Handling Non-Stationary Experts in Batch Inverse Reinforcement Learning: Como Lake Water System Control Case Study

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Amarildo Likmeta,
Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, Marcello Restelli
amarildo.likmeta2@unibo.it

POLITECNICO MILANO 1863

Problem and Motivation

- **Reinforcement Lerning** (RL, Sutton and Barto, 2018): find optimal policy π^* maximizing the *expected return* $J^\pi(\theta,\omega)$:

$$J_\pi(\theta,\omega) = \mathop{\mathbb{E}}_{\substack{S_0\sim\mu \\ A_t\sim\pi_\theta(\cdot|S_t) \\ S_{t+1}\sim\mathcal{P}(\cdot|S_t,A_t)}} \left[\sum_{t=0}^{+\infty} \gamma^i R_\omega(S_t,A_t) | S_0 = s\right]$$

- **Inverse Reinforcement Learning**
  1. Given a dataset D of demonstrations from an expert, find the unknown reward function being optimized:
  $$R^*_{\pi^E} \in \left\{ R \in \mathcal{R} : \pi^E \in \arg\max_{\pi\in\Pi} J(\pi,R) \right\}.$$
  2. We further assume linearity of the reward function in terms of a feature function ϕ
  $$\mathcal{R} = \left\{ R_{\boldsymbol{\omega}} = \boldsymbol{\omega}^T\boldsymbol{\phi} : \boldsymbol{\omega} \in \mathbb{R}^q_{\ge 0}, \|\boldsymbol{\omega}\|_1 = 1 \right\}$$

Model-Free No model of the environment is available

Batch We cannot further interact with the environment

Gradient Inverse Reinforcement Learning

- if $\pi_{\boldsymbol{\theta}^E}$ is optimal for the reward $R_{\boldsymbol{\omega}^E}$, $\boldsymbol{\theta}^E$ is a **stationary point** of the return $J(\boldsymbol{\theta},\boldsymbol{\omega}^E) = (\boldsymbol{\omega}^E)^T\boldsymbol{\psi}(\boldsymbol{\theta})$

  $\implies \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^E,\boldsymbol{\omega}^E) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta}^E)\boldsymbol{\omega}^E = 0$
- Find the reward weights $\boldsymbol{\omega}$ that lie in the **null space** of the jacobian

  $\implies$ Due to **estimation error** the null space of the jacobian might be **empty**

Contributions

- **Jacobian correction**: Account for the **uncertainty**, by modelling the sample distribution $\widehat{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta}) \sim \mathcal{N}(\mathrm{M}, \frac{1}{n}\boldsymbol{\Sigma})$

$$\min_{\substack{\boldsymbol{\omega}\in\mathbb{R}^q_{\ge 0} \\ \|\boldsymbol{\omega}\|_1=1}} \left\| \widehat{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\omega} \right\|^2_{\left[(\boldsymbol{\omega}\otimes\mathbf{I}_d)^T\boldsymbol{\Sigma}(\boldsymbol{\omega}\otimes\mathbf{I}_d)\right]^{-1}}, (\Sigma\text{-GIRL})$$

- **Non-Stationarity**: Account for **non-stationary** behavior in the demonstations due to **changing intentions**
  - **Detect** K intention change points and **identify** the reward functions $\{\boldsymbol{\omega}\}_{j=1}^K$

$$\min_{\substack{\boldsymbol{\omega}_{uv}\in\mathbb{R}^q_{\ge 0} \\ \|\boldsymbol{\omega}_{uv}\|_1=1}} (v-u)\sum_{i=u}^{v-1} \left\| \widehat{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\psi}_i(\boldsymbol{\theta})\boldsymbol{\omega}_{uv} \right\|^2_{\left[(\boldsymbol{\Sigma}(u,v)_i\right]^{-1}}$$

- **Real-World Application**: Application of IRL in real-world dataset of Lake Como dam operation

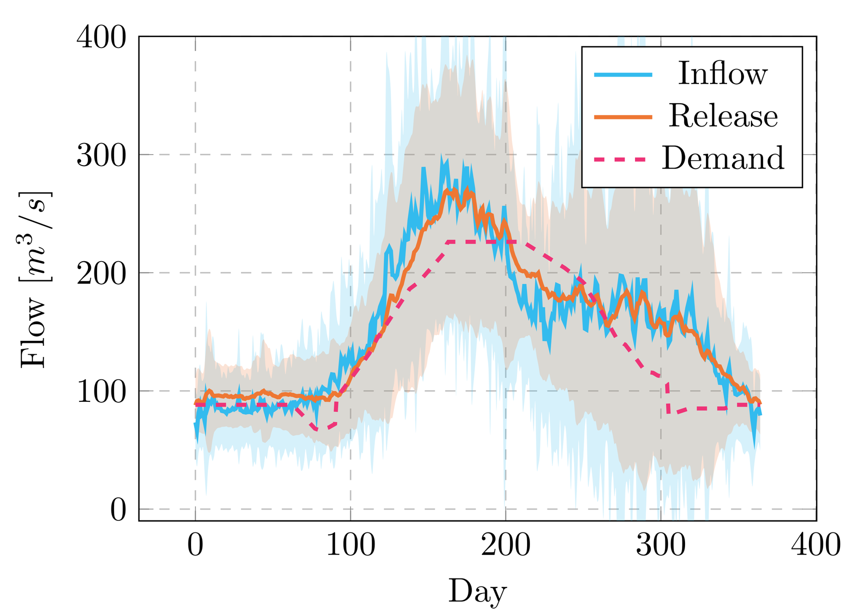
Como Use Case

System Modelling

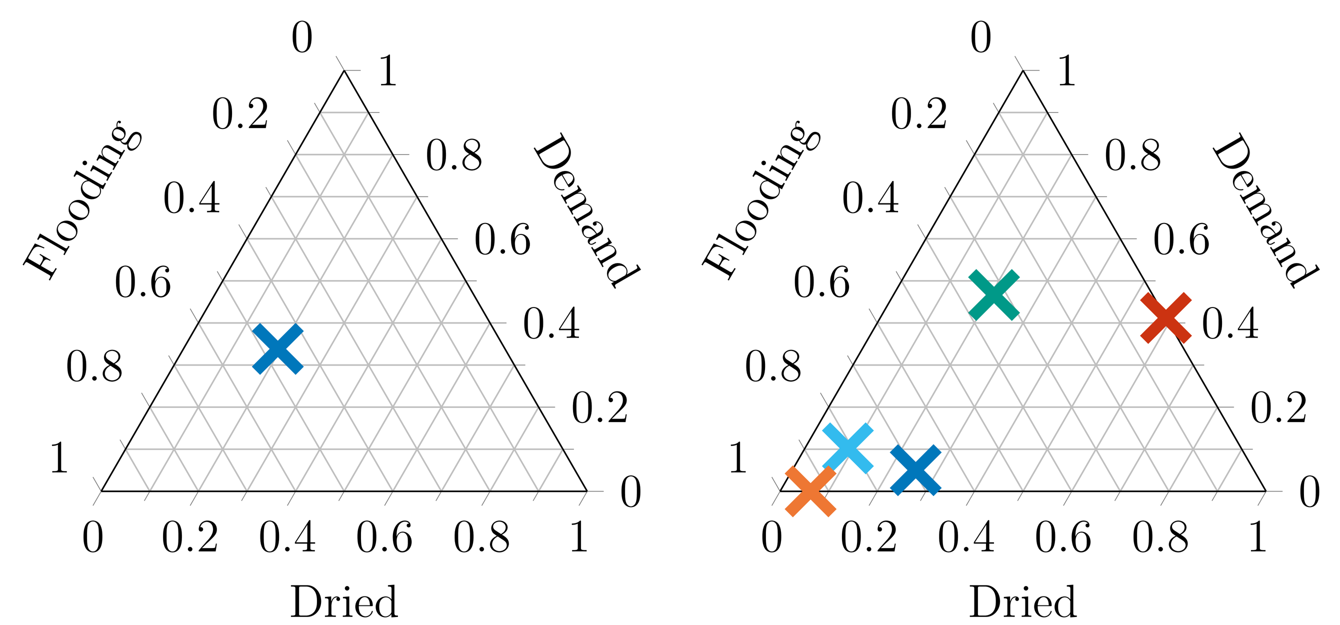
- **Problem**: Infer operator intentions from hystorical dam operation.
- **Problem**: The intentions of the operators change during 60 years
- Model as discrete-time, periodic, nonlinear, stochastic MDP
- Continuous state: water stored in the lake S_t, a continuous action: water released a_t, a state-transition function of: lake inflow q_{t+1}

$$S_{t+1} = S_t + q_{t+1} - r_{t+1}(S_t, a_t, q_{t+1})$$

- Three reward features representing conflicting objectives:
  - *Supply deficit* - (ϕ^D): deficit between the release and the demand
  - *Flood risk* (ϕ^F): penalize small releases associated to high lake levels
  - *Drought risk* (ϕ^L): penalize large releases with low lake levels



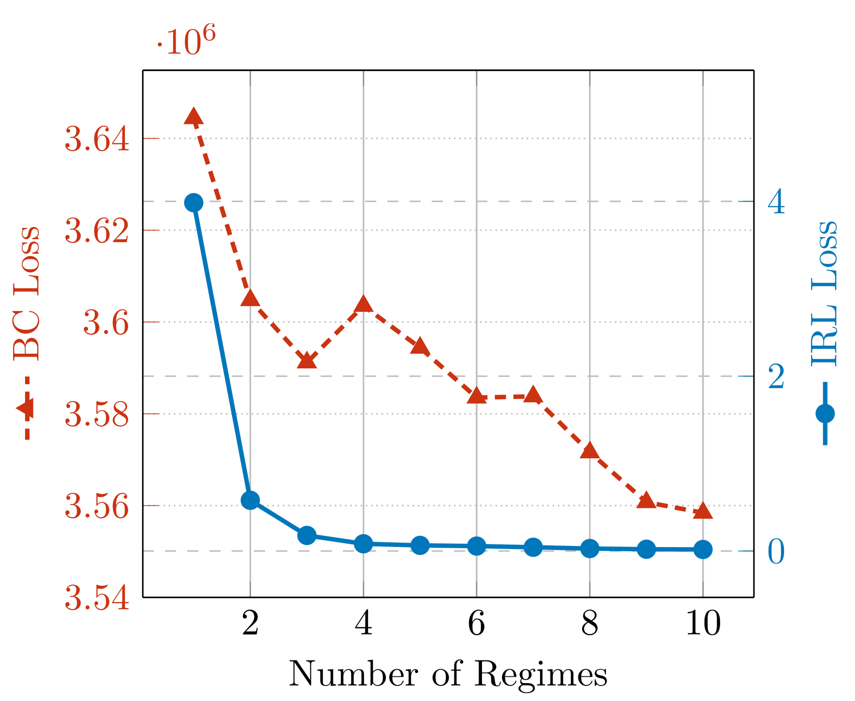
Single Reward IRL Results

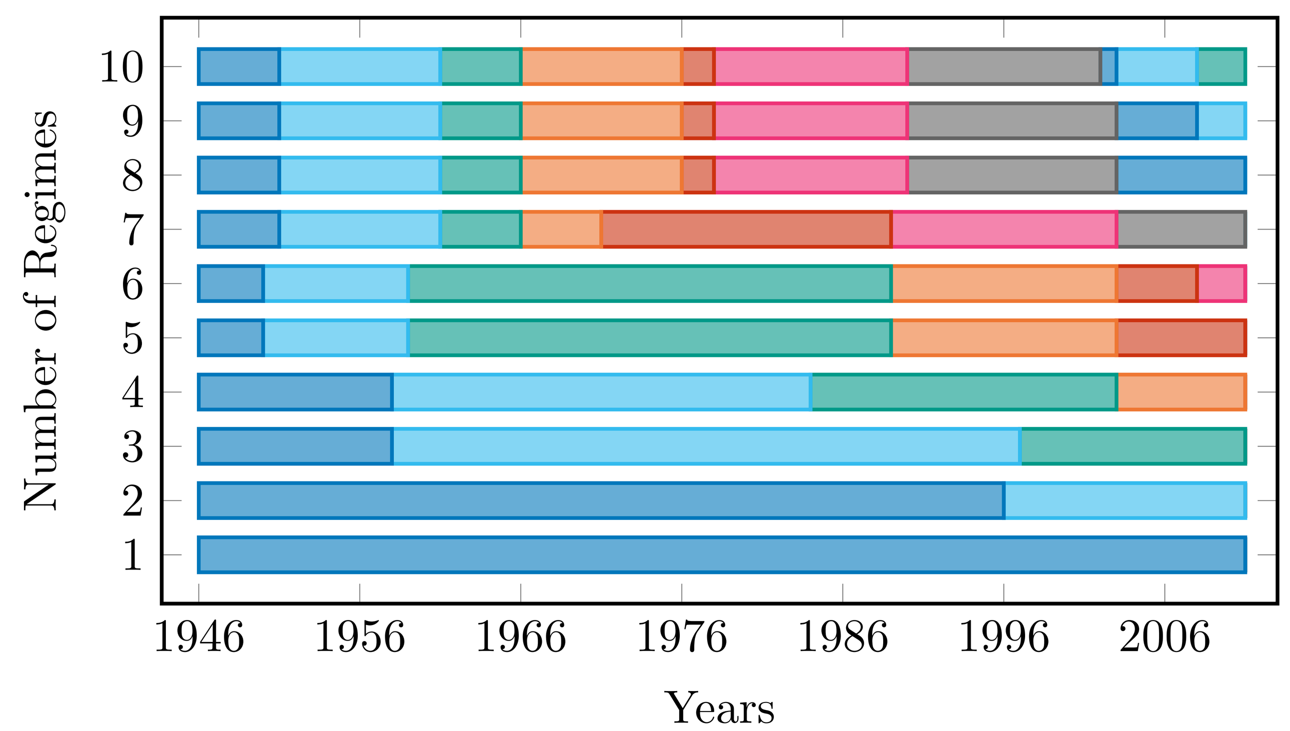


- Slight predominance of the interest in controlling the floods ($\omega^F = 0.47$)
- Remaining weight is divided between demand ($\omega^D = 0.34$) and drought control ($\omega^L = 0.19$)
- Results in line with literature results (Giuliani et al., 2019)
- Expert almost Pareto optimal

Non-Stationary IRL Results

- **Problem**: Lake Como is a **non-stationary**: system that has undergone several alterations
- **Idea**: Consider the dataset as a lifelong trajectory. Make subdivisions yearly. Find K **intention change points**.





- From an elbow analysis we can identify 4 or 5 distinct periods.
- Intervals match historical events such as, big flooding events (1987), or extreme droughts (2003)

Σ−GIRL (Ramponi et al., 2020)

input: dataset of demonstrations D
output: optimal parameters $\boldsymbol{\omega}^*$

1: Perform BC to find the policy parameters $\boldsymbol{\theta}^E$
2: Estimate $\widehat{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta})$ and Σ from the demonstration dataset
3: Find the weights $\boldsymbol{\omega}^*$ minimizing
$$\boldsymbol{\omega}^* = \arg\min \left\| \widehat{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\omega} \right\|^2_{\left[(\boldsymbol{\omega}\otimes\mathbf{I}_d)^T\boldsymbol{\Sigma}(\boldsymbol{\omega}\otimes\mathbf{I}_d)\right]^{-1}}$$
4: **return** $\boldsymbol{\omega}^*$

NS-Σ-GIRL (Likmeta et al., 2020)

input: dataset $\boldsymbol{\tau} = (\tau_1|\tau_2|\ldots|\tau_T)$,number of regimes k
output: optimal parameters $\Omega = (\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_k, t_1,\ldots,t_{k-1})$

```
1:  for u = 1, ..., T − 1 do
2:      for v = u + 1, ..., T do
3:          Define D_uv = {τ_u, ..., τ_{v−1}}
4:          Perform BC to find the policy parameters θ_uv
5:          Optimize ω*_uv ∈ arg max_ω log p(D_uv | ω)
6:          C_1(u, v) = log p(D_uv | ω*_uv)
7:      end for
8:  end for
9:  for l = 2, ..., k − 1 do
10:     for u = 1, ..., T − l do
11:         for v = u + l, ..., T do
12:             C_l(u, v) = max_{u+l−1≤t<v} {C_{l−1}(u, t) + C_1(t + 1, v)}
13:         end for
14:     end for
15: end for
16: t_k = T
17: for l = k, ..., 2 do
18:     t_{l−1} ∈ arg max_{l−1≤t<t_l} C_{l−1}(1, t) + C_1(t + 1, t_l)
19:     ω_l = ω*_{t_{l−1} t_l}
20: end for
21: return Ω
```

References

M. Giuliani, M. Zaniolo, A. Castelletti, G. Davoli, and P. Block. Detecting the state of the climate system via artificial intelligence to improve seasonal forecasts and inform reservoir operations. *Water Resources Research*, 55:9133–9147, 2019.

A. Likmeta, G. Ramponi, A. M. Metelli, M. Tirinzoni, Andrea Giuliani =, and M. Restelli. Dealing with multiple experts and non-stationarity in inverse reinforcement learning:an application to real-life problems. In *Machine Learning Journal (Under Revision)*, 2020.

G. Ramponi, A. Likmeta, A. M. Metelli, A. Tirinzoni, and M. Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *The 23nd International Conference on Artificial Intelligence and Statistics*, 2020.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.