Variational Inference for Inverse Reinforcement Learning with Gaussian Processes

Paulius Dilkas

26th January 2019

H. Kretzschmar, M. Spies, C. Sprunk and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

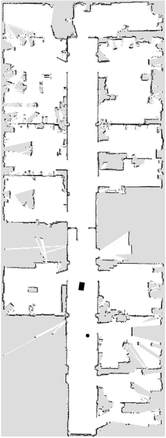
H. Kretzschmar, M. Spies, C. Sprunk and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

H. Kretzschmar, M. Spies, C. Sprunk and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

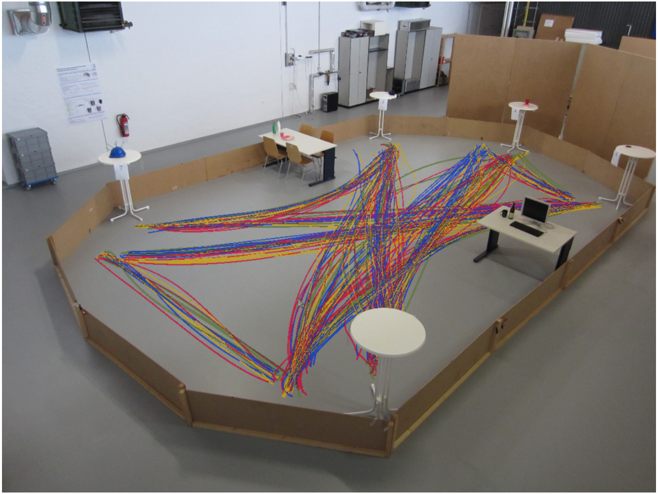
H. Kretzschmar, M. Spies, C. Sprunk and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

Inverse Reinforcement Learning (IRL)

Model (MDP)



Demonstrations



H. Kretzschmar, M. Spies, C. Sprunk and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

Definition (Markov Decision Process)

An MDP is a set $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathbf{r}\}$ that consists of:

- states $\mathcal{S}$
- actions $\mathcal{A}$
- transition function $\mathcal{T} \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$
- discount factor $\gamma \in [0, 1)$
- reward function/vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ (or $r \colon \mathcal{S} \to \mathbb{R}$)

Definition (Markov Decision Process)

An MDP is a set $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathbf{r}\}$ that consists of:

- states $\mathcal{S}$
- actions $\mathcal{A}$
- transition function $\mathcal{T} \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$
- discount factor $\gamma \in [0, 1)$
- reward function/vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ (or $r \colon \mathcal{S} \to \mathbb{R}$)

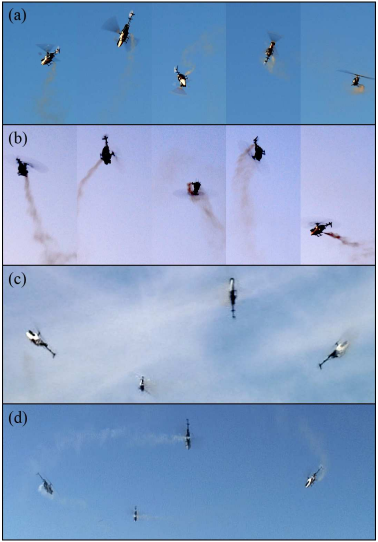
Definition (Inverse Reinforcement Learning (Russell 1998))

Given:

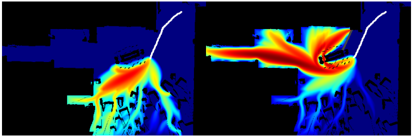
- $\mathcal{M} \setminus \{\mathbf{r}\}$,
- demonstrations $\mathcal{D} = \{\zeta_i\}_{i=1}^{N}$, where $\zeta_i = \{(s_{i,t}, a_{i,t})\}_{t=1}^{T}$,
- features $\mathbf{X} \in \mathbb{R}^{|\mathcal{S}| \times d}$,

find $\mathbf{r}$.

Other Applications



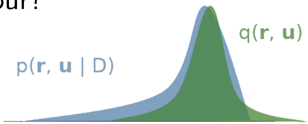
P. Abbeel, A. Coates, M. Quigley and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight", in *NIPS*, 2006



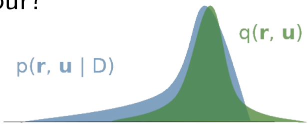
B. D. Ziebart, N. D. Ratliff, G. Gallagher, C. Mertz, K. M. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey and S. S. Srinivasa, "Planning-based prediction for pedestrians", in *IROS*, 2009

K. D. Bogert and P. Doshi, "Multi-robot inverse reinforcement learning under occlusion with interactions", in *AAMAS*, 2014

- Has the model learned optimal behaviour?
- Can it recognise its own weak spots?
- Solution: variational inference (VI)



- Has the model learned optimal behaviour?
- Can it recognise its own weak spots?
- Solution: variational inference (VI)



Outline for the rest of the talk

- Maximum causal entropy and stochastic policies
- Reward function as a Gaussian process (GP)
- Variational approximation of the posterior distribution
- Theoretical results: how can we compute the gradient?
- Empirical results: does it work?
- Conclusions: what have we achieved?

Maximum Causal Entropy

Standard MDP

$$V_{\mathbf{r}}(s) := r(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') V_{\mathbf{r}}(s')$$

Maximum Causal Entropy (Linearly Solvable) MDP[1]

$$V_{\mathbf{r}}(s) := \log \sum_{a \in \mathcal{A}} \exp \left(r(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') V_{\mathbf{r}}(s') \right)$$

[1] B. D. Ziebart, J. A. Bagnell and A. K. Dey, "Modeling interaction via the principle of maximum causal entropy", in *ICML*, 2010.

Reward Function as a Gaussian Process

Automatic Relevance Determination Kernel

For any two states $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \lambda_0 \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\intercal \boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j) - \mathbb{1}[i \neq j]\sigma^2 \operatorname{tr}(\boldsymbol{\Lambda})\right)$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$, $\sigma^2 = 10^{-2}/2$,

$$\mathbb{1}[b] = \begin{cases} 1 & \text{if } b \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

Reward Function as a Gaussian Process

Inducing Points

- $m \ll |\mathcal{S}|$ states,
- their features $\mathbf{X}_{\mathbf{u}}$
- and rewards $\mathbf{u}$.

The GP Then Gives Gives...

- Kernel/covariance matrices: $\mathbf{K}_{\mathbf{u},\mathbf{u}}$, $\mathbf{K}_{\mathbf{r},\mathbf{u}}$, $\mathbf{K}_{\mathbf{r},\mathbf{r}}$
- Prior probabilities:
  - $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}})$
  - $p(\mathbf{r} \mid \mathbf{u}) = \mathcal{N}(\mathbf{r}; \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{r},\mathbf{r}} - \mathbf{K}_{\mathbf{r},\mathbf{u}}^{\intercal}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{K}_{\mathbf{r},\mathbf{u}})$

Variational Inference

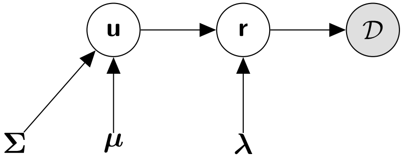
The posterior distribution

$$p(\mathbf{r}, \mathbf{u} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{r}) p(\mathbf{r} \mid \mathbf{u}) p(\mathbf{u})}{p(\mathcal{D})}$$

can be approximated with $q(\mathbf{r}, \mathbf{u}) = q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u})$, where

- $q(\mathbf{r} \mid \mathbf{u}) = p(\mathbf{r} \mid \mathbf{u})$
- $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

Variational Inference



Goal: minimise the *Kullback-Leibler divergence*:

$$D_{\mathrm{KL}}(q(\mathbf{r},\mathbf{u}) \parallel p(\mathbf{r},\mathbf{u} \mid \mathcal{D})) = \mathbb{E}_{q(\mathbf{r},\mathbf{u})}[\log q(\mathbf{r},\mathbf{u}) - \log p(\mathbf{r},\mathbf{u} \mid \mathcal{D})]$$

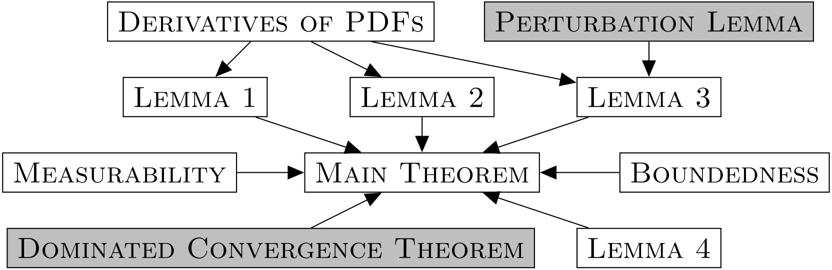
Equivalently, maximise the *evidence lower bound*:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\mathbf{r},\mathbf{u})}[\log p(\mathcal{D},\mathbf{r},\mathbf{u}) - \log q(\mathbf{r},\mathbf{u})] \\ &= \mathbf{t}^\intercal \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\boldsymbol{\mu} - \mathbb{E}_{q(\mathbf{r},\mathbf{u})}[v] - D_{\mathrm{KL}}(q(\mathbf{u}) \parallel p(\mathbf{u}))\end{aligned}$$

where

$$v = \sum_{i=1}^{N}\sum_{t=1}^{T} V_{\mathbf{r}}(s_{i,t}) - \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s_{i,t}, a_{i,t}, s') V_{\mathbf{r}}(s').$$

Theoretical Results



Main Theorem

Theorem

Whenever the derivative exists,

$$\frac{\partial}{\partial t} \iint V_{\mathbf{r}}(s) q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u}) \, d\mathbf{r} \, d\mathbf{u} = \iint \frac{\partial}{\partial t} [V_{\mathbf{r}}(s) q(\mathbf{r} \mid \mathbf{u}) q(\mathbf{u})] \, d\mathbf{r} \, d\mathbf{u},$$

where t is any element of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, or $\boldsymbol{\lambda}$.

Dominated Convergence Theorem

Theorem

Let $(X, \mathcal{M}, \mu)$ be a measure space and $\{f_n\}$ a sequence of measurable functions on X for which $\{f_n\} \to f$ pointwise almost everywhere on X, and the function f is measurable. Assume there is a non-negative function g that is integrable over X and dominates the sequence $\{f_n\}$ on X in the sense that

$$|f_n| \leq g \text{ almost everywhere on } X \text{ for all } n.$$

Then f is integrable over X and

$$\lim_{n\to\infty} \int_X f_n \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu.$$

Other Results

Seeing V as $V\colon \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$...

Proposition (Measurability)

MDP value functions $V(s)\colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}$ (for $s \in \mathcal{S}$) are Lebesgue measurable.

Proposition (Boundedness)

If the initial values of the MDP value function satisfy the following bound, then the bound remains satisfied throughout value iteration:

$$|V_{\mathbf{r}}(s)| \leq \frac{\|\mathbf{r}\|_\infty + \log|\mathcal{A}|}{1-\gamma}.$$

Other Results

Lemma

Let $i, j = 1, \ldots, m$, *and let*

$$c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (\Sigma_{i,j} - \epsilon, \Sigma_{i,j} + \epsilon) \subset \mathbb{R}$$

be a function with a codomain arbitrarily close to $\Sigma_{i,j}$. *Then every element of*

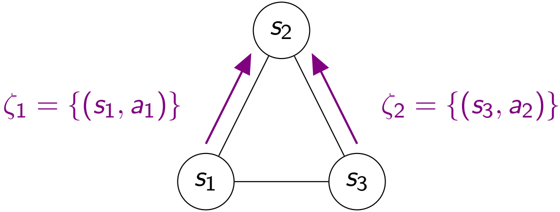
$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\Sigma}}\right|_{\Sigma_{i,j} = c(\mathbf{r}, \mathbf{u})}$$

has upper and lower bounds of the form $q(\mathbf{u})d(\mathbf{u})$, *where* $d(\mathbf{u}) \in \mathbb{R}_2[\mathbf{u}]$.

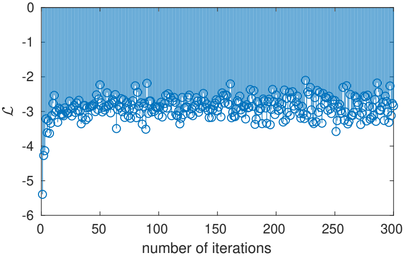
Experiments

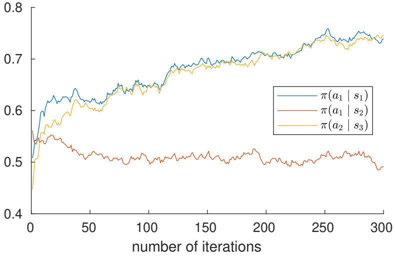
Source: https://www.flickr.com/photos/nationaleyeinstitute/9955408003/, Creative Commons Attribution 2.0 Generic license

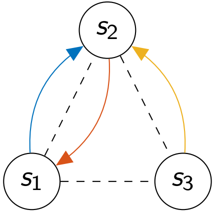
Experimental Scenario



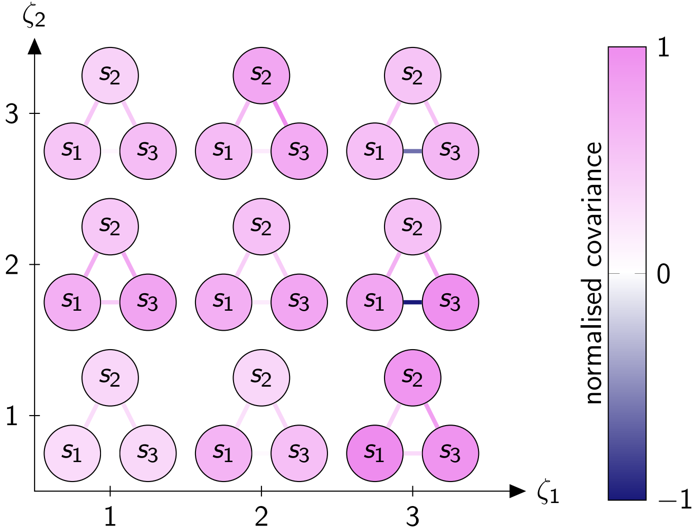
Convergence



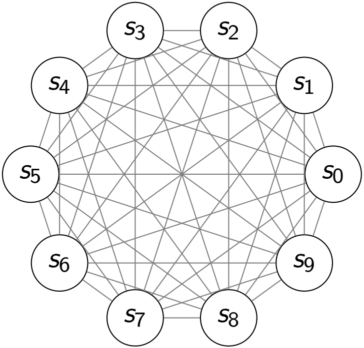


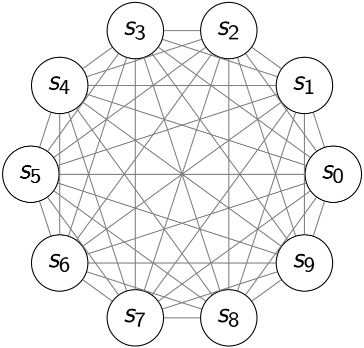


Covariance: Attempt 1

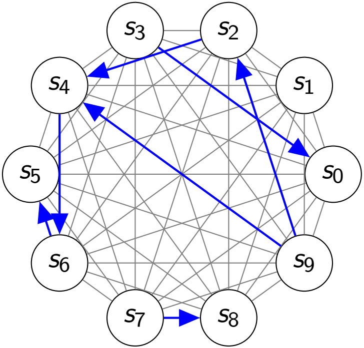


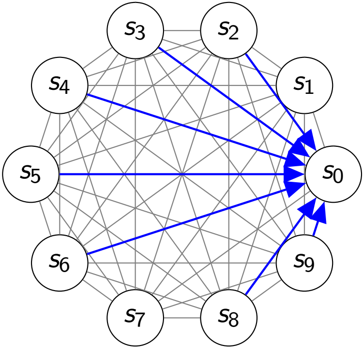
Covariance: Attempt 2 (with Cliques!)



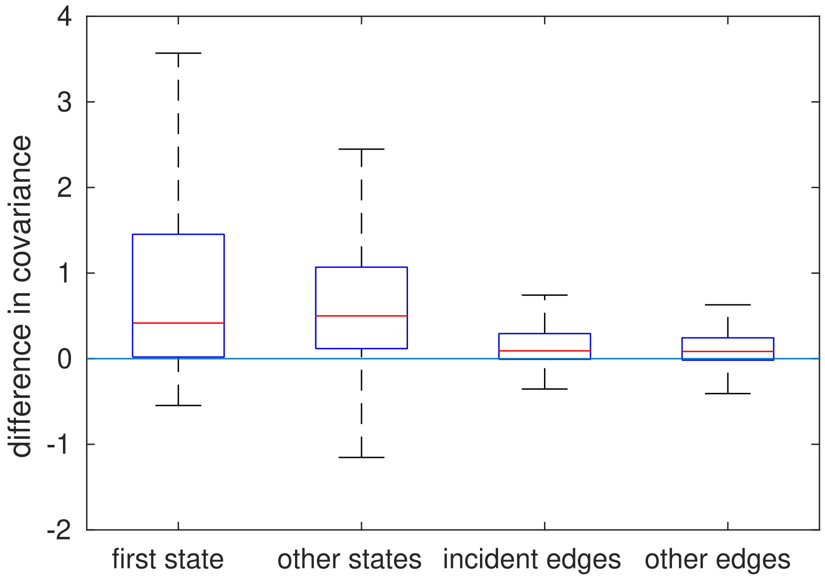


Covariance: Attempt 2 (with Cliques!)





Covariance: Attempt 2 (with Cliques!)



Conclusions

- VI can be applied to IRL without additional assumptions
  - proof
  - other theoretical results
  - implementation
- Covariances can be used to compare clear vs. noisy data
  - but not the amount of data

Conclusions

- VI can be applied to IRL without additional assumptions
  - proof
  - other theoretical results
  - implementation
- Covariances can be used to compare clear vs. noisy data
  - but not the amount of data

Thank You!