Variational Inference for Inverse Reinforcement Learning with Gaussian Processes

Paulius Dilkas

27th March 2019

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

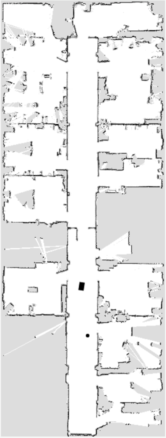
H. Kretzschmar, M. Spies, C. Sprunk *et al.*, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

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

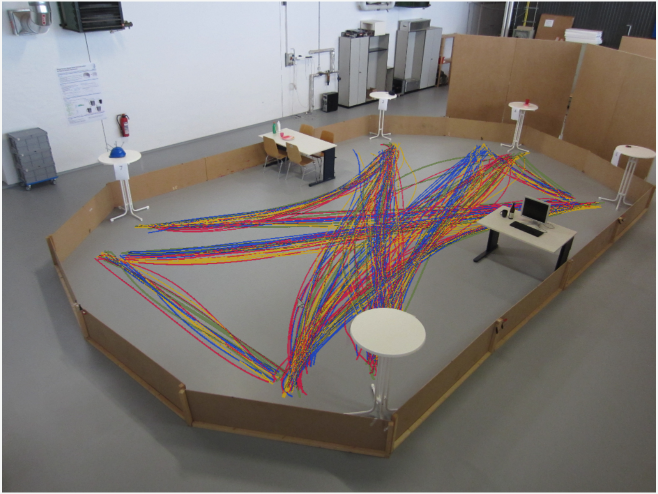
H. Kretzschmar, M. Spies, C. Sprunk *et al.*, "Socially compliant mobile robot navigation via inverse reinforcement learning", *I. J. Robotics Res.*, 2016

Inverse Reinforcement Learning (IRL)

Model (MDP)



Demonstrations



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

Inverse Reinforcement Learning (COLT 1998)

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

Inverse Reinforcement Learning (COLT 1998)

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

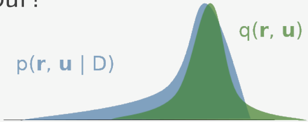
Definition (Inverse Reinforcement Learning Problem)

Given:

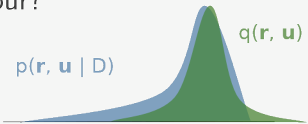
- $\mathcal{M} \setminus \{\mathbf{r}\}$,
- demonstrations $\mathcal{D} = \{\zeta_i\}_{i=1}^N$, where $\zeta_i = \{(s_{i,t}, a_{i,t})\}_{t=1}^T$,
- features $\mathbf{X} \in \mathbb{R}^{|\mathcal{S}| \times d}$,

find $\mathbf{r}$.

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



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



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?
- Further work: what comes next?

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 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 $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$,

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

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

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

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 Approximation

Previous Work

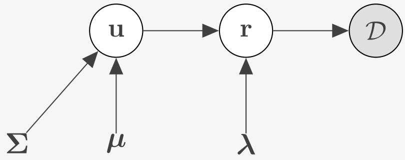
- Levine et al. (2011) assume that $\mathbf{r} = \mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}$ and maximise the likelihood
- Jin et al. (2017) add more assumptions and use a deep GP model
- Wulfmeier et al. (2015) use a neural network

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

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 Approximation



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*:

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

where

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

Mathematical Preliminaries

Vector norms	Matrix norms
$\|\mathbf{x}\|_1 = \sum_i \lvert x_i \rvert$	$\|\mathbf{A}\|_p = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$
$\|\mathbf{x}\|_\infty = \max_i \lvert x_i \rvert$	$\|\mathbf{A}\|_\infty = \max_i \sum_j \lvert A_{i,j} \rvert$

Lemma (Perturbation Lemma)

Let $\|\cdot\|$ *be any matrix norm, and let* $\mathbf{A}$ *and* $\mathbf{E}$ *be matrices such that* $\mathbf{A}$ *is invertible and* $\|\mathbf{A}^{-1}\|\|\mathbf{E}\| < 1$*, then* $\mathbf{A} + \mathbf{E}$ *is invertible, and*

$$\|(\mathbf{A} + \mathbf{E})^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{E}\|}.$$

Theoretical Results

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

Proposition

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

Proposition

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

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

Theoretical Results

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 scalar part of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, or $\boldsymbol{\lambda}$.

A Note on Polynomials

Definition

Let $\mathbb{R}_d[\mathbf{x}]$ denote the vector space of polynomials with degree at most d, where variables are elements of $\mathbf{x}$, and coefficients are in $\mathbb{R}$.

Example

$$\begin{aligned}\mathbb{R}_2[\mathbf{x}] \supset \{&2x_1^2 + \pi x_2,\\ &x_1 x_2,\\ &-3x_1 + 1,\\ &0\}\end{aligned}$$

Helpful Lemmas

Lemma

$$\int \|\mathbf{r}\|_\infty q(\mathbf{r} \mid \mathbf{u})\, d\mathbf{r} \le a + \|\mathbf{K}_{\mathbf{r},\mathbf{u}}^\intercal \mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1} \mathbf{u}\|_1,$$

where a is a constant independent of $\mathbf{u}$.

Lemma

Let $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ be an arbitrary bounded function. Then, for $i = 0, \ldots, d$,

$$\left. \frac{\partial q(\mathbf{r} \mid \mathbf{u})}{\partial \lambda_i} \right|_{\lambda_i = c(\mathbf{r},\mathbf{u})}$$

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

Helpful Lemmas

Lemma

Let $c : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m \to (a, b) \subset \mathbb{R}$ *be an arbitrary bounded function. Then, for* $i = 1, \ldots, m$*, every element of*

$$\left.\frac{\partial q(\mathbf{u})}{\partial \boldsymbol{\mu}}\right|_{\mu_i = c(\mathbf{r}, \mathbf{u})}$$

has upper and lower bounds of the form $q(\mathbf{u})d(\mathbf{u})$*, where* $d(\mathbf{u}) \in \mathbb{R}_1[\mathbf{u}]$.

Helpful Lemmas

Lemma

Let $i, j = 1, \ldots, m$*, and let* $\epsilon > 0$ *be arbitrary. Furthermore, let*

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

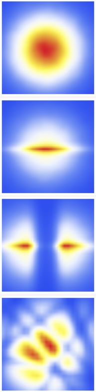
be a function with a codomain arbitrarily close to $\Sigma_{i,j}$*. Then every element of*

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

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

Further Work

- More flexible models using...
  - normalizing flows[1]
  - spectral kernels[2]
- Faster GP inference and MDP solving
- IRL in the context of reinforcement learning
- Interplay between rewards and stochastic/deterministic policies



---

[1] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows", in *ICML*, 2015.

[2] A. Wilson and R. Adams, "Gaussian process kernels for pattern discovery and extrapolation", in *ICML*, 2013.