Efficient Sampling-Based Maximum Entropy Inverse Reinforcement Learning with Application to Autonomous Driving

04/11/21

Authors: Zheng Wu, Liting Sun, Wei Zhan, Chenyu Yang, and Masayoshi Tomizuka
Presented by: Xinyi Yan

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Introduction

- What is the paper about ?
  - Autonomous driving
  - Learn driving policies
  - What to optimize ?
  - Extract what human drivers try to optimize from real traffic data

Introduction

- What is the problem tackled ?
  - Extract what human drivers try to optimize from real traffic data
  - Challenges:
    - high dimensional continuous space with long horizons
    - Vehicle kinematics: distance, speed, acceleration. etc
    - Uncertainties
    - Interpretable, generalizable

Introduction

- What is the solution proposed ?
  - Sampling-based maximum entropy IRL (SMIRL)

Background

- The problem tackled
  - Extract what human drivers try to optimize from real traffic data
  - Challenges:
    - high dimensional continuous space with long horizons
    - Vehicle kinematics: distance, speed, acceleration. etc
    - Uncertainties
    - Interpretable, generalizable
  - Learn reward functions from real driving data
  - The principle of maximum entropy
  - Trajectory sampling

Background

- Necessary background
  - Principle of maximum entropy
  - Maximum Entropy Inverse Reinforcement Learning
  - Assumptions:
    - the reward function is roughly consistent

Content

- SMIRL at a high level
  - A set of demonstrations: $\Xi_D = \{\xi_i\}$
  - $R(\xi, \theta) = \theta^T \mathbf{f}(\xi)$
  - Boltzmann rationality: $P(\xi, \theta) \propto e^{\beta R(\xi, \theta)}$
  - $P(\Xi_D|\theta) = \prod_{i=1}^{M} \frac{e^{\beta R(\xi_i, \theta)}}{\int_{\tilde{\xi} \in \Phi_{\xi_i}} e^{\beta R(\tilde{\xi}, \theta)} d\tilde{\xi}} = \prod_{i=1}^{M} \frac{1}{Z_{\xi_i}} e^{\beta R(\xi_i, \theta)}$
  - $\theta^\star = \arg\max_\theta \frac{1}{M} \log P(\Xi_D|\theta) = \arg\max_\theta \frac{1}{M} \sum_{i=1}^{M} \log P(\xi_i|\theta)$
  - $Z_{\xi_i} \approx \sum_{m=1}^{K} e^{\beta R(\tau_m^i, \theta)}$

Content

- SMIRL at a high level
  - The Sampler
    - Discrete Elastic Band
    - Path Smoothing
    - Two-step Speed Sampling

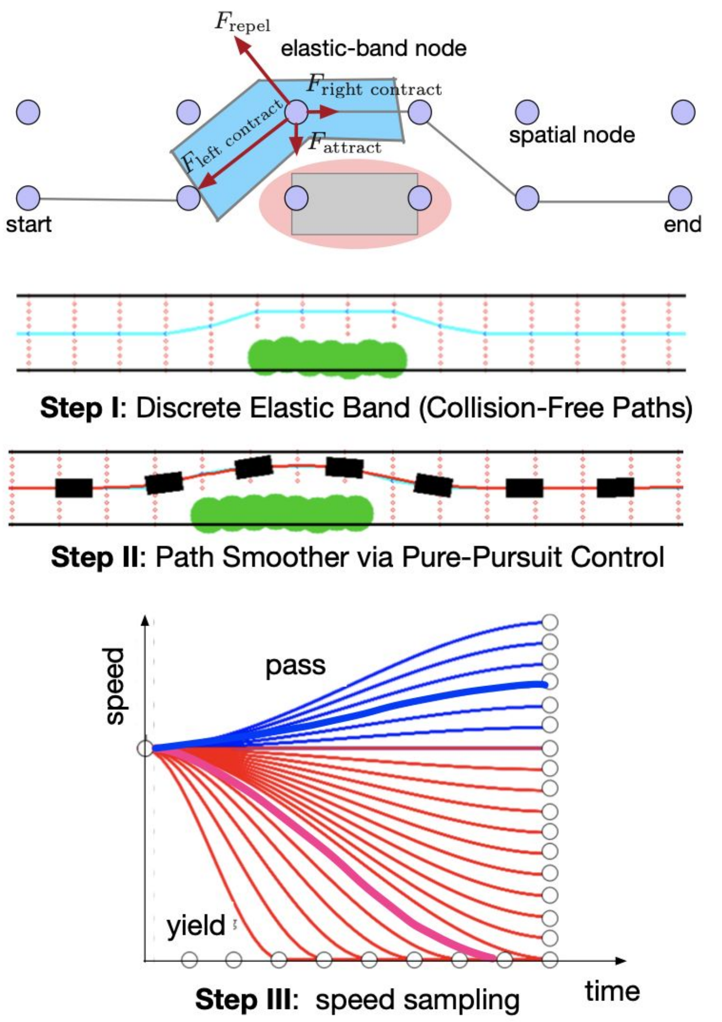


Fig. 1: The overview of the sampling process.

Content

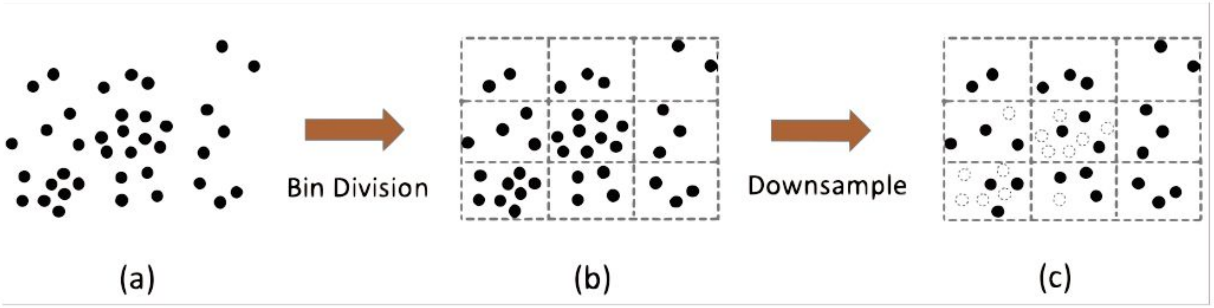
- SMIRL at a high level
  - Re-Distribution of Samples
    - $$P(\Xi_D|\theta) = \prod_{i=1}^{M} \frac{e^{\beta R(\xi_i,\theta)}}{\int_{\tilde{\xi}\in\Phi_{\xi_i}} e^{\beta R(\tilde{\xi},\theta)} d\tilde{\xi}} = \prod_{i=1}^{M} \frac{1}{Z_{\xi_i}} e^{\beta R(\xi_i,\theta)}$$
    - $$Z_{\xi_i} \approx \sum_{m=1}^{K} e^{\beta R(\tau_m^i,\theta)}$$
    - 

Fig. 2: Re-distribution of samples

Content

- SMIRL at a high level
  - **Algorithm 1:** The Proposed Sampling-based Maximum Entropy IRL for Driving

Result: optimized reward function parameters $\theta^\star$

Input: The demonstration dataset $\mathcal{D}_M = \{\xi_i\}_{i=1:M}$, the convergence threshold ϵ and the learning rate α.

Initialize θ_0, $k = 0$ and compute expected expert feature count $\bar{\mathbf{f}}(\mathcal{D}_M) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{f}(\xi_i)$;

Generate the sample set $\mathcal{D}_s^0 = \{\tau_m^i\}_{m=1:K,i=1:M}$ using the sampler in Section II-B;

Re-distribute the samples according to their similarities as discussed in Section II-C, and generate a new sample set $\mathcal{D}_s$;

Compute the initial expected feature count over all samples $\widetilde{\mathbf{f}}_0(\mathcal{D}_s) = \dfrac{1}{M}\sum_{i=1}^{M}\widetilde{\mathbf{f}}_0(\xi_i) = \dfrac{1}{M}\sum_{i=1}^{M}\frac{1}{K}\sum_{m=1}^{K}\frac{\exp R(\tau_m^i,\theta_0)}{\sum_{m=1}^{M}\exp R(\tau_m^i,\theta_0)}\mathbf{f}(\tau_m^i)$;

while $\|\bar{\mathbf{f}}(\mathcal{D}_M) - \widetilde{\mathbf{f}}_k(\mathcal{D}_s)\|_2 \geq \epsilon$ **do**

  Update θ_k using gradient decent, i.e., $\theta_{k+1} = \theta_k + \nabla_{\theta_k}L = \theta_k + \alpha(\bar{\mathbf{f}}(\mathcal{D}_M) - \widetilde{\mathbf{f}}(\mathcal{D}_s))$;

  Compute the expected feature count based on θ_{k+1} over all samples $\widetilde{\mathbf{f}}_{k+1}(\mathcal{D}_s) = \dfrac{1}{M}\sum_{i=1}^{M}\widetilde{\mathbf{f}}_{k+1}(\xi_i) = \dfrac{1}{M}\sum_{i=1}^{M}\frac{1}{K}\sum_{m=1}^{K}\frac{\exp R(\tau_m^i,\theta_{k+1})}{\sum_{m=1}^{M}\exp R(\tau_m^i,\theta_{k+1})}\mathbf{f}(\tau_m^i)$;

  $k = k + 1$;

end

$\theta^* = \theta_k$;

Content

- Advantages and disadvantages of the proposed solution compared to other work
  - Scale well in large-scale continuous domain with long horizons
  - Interpretable and generalizable features
    - Non-interactive features: Speed, longitudinal and lateral accelerations, etc
    - Interactive features: Future distance, etc
  - Less sensitive to either noise and feature selection
  - Need manually crafted features
    - Speed, Longitudinal and lateral accelerations, etc.

Empirical evaluation

- Two types of driving scenarios
    - non-interactive driving when moving through the roundabout
    - interactive driving when moving through the roundabout

(a) The NID scenario with DR_USA_Roundabout_SR (b) The ID scenario with DR_USA_Roundabout_FT

Fig. 3: Two roundabout scenarios.

Empirical evaluation

- Evaluation Metrics
  - Feature Deviation: $\mathcal{E}_{FD} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\frac{|\mathbf{f}(\xi_i^{gt}) - \mathbf{f}(\xi_i^{pred})|}{\mathbf{f}(\xi_i^{gt})}$
  - Mean Euclidean distance: $\mathcal{E}_{MED} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}||\xi_i^{gt} - \xi_i^{pred}||_2$
  - Probabilistic Metrics: $P(\xi|\theta, \{\mathcal{T}\}) = \frac{\exp\left(R(\xi,\theta)\right)}{\exp\left(R(\xi,\theta)\right) + \sum_{i=1}^{M}\exp\left(R(\tau_i,\theta)\right)}$

Empirical evaluation

▪

TABLE I: A summary of the IRL algorithms in the non-interactive driving scenario.

	a_lon	j_lon	v_des	a_lat	MED	Win Count	Log Likelihood
Ours	0.16±0.12	0.20±0.15	0.09±0.04	0.09± 0.03	0.21± 0.06	33	-238.98
Opt-IRL	0.19± 0.19	0.32± 0.19	0.13± 0.06	0.11± 0.03	0.29± 0.09	0	-398.93
CIOC	0.48± 0.42	0.23± 0.17	0.10± 0.07	0.06± 0.05	0.23± 0.09	0	-662.16
GCL	—	—	—	—	3.73± 1.95	0	-1377.65

TABLE II: A summary of the IRL algorithms in the interactive driving scenario.

	a_lon	j_lon	a_lat	v_des	fut_dis	fut_int_dis	MED	Win Count	Log Likelihood
Ours	0.15± 0.24	0.54± 0.19	0.19± 0.24	0.034± 0.026	0.012± 0.0078	0.032± 0.045	0.066± 0.038	63	-515.97
Opt-IRL	0.69± 1.04	0.55± 0.40	0.20± 0.23	0.083± 0.11	0.021± 0.018	0.043± 0.066	0.14± 0.16	4	-802.01
CIOC	0.42± 0.77	0.69± 0.26	0.26± 0.23	0.064 ± 0.10	0.023± 0.012	0.045± 0.10	0.14± 0.14	9	-595.27
GCL	—	—	—	—	—	—	1.53± 1.16	0	-1196.75

Empirical evaluation

- Performance on Test Sets in Unseen Environments

TABLE IV: Generalization results of different IRL algorithms under the MED metric. The results are in meters.

	Seen NID	Unseen NID	Seen ID	Unseen ID
Ours	0.21	0.74	0.066	0.072
Opt-IRL	0.29	0.89	0.14	0.17
CIOC	0.23	0.90	0.14	0.16
GCL	3.73	46.70	1.53	4.69

TABLE V: Generalization results of different IRL algorithms under the probabilistic metric.

	Seen NID	Unseen NID	Seen ID	Unseen ID
Ours	-238.98	-399.85	-515.97	-571.60
Opt-IRL	-398.93	-472.51	-802.01	-870.72
CIOC	-662.16	-1153.74	-595.27	-621.13
GCL	-1377.65	-3140.24	-1196.75	-2898.64

Empirical evaluation

- Computation Complexity

TABLE VI: The time cost of the three algorithms for both non-interactive and interactive scenarios. Results are in minutes

	Ours	CIOC	Opt-IRL	GCL
Non-interactive	6	60	1800	40
Interactive	5	90	1260	30

Empirical evaluation

- The Effect of Sample Re-Distribution

TABLE VII: Experiment results of the non-interactive scenario with and without the step of sample re-distribution

	a_lon	j_lon	v_des	a_lat	MED	Win Count	Log Likelihood
w/ sample re-distribution	0.16± 0.12	0.20± 0.15	0.09± 0.04	0.09± 0.03	0.21± 0.06	33	-238.982
w/o sample re-distribution	0.18± 0.10	0.30± 0.16	0.12± 0.05	0.11± 0.04	0.26± 0.08	0	-259.064

TABLE VIII: Experiment results of the interactive scenario with and without the step of sample re-distribution

	a_lon	j_lon	a_lat	v_des	fut_dis	fut_int_dis	MED	Win Count	Log Likelihood
w/ sample re-distribution	0.14 ± 0.24	0.53 ± 0.18	0.19 ± 0.23	0.032 ± 0.026	0.012 ± 0.0074	0.027 ± 0.044	0.072 ± 0.043	76	-515.965
w/o sample re-distribution	0.23 ± 0.53	0.55 ± 0.18	0.19 ± 0.23	0.031 ± 0.028	0.012 ± 0.0062	0.027 ± 0.045	0.067 ± 0.041	0	-557.307

Contribution

- Proposed a sampling-based maximum entropy inverse reinforcement learning algorithm
- Efficient sampler and sample re-distribution
- Better generalization ability and converge significantly faster

Take home message

- Extract human behaviors from real traffic data
- The principle of maximum entropy
- A uniformed distribution of samples

Future Work

- General robotic systems with higher dimensions
- Explore better metrics other than the Euclidean distance

UNIVERSITY OF WATERLOO

FACULTY OF MATHEMATICS

Thank you