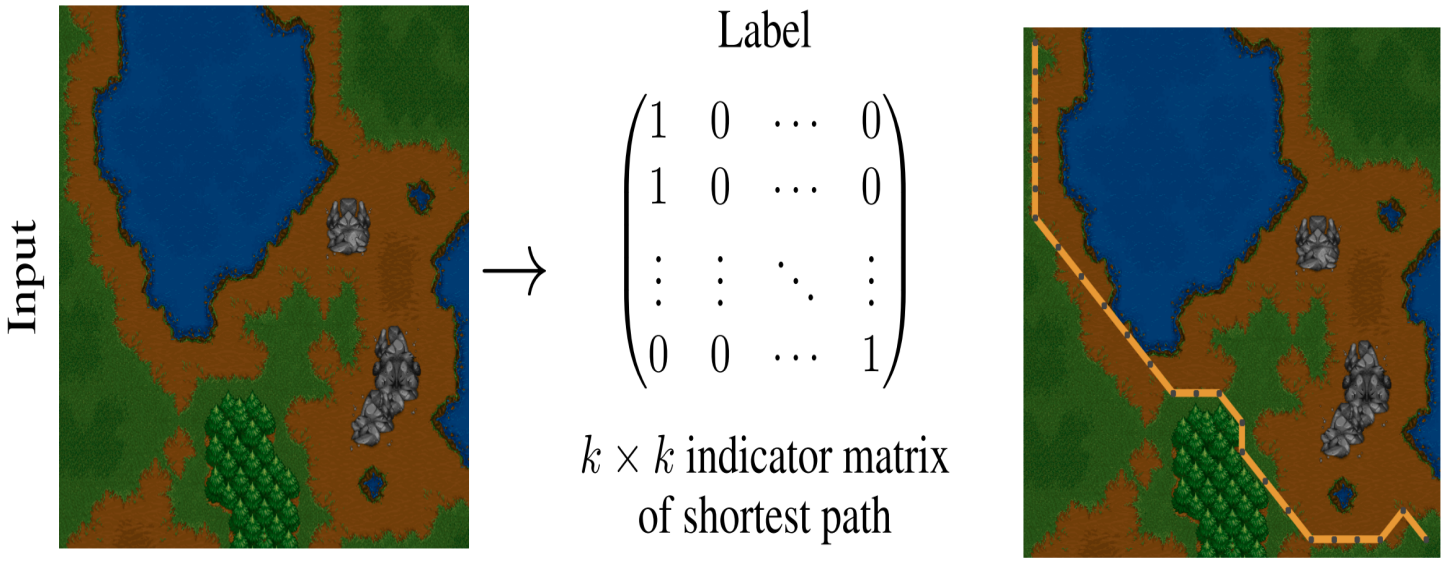
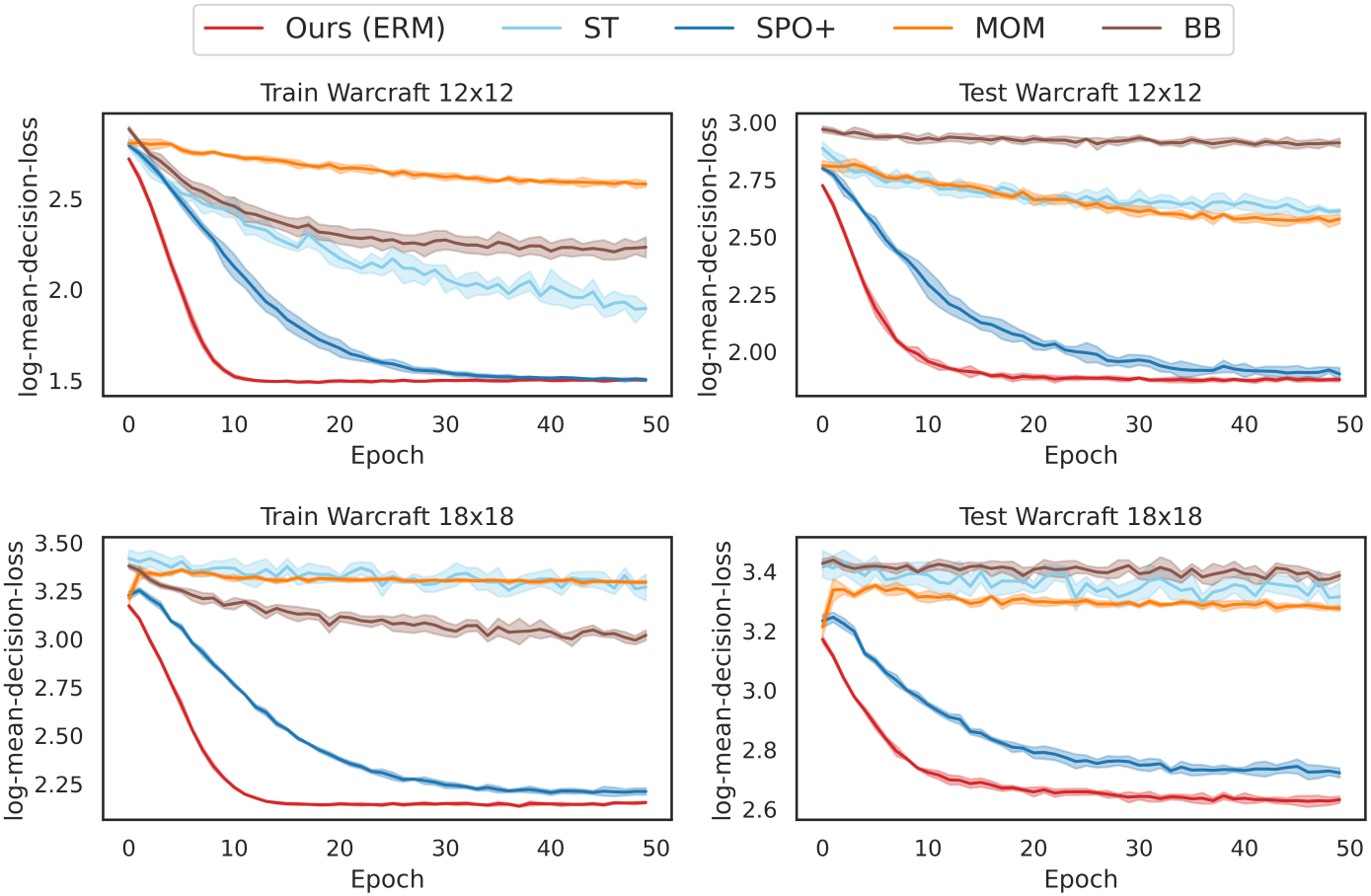


Figure: Warcraft Shortest Path [VPM$^+$19]

Experiment Results



$$\text{Decision-Loss}(\theta) = \sum_{i=1}^{N} \|\hat{x}(f_\theta(z_i)) - x_i^*\|^2 \tag{6}$$

Conclusion

- Reduction the problem of CIO to Feasibility
- Further, presented the reduction to ERM
- Presented the convergence guarantees and generalization guarantees
- Paper: https://arxiv.org/abs/2402.17890
- Contact: skm24@sfu.ca, vaswani.sharan@gmail.com

References I

- Harold W Kuhn and Albert W Tucker, *Nonlinear programming, paper presented at proceedings of the second berkeley symposium on mathematical statistics and probability*, 1951.
- Boris T Polyak, *Gradient methods for solving equations and inequalities*, USSR Computational Mathematics and Mathematical Physics **4** (1964), no. 6, 17–32.
- Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek, *Differentiation of blackbox combinatorial solvers*, arXiv preprint arXiv:1912.02175 (2019).
- Bryan Wilder, Bistra Dilkina, and Milind Tambe, *Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 1658–1665.

References II


Dariush Wahdany, Carlo Schmitt, and Jochen L Cremer, *More than accuracy: end-to-end wind power forecasting that optimises the energy system*, Electric Power Systems Research **221** (2023), 109384.