The 22nd Behavior Modeling Summer School

(10:45-11:15 Sep. 19, 2023)

Markov decision process for modeling behavior in transportation networks

Yuki Oyama

Shibaura Institute of Technology

Activity Landscape Design Lab.

oyama@shibaura-it.ac.jp

Contents

1. Markov decision process
2. MDP in networks
3. Reward inference as inverse problem
4. Adversarial Inverse reinforcement learning
5. Summary

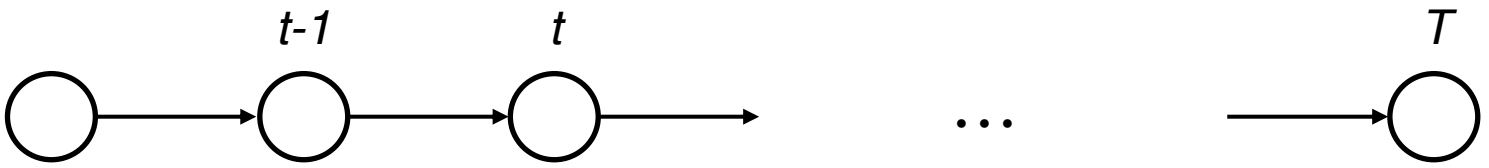
Markov decision process (MDP)

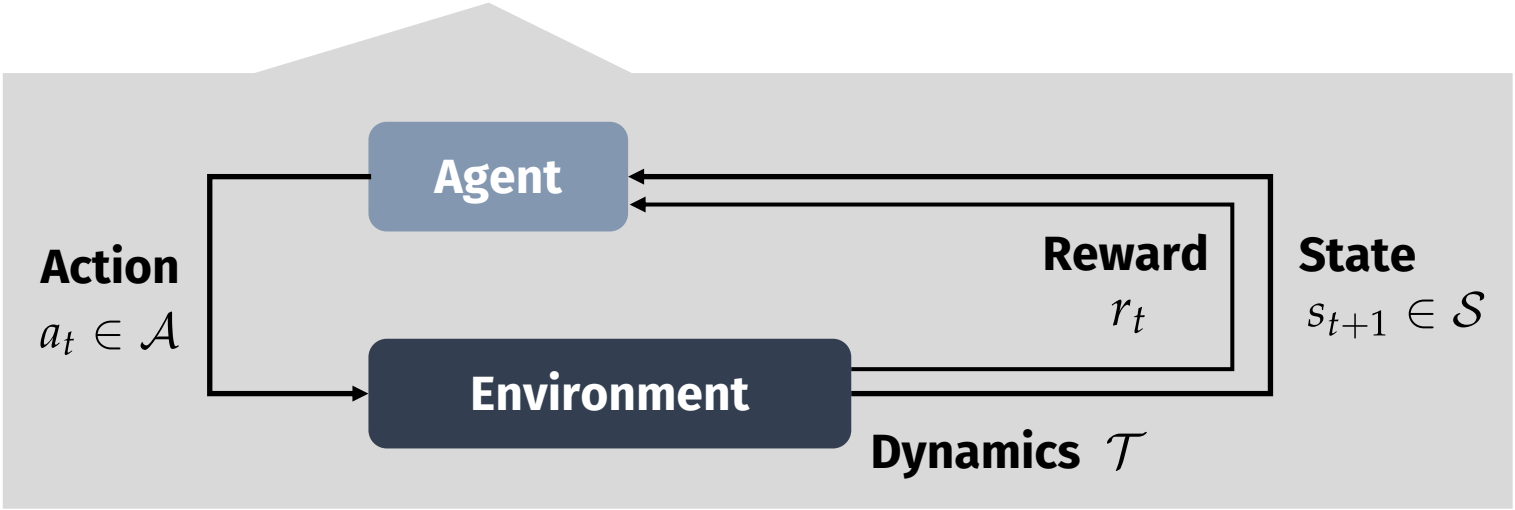
Modeling **sequential decision-making** of an agent under uncertainty

MDP is defined as

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$$

- $\mathcal{S}$: State space
- $\mathcal{A}$: Action space
- $\mathcal{T}$: Dynamics (state transition)
- r : Reward function





Markov decision process (MDP)

Agent's behavior

Decide **policy** (**action probability**) to maximize expected sum of rewards

$$\pi^\star = \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \middle| a_t = \pi(s_t) \right]$$

Recursive formulation:

$$V(s_t) = \max_{a_t \in \mathcal{A}(s_t)} \left\{ r(s_t, a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{T}(s_{t+1} | a_t, s_t) V(s_{t+1}) \right\}$$

Different **reward specifications** lead to different MDP models (Mai & Jaillet, 2020)

- Regularized MDP

$$\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \left(v(s_t, a_t) + \phi(s_t) \right) \right]$$

- **Entropy regularized MDP (Maximum causal entropy)**

$$\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \left(v(s_t, a_t) - \ln \pi(a_t | s_t) \right) \right]$$

- **Dynamic discrete choice model**

$$\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \left(v(s_t, a_t) + \epsilon(a_t) \right) \right]$$

Mai, T., & Jaillet, P. (2020). A relation analysis of Markov decision process frameworks. *arXiv preprint arXiv:2008.07820*.

Dynamic discrete choice model (Rust, 1987)

Connection of MDP with a **choice model**

- State ***s*** : **situation** for the choice (e.g., attributes of alternatives/agents)
- Action ***a*** : **choice** among alternatives
- Reward ***r*** : **utility** that agent gains from choice
- Policy ***π*** **: choice probability**

Under (additive) **random utility theory**:

$$\max_{p} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(v(a_t|s_t) + \epsilon(a_t)\right)\right] \quad \text{with} \quad r(a|s) = v(a|s) + \epsilon(a)$$

Systematic & **error** components

Value function (with known dynamics) is then

$$V(s_t) = \mathbb{E}\left[\max_{a_t \in \mathcal{A}(s_t)} \left\{ v(a_t|s_t) + \epsilon(a_t) + \gamma \sum_{s_{t+1} \in \mathcal{S}} \mathcal{T}(s_{t+1}|a_t, s_t) V(s_{t+1}) \right\}\right]$$

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica:* 999-1033.

MDP in (static) networks

Network route choice MDP models

- Often define **link** (node) **as state** so that action always and directly leads to the same next state (**deterministic dynamics**)
- Have **destinations** where agents terminate their actions (**episodic MDP**)

Dynamic discrete choice model (same as previous):

$$V(s) = \mathbb{E}\left[\max_{a \in \mathcal{A}(s)} \left\{ v(a|s) + \epsilon(a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s'|a,s) V(s') \right\}\right]$$

$$\Downarrow$$

Recursive logit (route choice) model (Fosgerau et al., 2013):

$$V(s) = \mathbb{E}\left[\max_{a \in \mathcal{A}(s)} \{ v(a|s) + \epsilon(a) + V(a) \}\right] = \ln \sum_{a \in \mathcal{A}(s)} e^{v(a|s)+V(a)} \quad \text{(i.i.d. Gumbel)}$$

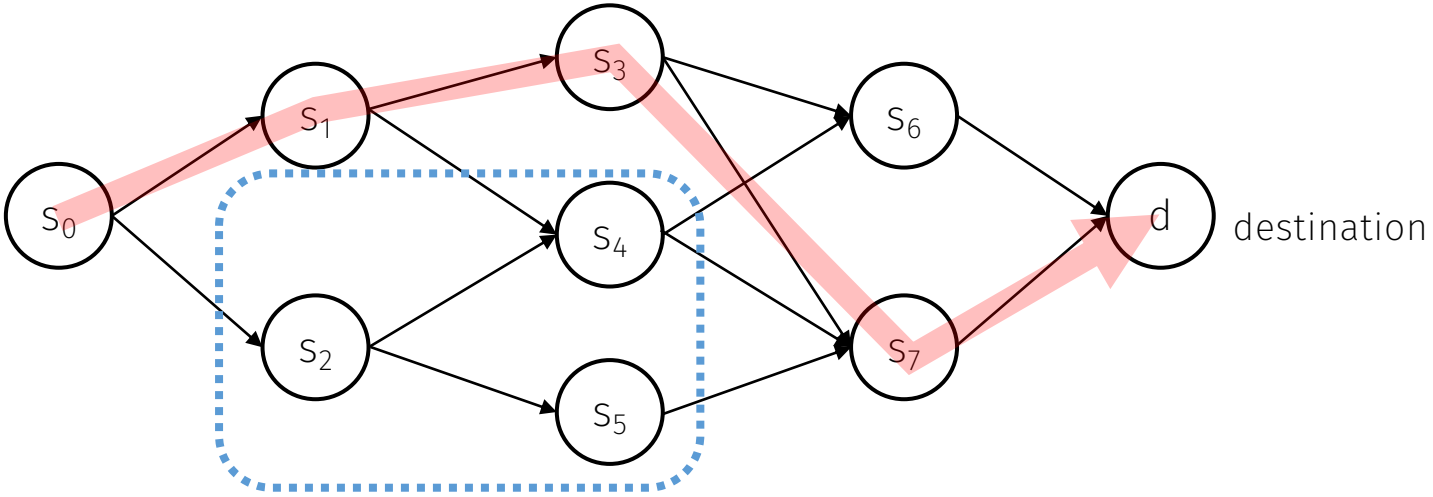
$$p(a|s) = \frac{e^{v(a|s)+V(a)}}{\sum_{a' \in \mathcal{A}(s)} e^{v(a'|s)+V(a')}} \quad \text{(action choice probability)}$$

*Equivalent to **Maximum Causal Entropy MDP** (Ziebart et al., 2008; Zierbart, 2010)

Fosgerau, M., Frejinger, E., & Karlstrom, A. (2013). A link based network route choice model with unrestricted choice set. *Transp. Res. Part B 56*: 70-80.

MDP in networks

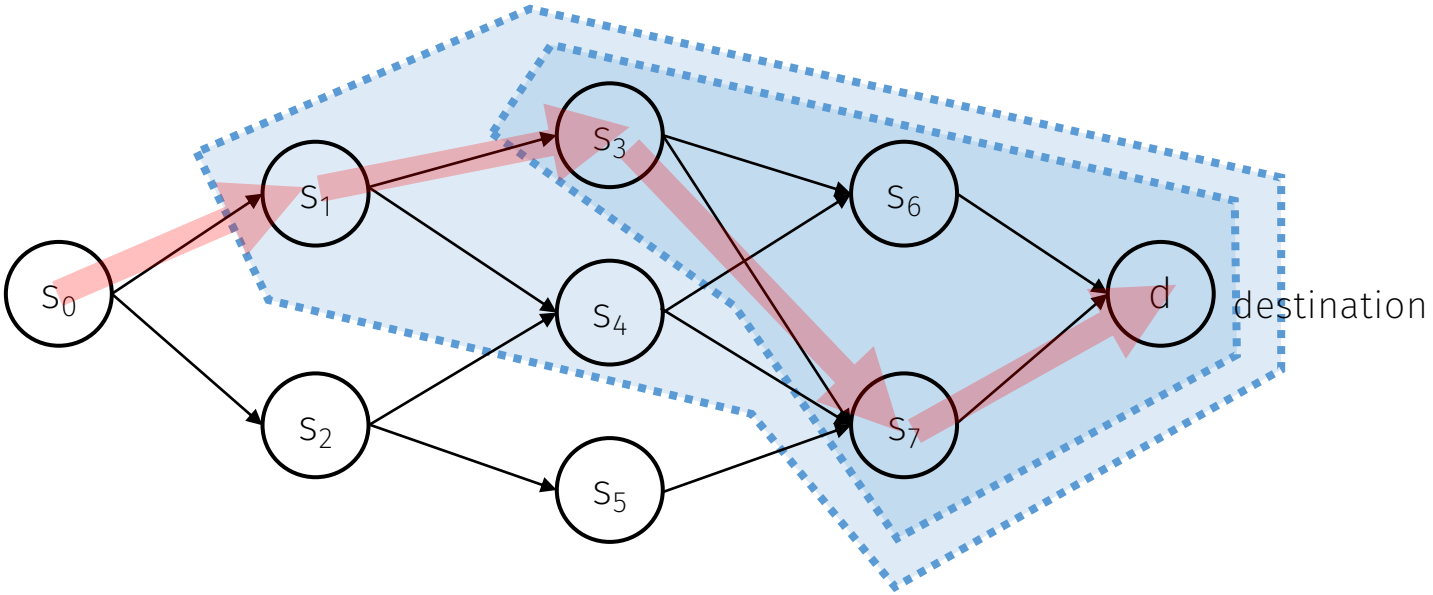
Define choice elements as **state**s and their relationship as a **network**



- A sequence of states is described as a **path** in the network
- An elemental choice problem corresponds to a **substructure of network**
  - E.g., choice (action) set available to S_2 is $\{S_4, S_5\}$

MDP in networks

Define choice elements as **state**s and their relationship as a **network**

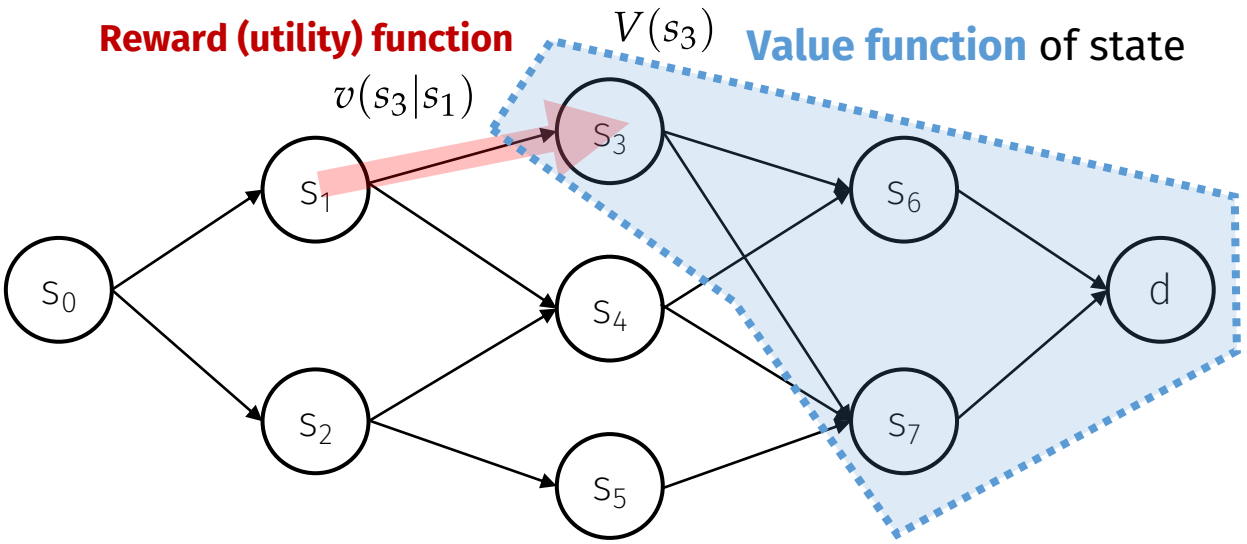


Recursive structure of decision-making:

1. Not direct choice of a sequence, but **sequential choices of states**
   - Elemental choice is the choice of next state given the current state
2. When choosing the next state, traveler **considers possible future states**
   - Trade-off between the current stage utility and the future expected utility

MDP in networks

Define choice elements as **state**s and their relationship as a **network**



Recursive (link-based) path choice model

$$V^d(k) \equiv \mathbb{E}\left[\max_{a \in A(k)} \{v(a|k) + V^d(a) + \mu\epsilon(a|k)\}\right] \quad \Rightarrow \quad \frac{e^{\frac{1}{\mu}V^d(k)}}{z_k} = \sum_{a \in A(k)} \frac{e^{\frac{1}{\mu}\{v(a|k)+V^d(a)\}}}{M_{ka} z_a}$$

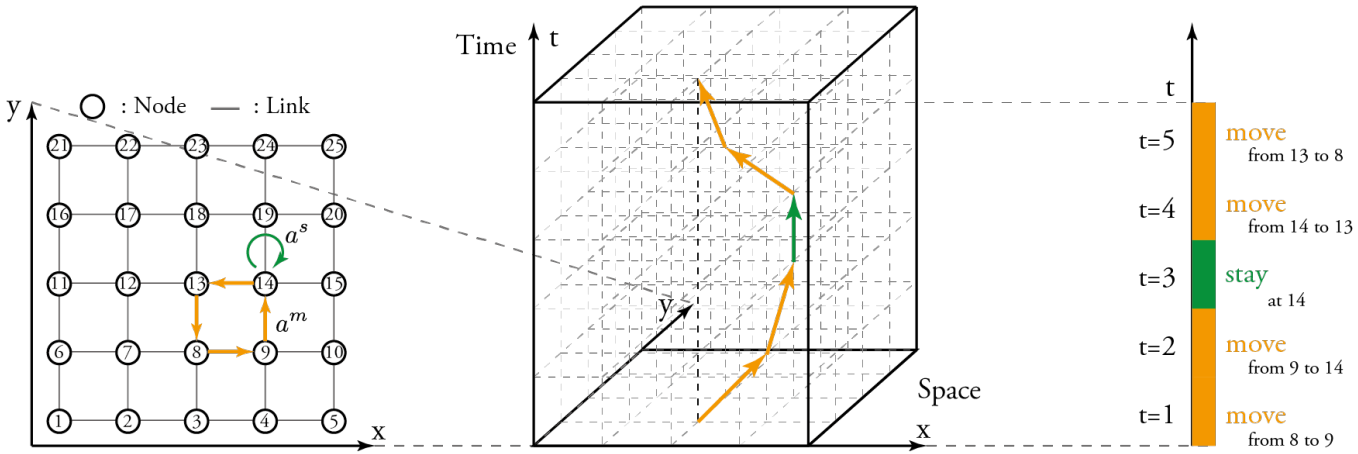
$$\boldsymbol{z} = \mathbf{M}\boldsymbol{z} + \boldsymbol{b} \Leftrightarrow \boldsymbol{z} = (\mathbf{I} - \mathbf{M})^{-1}\boldsymbol{b}$$ System of linear equations

Dynamic decision-making description & **Efficient computation**

Different types of network behavior modeling

MDP path models can describe various types of network path choice problem based on the **definition of states.**

Network	Application	State	Reference
General transportation network	Road, bicycle, walking	Link	Fosgerau+ (2013) Zimmerman+ (2017)
Multi-modal transportation network	MaaS, tourism	(Link, Mode)	De Freitas+ (2019) Tabuchi & Fukuda (2020)
Time-space network	Transit, time-dependent road NW	(Link, Time)	De Moraes Ramos+ (2020) Akamatsu+ (2023)
Activity network	Daily schedule, trip-chain	(Place, Time, Type, Stay/move)	Zimmerman et al. (2018) Oyama & Hato (2016)

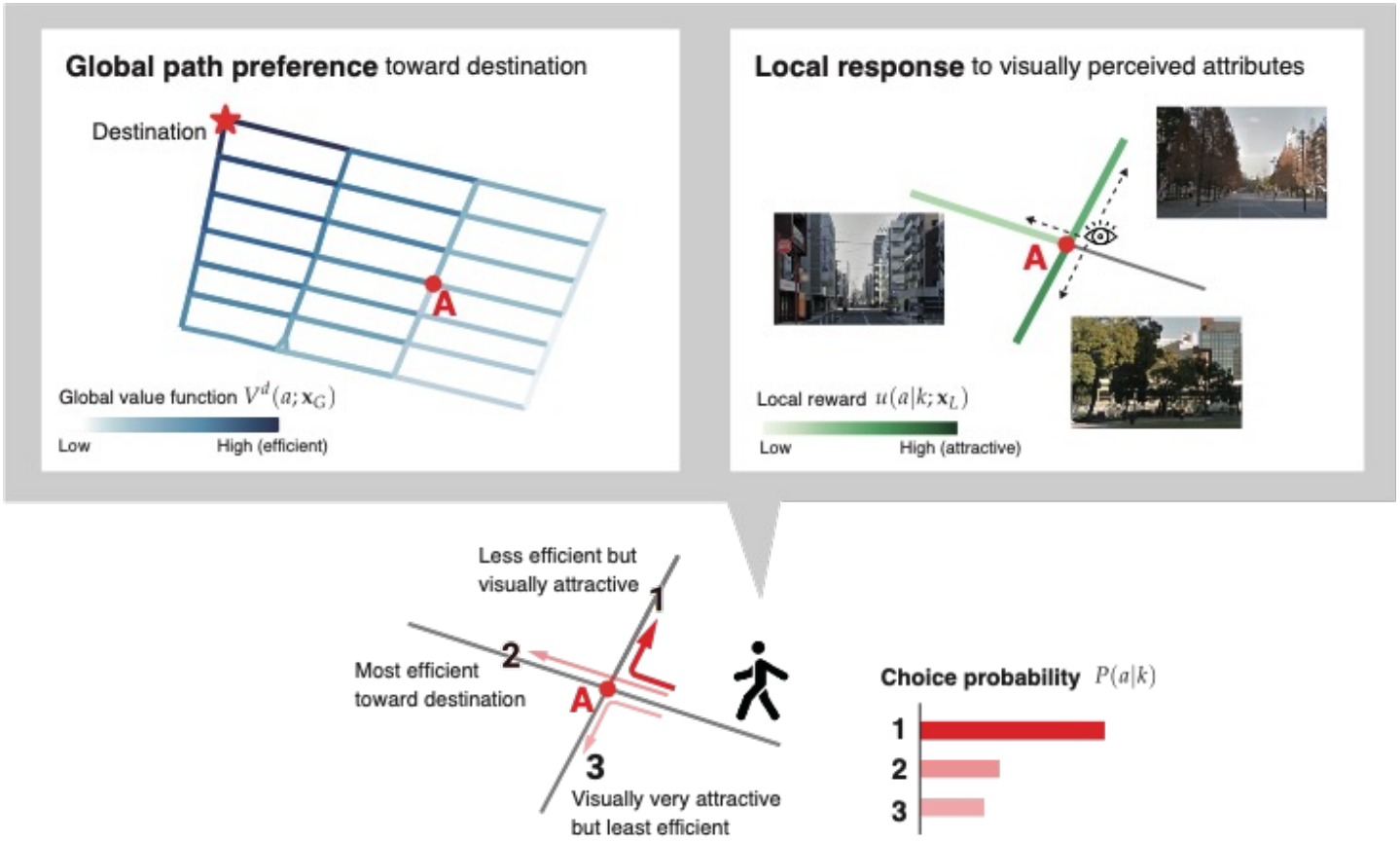


Integrated choice of route, activity places and durations by activity path modeling (Oyama & Hato, 2016)

Global routing and local response

Reward decomposition approach (Oyama, 2023)

$$r(a|s) = r_G(a|s, \mathbf{x}_G) + r_L(a|s, \mathbf{x}_L) \quad \text{and} \quad V(s) = V_s(\mathbf{r}_G)$$



Oyama (2023) Global path preference and local response: A reward decomposition approach for network path choice analysis in the presence of locally perceived attributes. arXiv.

Reward inference as **inverse** problem

Forward decision-making problem: **Reinforcement Learning (RL)**

Find optimal policy p given reward functions r

$$\max_p \mathbb{E}_p \left[\sum_{t=0}^{T} \gamma^t r(a_t|s_t) \right]$$

Inverse estimation problem: **Inverse Reinforcement Learning (IRL)**

Recover reward functions r from expert demonstrations $\mathcal{D}$

Maximum likelihood:

$$\max_\theta \mathbb{E}_{\tau\sim\mathcal{D}}[\log p_\theta(\tau)] = \mathbb{E}_{\tau\sim\mathcal{D}} \left[\sum_{t=0}^{T} \gamma^t r_\theta(s_t, a_t) \right] - \log Z_\theta$$

where $\mathcal{D} = \{\tau_j\}_{j=1}^{N} \qquad \tau_j = \{(s_t^j, a_t^j)\}_{t=0}^{T}$

(Observed sequence of state-action pairs)

Inverse Reinforcement Learning (IRL)

Recursive logit model / Maximum Entropy IRL

- Assuming **deterministic dynamics**
- **Linear-in-parameter** reward functions, i.e., $v_\theta(a|s) = \sum_l \theta_l x_{l,a|s}$

$$\max_\theta \mathbb{E}_{\tau\sim\mathcal{D}}[\log p_\theta(\tau)] = \sum_j \left(\sum_{t=0}^{T-1} v_\theta(s^j_{t+1}|s^j_t) - V_\theta(s^j_0) \right)$$

where

$$p_\theta(a|s) = e^{v_\theta(a|s)+V_\theta(a)-V_\theta(s)}$$

$$V_\theta(s) = \ln \sum_{a\in\mathcal{A}(s)} e^{v_\theta(a|s)+V_\theta(a)}$$

Solution algorithm (e.g., NXFP method; also see my previous year's lecture slides):

- Outer loop: **Parameter update** by non-linear optimization algorithm
  - Econometrics – newton-type methods
  - Machine learning – gradient descent methods
- Inner loop: **Compute value function** for each updated parameter
  - Recursive logit – solving the system of linear equation
  - MaxEnt IRL / nonlinear recursive models – value iteration

Estimation of linear rewards and interpretability

Example: Pedestrian route choice analysis (Oyama, 2023)

$$\begin{cases} v_G(a|k) &= (\beta^G_{\text{len}} + \beta^G_{\text{walk}} x_a^{\text{walk}} + \beta^G_{\text{green}} x_a^{\text{green}}) x_a^{\text{len}} + \beta^G_{\text{cross}} x_a^{\text{cross}} - 20 x_{a|k}^{\text{uturn}} \\ v_L(a|k) &= \beta^L_{\text{green}} x_a^{\text{green}} x_a^{\text{len}} \end{cases}$$

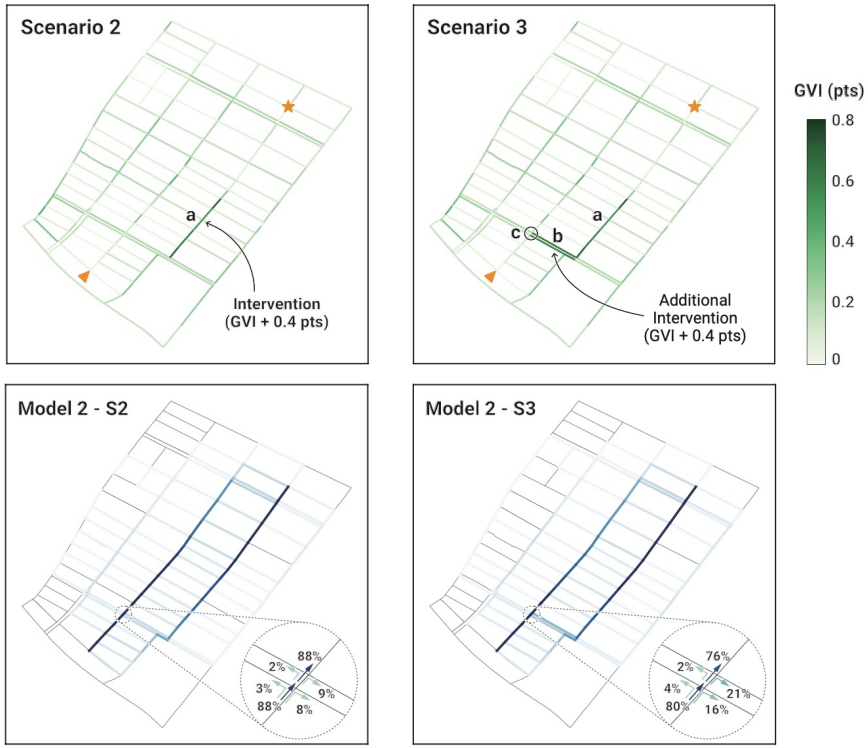
Interaction terms

	Parameter	Estimate	std. err.	t-stat
Global	$\hat{\beta}_{\text{len}}$	-0.290	0.016	-18.62***
	$\hat{\beta}_{\text{cross}}$	-0.816	0.066	-12.35***
	$\hat{\beta}_{\text{walk}}$	0.062	0.010	6.29***
	$\hat{\beta}_{\text{green}}$	-	-	-
Local	$\hat{\beta}_{\text{green}}$	0.096	0.044	2.18**
Scale	$\hat{\mu}_G$	1.280	0.142	1.98**
Path observations				410
Log-likelihood				-1689.6
AIC				3389.2

Willingness-to-walk (+m/100m)

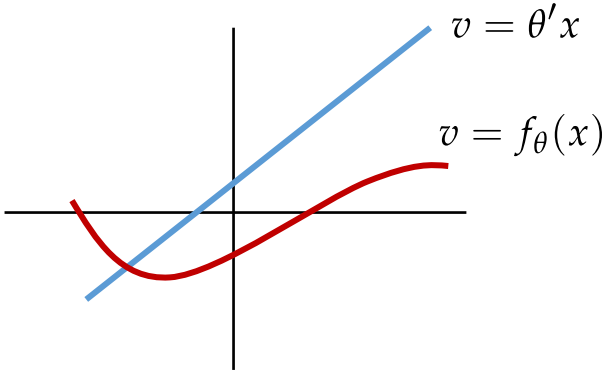
	Model 3	
Variable	Mean	CI
One extra crosswalk along path	-28.2	[-31.5, -25.0]
One meter increase in sidewalk width	21.7	[17.0, 26.0]
10% (0.1 pts) increase in GVI (Global)	-	-
10% (0.1 pts) increase in GVI (Local)	3.65	[0.56, 6.65]

Policy simulation (flow increase)

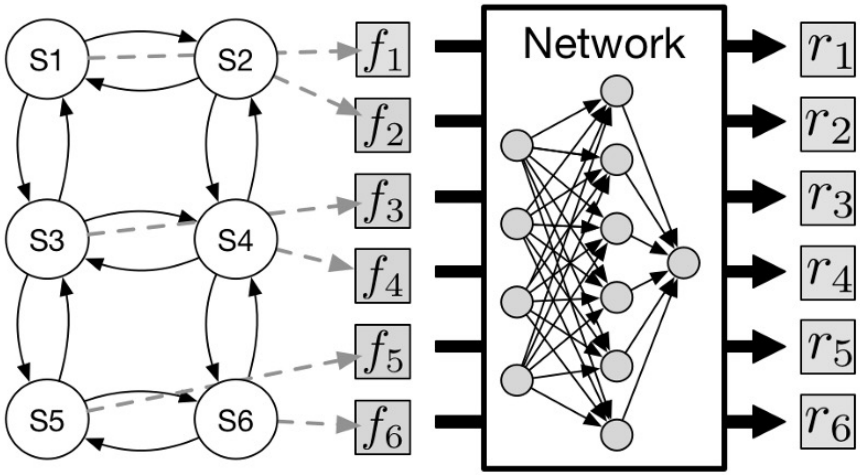


Capturing non-linear rewards by Deep IRL

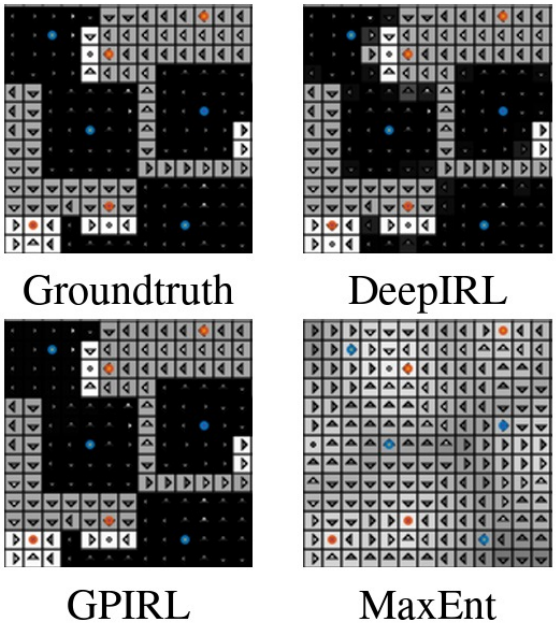
Linear functions often assumed in behavior modeling have **limited expressiveness** of complex travelers' utilities



MaxEnt **Deep IRL** (Wulfmeier et al., 2015)



$$r \approx g(f, \theta_1, \theta_2, ..., \theta_n)$$
$$= g_1(g_2(...(g_n(f, \theta_n), ...), \theta_2), \theta_1).$$

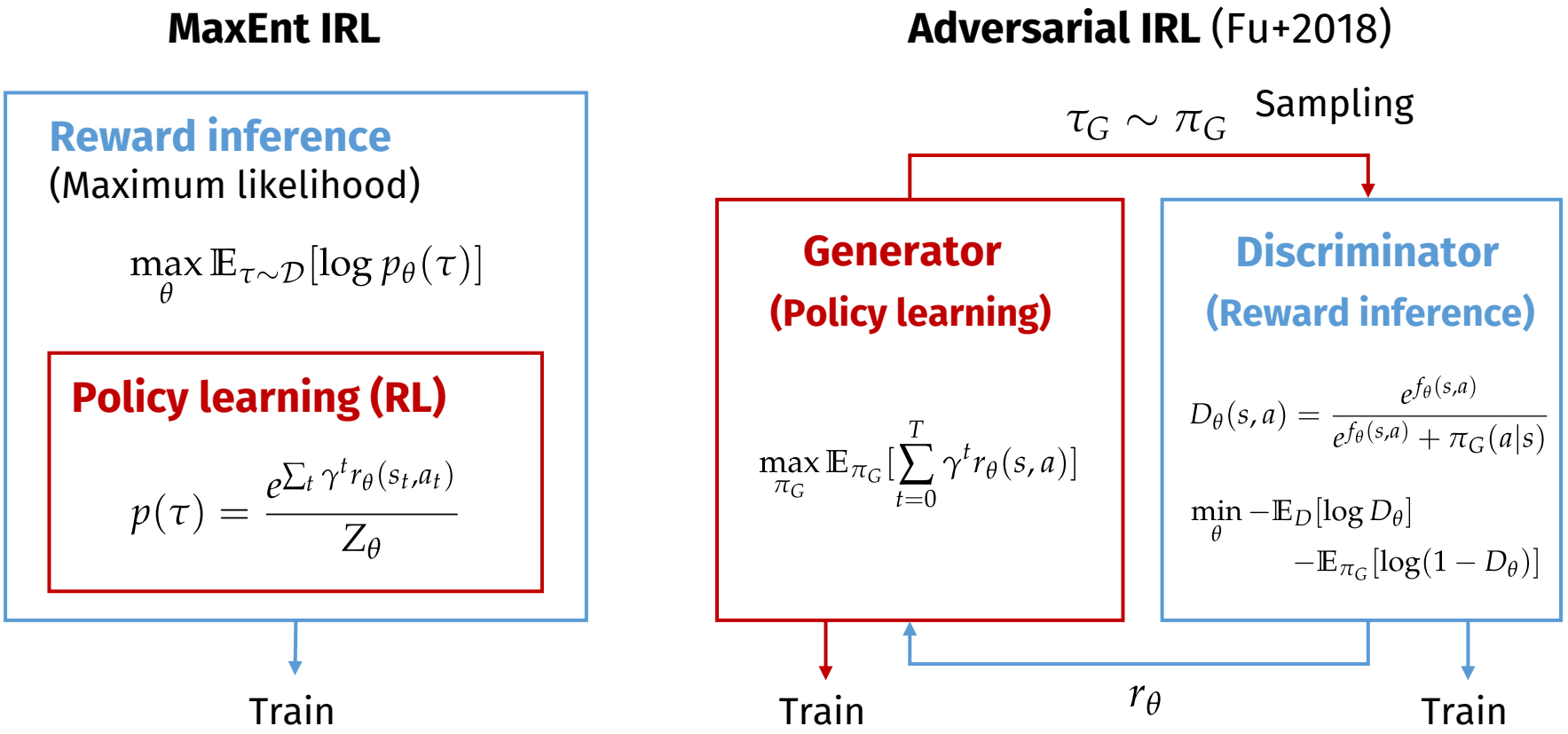


Wulfmeier, M., Ondruska, P., & Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888.

Adversarial Inverse Reinforcement Learning (AIRL)

Partition function *Z* for IRL is intractable (cannot be computed) when

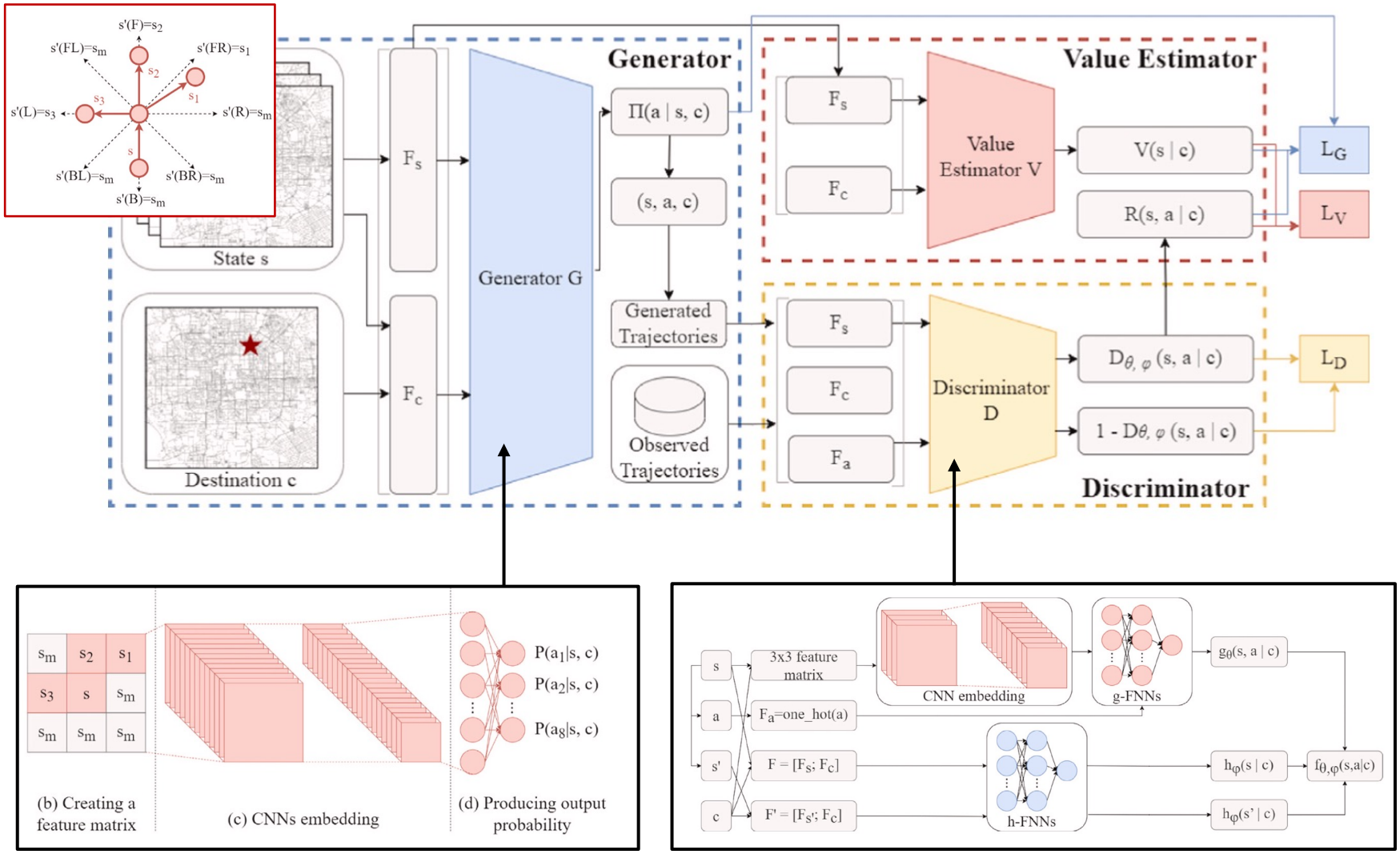
- State-action spaces are large or continuous
- Environment dynamics are stochastic and unknown



Fu, J., Luo, K., & Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. arXiv preprint arXiv:1710.11248.

Adversarial Inverse Reinforcement Learning (AIRL)

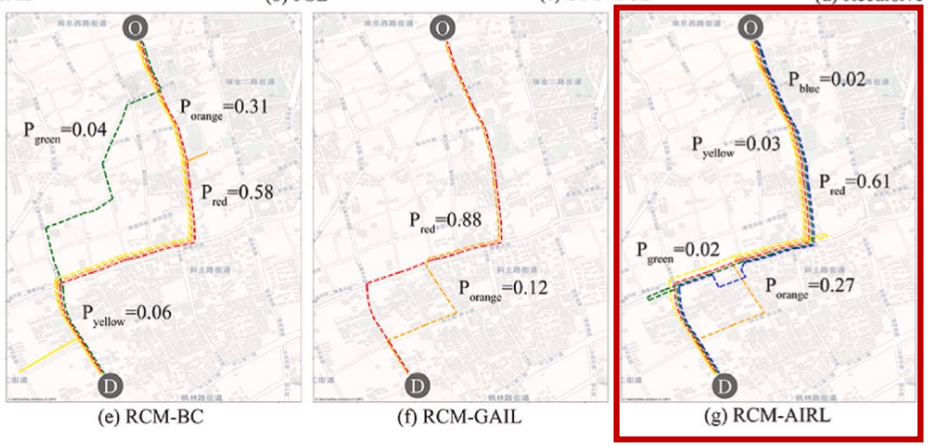
Route choice application by Zhao and Liang (2023)



Zhao and Liang (2023) A deep inverse reinforcement learning approach to route choice modeling with context-dependent rewards. TRC: 104079.

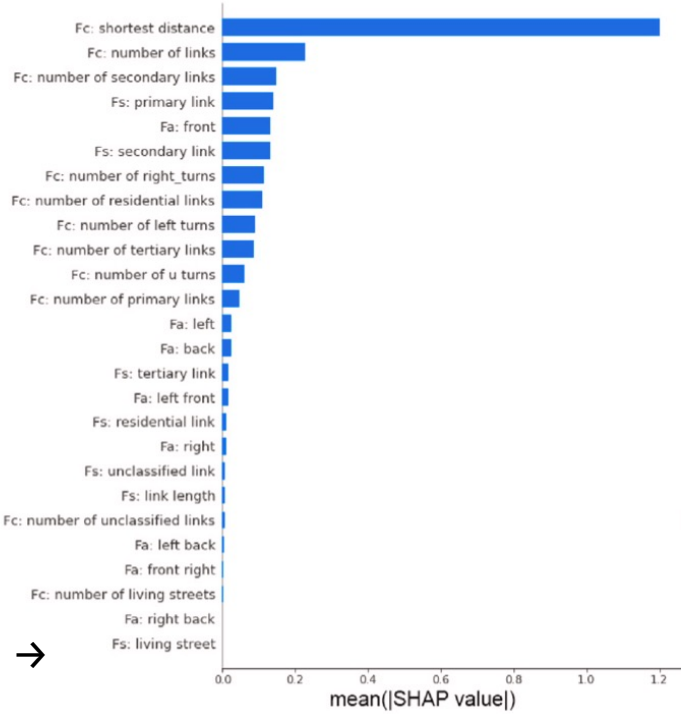
Adversarial Inverse Reinforcement Learning (AIRL)

Route choice application by Zhao and Liang (2023)



← Predictability

Interpretability by SHAP →



Multi-agent AIRL (MA-AIRL)

Reward inference considering **interactions between agents** (Yu+, 2019)

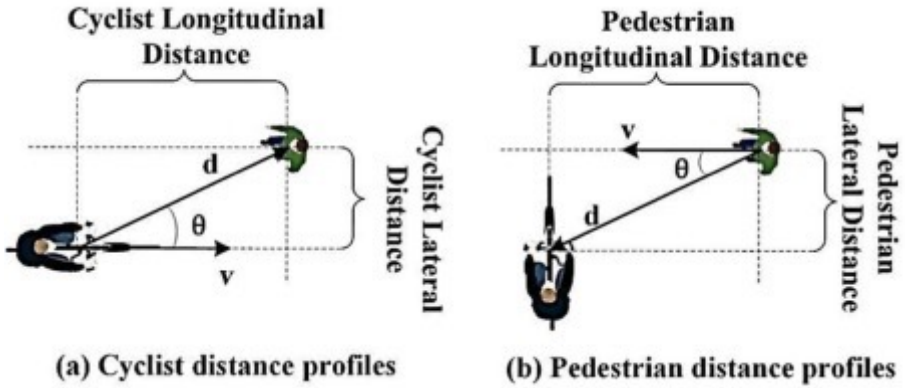
Logit stochastic best response equilibrium

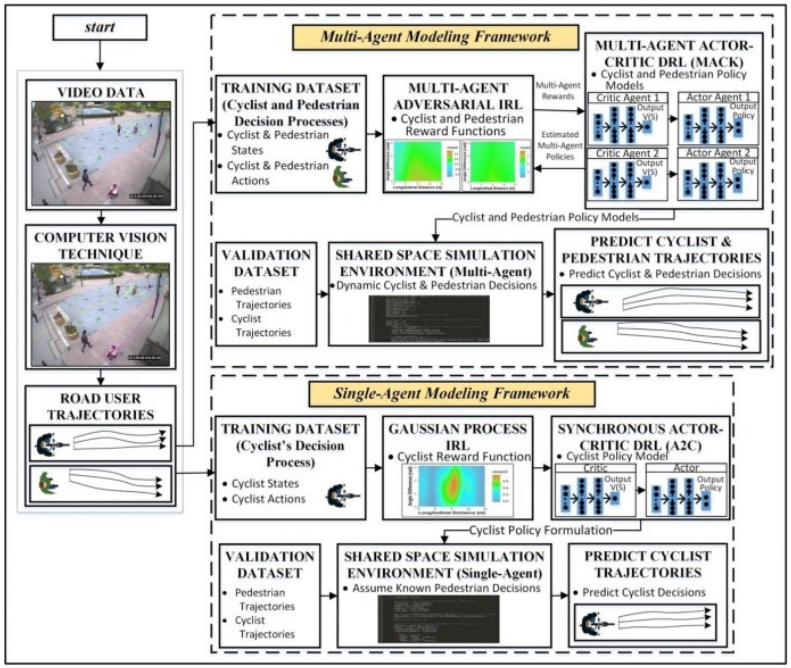
$$p_i(a_i|\mathbf{a}_{-i} = \mathbf{z}_{-i}) = \frac{\exp(\mu Q_i^{\pi}(s, a_i, \mathbf{z}_{-i}))}{\sum_{a_i'} \exp(\mu Q_i^{\pi}(s, a_i', \mathbf{z}_{-i}))}$$

Maximum pseudolikelihood:

$$\max_{\theta} \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=1}^{T} \sum_{i=1}^{N} \log \pi_i^t(a_i^t | \mathbf{a}_{-i}^t, s^t) \right]$$

Cyclist-pedestrian interaction modeling (Alsaleh & Sayed, 2021)



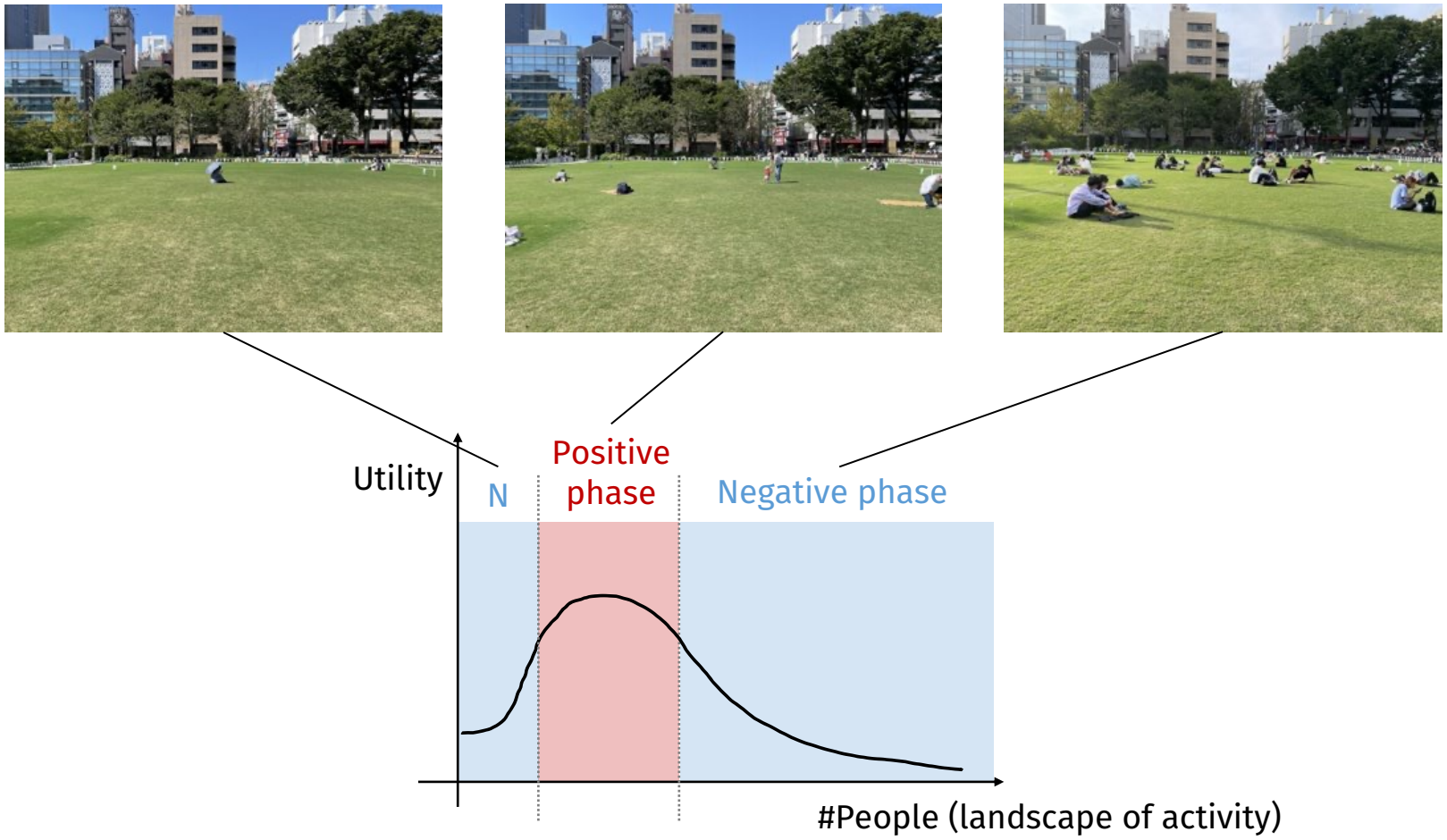


Yu, L., Song, J., & Ermon, S. (2019). Multi-agent adversarial inverse reinforcement learning. *International Conference on Machine Learning*: 7194-7201.

Alsaleh, R., & Sayed, T. (2021). Markov-game modeling of cyclist-pedestrian interactions in shared spaces: A multi-agent adversarial inverse reinforcement learning approach. *Transportation research part C: emerging technologies*, *128*, 103191.

Multi-agent AIRL (MA-AIRL)

Deep MA-AIRL may allow us to analyze **complex (non-linear & non-monotone) social interactions between people** who share the same space, as well as complex structure of spatial attractiveness.



古川智浩 (2023) 「人のいる風景」が生み出す公共空間の魅力とその両面性. 2022年度 芝浦工業大学 卒業論文.

Summary

1. **Markov decision process**

   Modeling framework of agent's sequential decision-making

2. **MDP in networks**

   Recursive logit as an efficient & consistent path choice modeling

3. **Reward inference as inverse problem**

   Learn rewards based on observed trajectories

4. **Adversarial Inverse reinforcement learning**

   Scalable method applicable to various problems including route choice and multi-agent interactions

Questions ?

oyama@shibaura-it.ac.jp