University of Stuttgart
Germany

Maximum Entropy Inverse Reinforcement Learning

Algorithms for Imitation Learning

Maximilian Luz

Summer Semester 2019

MLR / IPVS

Outline

Nomenclature

Nomenclature (i)

Markov Decision Process (MDP)

$$
\begin{aligned}
S &= \{s_i\}_i && \text{States} \\
A &= \{a_i\}_i && \text{Actions} \\
T &= p(s_{t+1} \mid s_t, a_t) && \text{Transition dynamics} \\
R &: S \to \mathbb{R} && \text{Reward}
\end{aligned}
$$

Trajectories & Demonstrations

$$
\begin{aligned}
\tau &= ((s_1, a_1), (s_2, a_2), \ldots, s_{|\tau|}) && \text{Trajectory} \\
\mathcal{D} &= \{\tau_i\}_i && \text{Demonstrations}
\end{aligned}
$$

Nomenclature (ii)

Features

$$\phi : S \rightarrow \mathbb{R}^d \qquad \text{with} \qquad \phi(\tau) = \sum_{s_t \in \tau} \phi(s_t)$$

Policies

$\pi(a_j \mid s_i)$	Policy (stochastic)
π^L	Learner Policy
π^E	Expert Policy

Basis

Feature Expectation Matching

Idea: Learner should visit same features as expert (in expectation).

Feature Expectation Matching [Abbeel and Ng 2004]

$$\mathbb{E}_{\pi^E}\left[\phi(\tau)\right] = \mathbb{E}_{\pi^L}\left[\phi(\tau)\right]$$

Note: We want to find reward $R : S \to \mathbb{R}$ defining $\pi^L(a \mid s)$ and thus $p(\tau)$.

$$\mathbb{E}_{\pi^L}\left[\phi(\tau)\right] = \sum_{\tau \in \mathcal{T}} p(\tau) \cdot \phi(\tau)$$

Observation: Optimality for *linear* (unknown) reward [Abbeel and Ng 2004].

$\Rightarrow \qquad R(s) = \boldsymbol{\omega}^\top \phi(s), \qquad \boldsymbol{\omega} \in \mathbb{R}^d :$ Reward parameters

Feature Expectation Matching: Problem

Problem: Multiple (infinite) solutions $\Rightarrow$ ill-posed (Hadamard).

- Reward shaping [Ng et al. 1999]:
  - Multiple reward functions R lead to same policy π.

Idea (Ziebart et al. 2008):

- Regularize by maximizing entropy $H(p)$.
  - But why?

Shannon's Entropy

Entropy $H(p)$

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

$$\begin{aligned} x \in \mathcal{X} :& \quad \text{Event} \\ p(x) :& \quad \text{Probability of occurrence} \\ -\log_2 p(x) :& \quad \text{Optimal encoding length} \end{aligned}$$

Expected information received when observing $x \in \mathcal{X}$.

$\Rightarrow$ Measure of uncertainty.

No uncertainty, $H(p)$ minimal.

Uniformly random, $H(p)$ maximal.

Principle of Maximum Entropy [Jaynes 1957]

Consider: A problem with solutions $p, q, \ldots$ (e.g. feature expectation matching)

$\Rightarrow$ p, q represent *partial* information.

$\Rightarrow$ Maximizing entropy minimizes bias.

Maximum Entropy IRL

Constrained Optimization Problem

Problem Formulation

$$
\begin{aligned}
&\arg\max_p && H(p) && \text{(entropy)} \\
&\text{subject to} && \mathbb{E}_{\pi^E}\left[\phi(\tau)\right] = \mathbb{E}_{\pi^L}\left[\phi(\tau)\right], && \text{(feature matching)} \\
& && \textstyle\sum_{\tau \in \mathcal{T}} p(\tau) = 1, \quad \forall \tau \in \mathcal{T} : p(\tau) > 0 && \text{(prob. distr.)}
\end{aligned}
$$

Solution: Deterministic Dynamics

Solution via Lagrange multipliers [Ziebart et al. 2008]:

$$p(\tau) \propto \exp\left(R(\tau)\right) \qquad \text{where} \qquad R(\tau) = \boldsymbol{\omega}^\top \boldsymbol{\phi}(\tau)$$

Lagrange multipliers for feature matching

Deterministic transition dynamics:

$$p(\tau \mid \boldsymbol{\omega}) = \underbrace{\frac{1}{Z(\boldsymbol{\omega})}}_{\text{normalization}} \exp\Big(\underbrace{\boldsymbol{\omega}^\top \boldsymbol{\phi}(\tau)}_{\text{reward}}\Big) \qquad \text{with} \qquad \underbrace{Z(\boldsymbol{\omega}) = \sum_{\tau \in \mathcal{T}} \exp\Big(\boldsymbol{\omega}^\top \boldsymbol{\phi}(\tau)\Big)}_{\text{partition function}}$$

Solution: Stochastic Dynamics

Stochastic transition dynamics:

assumes limited transition randomness

$$p(\tau \mid \boldsymbol{\omega}) = \underbrace{\frac{1}{Z_s(\boldsymbol{\omega})} \exp\left(\boldsymbol{\omega}^\top \phi(\tau)\right)}_{\propto \text{ deterministic}} \underbrace{\prod_{s_t, a_t, s_{t+1} \in \tau} p\left(s_{t+1} \mid s_t, a_t\right)}_{\text{combined transition probability}}$$

via adaption of deterministic solution [Ziebart et al. 2008].

Problem: Adaption introduces bias [Osa et al. 2018; Ziebart 2010]:

$$\tilde{R}(\tau) = \boldsymbol{\omega}^\top \phi(\tau) + \sum_{s_t, a_t, s_{t+1} \in \tau} \log p\left(s_{t+1} \mid s_t, a_t\right)$$

Solution: Maximum Causal Entropy IRL (Ziebart 2010, not covered here).

Likelihood and Gradient

Obtain parameters by maximizing **Likelihood:**

$$\omega^* = \arg\max_{\omega} \mathcal{L}(\omega) = \arg\max_{\omega} \sum_{\tau \in \mathcal{D}} \log p(\tau \mid \omega)$$

Observation:

- Maximizing likelihood equiv. to minimizing KL-divergence [Bishop 2006].
  $\Rightarrow$ M-projection onto manifold of maximum entropy distributions [Osa et al. 2018].
- Convex, can be optimized via gradient ascent.

Gradient [Ziebart et al. 2008]:

computationally infeasible

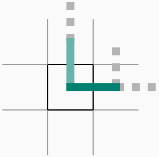
$$\begin{aligned} \nabla \mathcal{L}(\omega) &= \mathbb{E}_{\mathcal{D}}[\phi(\tau)] - \sum_{\tau \in \mathcal{T}} p(\tau \mid \omega)\, \phi(\tau) \\ &= \mathbb{E}_{\mathcal{D}}[\phi(\tau)] - \sum_{s_i \in S} D_{s_i}\, \phi(s_i) \end{aligned}$$

"count" features in $\mathcal{D}$

state visitation frequency

Observation:

$$\pi_{\mathrm{ME}}(a_j \mid s_i, \boldsymbol{\omega}) \;\propto \sum_{\tau \in \mathcal{T}:\; s_i, a_j \in \tau_{t=1}} \overbrace{p\,(\tau \mid \boldsymbol{\omega})}^{\text{can be computed via } R}$$



Idea: Split into sub-problems.

1. *Backward Pass:* Compute policy $\pi_{\mathrm{ME}}(a \mid s, \boldsymbol{\omega})$.
2. *Forward Pass:* Compute state visitation frequency from $\pi_{\mathrm{ME}}(a \mid s, \boldsymbol{\omega})$.

State Visitation Frequency: Backward Pass

Observation:

$$\pi_{\text{ME}}(a_j \mid s_i, \boldsymbol{\omega}) \propto \sum_{\tau \in \mathcal{T}:\ s_i, a_j \in \tau_{t=1}} p(\tau \mid \boldsymbol{\omega})$$

Idea:

$$\pi_{\text{ME}}(a_j \mid s_i, \boldsymbol{\omega}) = \frac{Z_{s_i,a_j}}{Z_{s_i}}$$

normalization

recursively expand observation

$$Z_{s_i,a_j} = \sum_{s_k \in S} p(s_k \mid s_i, a_j) \cdot \exp\left(\boldsymbol{\omega}^\top \phi(s_i)\right) \cdot Z_{s_k}, \qquad Z_{s_i} = \sum_{a_j \in A} Z_{s_i,a_j}$$

Algorithm:

1. Initialize $Z_{s_k} = 1$ for all terminal states $s_k \in S_{\text{terminal}}$.
2. Compute Z_{s_i,a_j} and Z_{s_i} by recursively backing-up from terminal states.
3. Compute $\pi_{\text{ME}}(a_i \mid s_i, \boldsymbol{\omega})$.

Parallels to value-iteration.

Idea: Propagate starting-state probabilities $p_0(s)$ forward via policy $\pi_{\mathrm{ME}}(a \mid s, \boldsymbol{\omega})$.

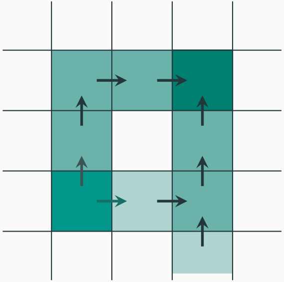
Algorithm:

1. Initialize $D_{s_i,0} = p_0(s) = p(\tau \in \mathcal{T} : s \in \tau_{t=1})$.
2. Recursively compute

$$D_{s_k,t+1} = \sum_{s_i \in S} \sum_{a_j \in A} D_{s_i,t} \cdot \pi_{\mathrm{ME}} \left(a_j \mid s_i\right) \cdot p \left(s_k \mid a_j, s_i\right)$$

3. Sum up over t, i.e.

$$D_{s_i} = \sum_{t=0,\dots} D_{s_i,t}$$



Summary

Algorithm: Iterate until convergence:

1. Compute policy $\pi_{\mathrm{ME}}(a \mid s, \boldsymbol{\omega})$ *(forward pass)*.
2. Compute state visitation frequency D_{s_i} *(backward pass)*.
3. Compute gradient $\nabla \mathcal{L}(\boldsymbol{\omega})$ of likelihood.
4. Gradient-based optimization step, e.g.: $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} + \eta \nabla \mathcal{L}(\boldsymbol{\omega})$.

Assumptions:

- Known transition dynamics $T = p(s_{t+1} \mid s_t, a_t)$.
- Limited transition randomness.
- Linear reward $R(s) = \boldsymbol{\omega}^\top \phi(s)$.

Other Drawbacks:

- Need to "solve" MDP once per iteration.
- Reward bias for stochastic transition dynamics.

Extensions

- Maximum Causal Entropy IRL [Ziebart 2010]
- Maximum Entropy Deep IRL [Wulfmeier et al. 2015]
- Maximum Entropy IRL in Continuous State Spaces with Path Integrals [Aghasadeghi and Bretl 2011]

Demonstration

github.com/qzed/irl-maxent

References

Abbeel, Pieter and Andrew Y. Ng (2004). “Apprenticeship Learning via Inverse Reinforcement Learning”. In: *Proc. 21st Intl. Conference on Machine Learning (ICML ’04).*

Aghasadeghi, N. and T. Bretl (Sept. 2011). “Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals”. In: *Intl. Conference on Intelligent Robots and Systems (IORS 2011)*, pp. 1561–1566.

Bishop, Christopher M. (Aug. 17, 2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc.

Jaynes, E. T. (May 1957). “Information Theory and Statistical Mechanics”. In: *Physical Review* 106.4, pp. 620–630.

Ng, Andrew Y., Daishi Harada, and Stuart J. Russell (1999). “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”. In: *Proc. 16th Intl. Conference on Machine Learning (ICML ’99)*, pp. 278–287.

Osa, Takayuki et al. (2018). “An Algorithmic Perspective on Imitation Learning”. In: *Foundations and Trends in Robotics* 7.1-2, pp. 1–179.

Wulfmeier, Markus, Peter Ondruska, and Ingmar Posner (2015). “Deep Inverse Reinforcement Learning”. In: *Computing Research Repository*. arXiv: `1507.04888`.

Ziebart, Brian D. (2010). “Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy”. PhD thesis. Carnegie Mellon University.

Ziebart, Brian D. et al. (2008). “Maximum Entropy Inverse Reinforcement Learning”. In: *Proc. 23rd AAAI Conference on Artificial Intelligence (AAAI ’08)*, pp. 1433–1438.