Journal Club: "Continuous Inverse Optimal Control with Locally Optimal Examples"

Stéphane Caron

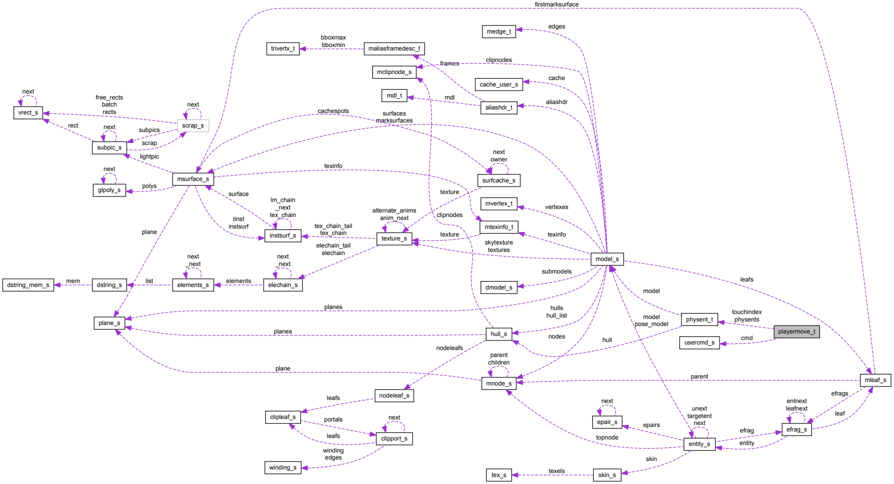
October 31, 2012

MaxEnt

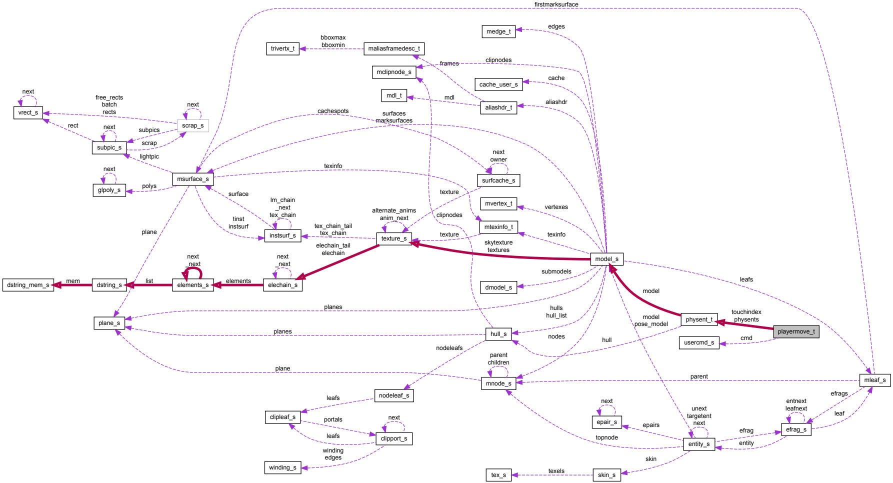
Paper

Maximum Entropy Inverse Reinforcement Learning
Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey.
AAAI Conference on Artificial Intelligence (AAAI 2008).

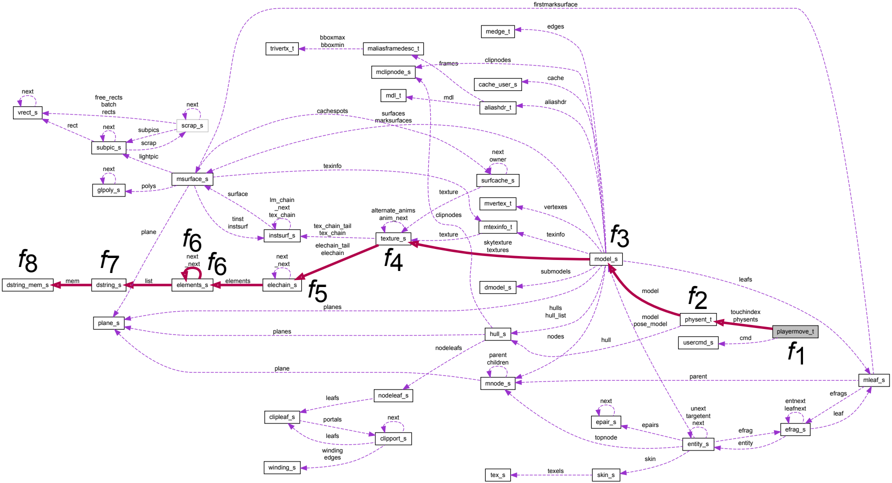
Environment



Environment



Environment



MaxEnt Framework

- States s
- Actions a
- Transition distribution $T : \{P_T(s'|s, a)\}$
- Feature vector for state s: $\mathbf{f}_s$
- Path $\zeta = ((s_0, a_0), \ldots, (s_T, a_T))$
- Feature counts $\mathbf{f}_\zeta = \sum_{s \in \zeta} \mathbf{f}_s$

MaxEnt Framework

- States s
- Actions a
- Transition distribution $T : \{P_T(s'|s, a)\}$
- Feature vector for state s: $\mathbf{f}_s$
- Path $\zeta = ((s_0, a_0), \ldots, (s_T, a_T))$
- Feature counts $\mathbf{f}_\zeta = \sum_{s \in \zeta} \mathbf{f}_s$

Rewards

Linear model: $r(\mathbf{f}_\zeta) := \theta^\top \mathbf{f}_\zeta$.

Goal: learn a pdf P over trajectories $\{\zeta\}$.

Demonstrations

- Trajectories $\left\{\tilde{\zeta}_i,\ 1 \leq i \leq m\right\}$
- Empirical feature count: $\tilde{\mathbf{f}} := \frac{1}{m} \sum_i \mathbf{f}_{\tilde{\zeta}_i}$

Demonstrations

- Trajectories $\left\{\tilde{\zeta}_i,\ 1 \leq i \leq m\right\}$
- Empirical feature count: $\tilde{\mathbf{f}} := \frac{1}{m}\sum_i \mathbf{f}_{\tilde{\zeta}_i}$

Matching on feature expectations

Choose P s.t., for $\zeta \sim P$,

$$\mathbb{E}(\mathbf{f}_\zeta) := \sum_{\zeta} P(\zeta)\mathbf{f}_\zeta \ = \ \tilde{\mathbf{f}}.$$

Observation: infinitely many solutions.

Enters Entropy

$$P(\zeta|\theta) = \frac{1}{Z(\theta)} e^{\theta^\top \mathbf{f}_\zeta}$$

Enters Entropy

$$P(\zeta|\theta) = \frac{1}{Z(\theta)} e^{\theta^\top \mathbf{f}_\zeta}$$

Nice:

- Same reward $\Rightarrow$ same probability
- Higher reward $\Rightarrow$ exponentially preferred

Enters Entropy

$$P(\zeta|\theta) = \frac{1}{Z(\theta)} e^{\theta^\top \mathbf{f}_\zeta}$$

Nice:

- Same reward $\Rightarrow$ same probability
- Higher reward $\Rightarrow$ exponentially preferred

Not nice:

- $Z(\theta) := \sum_\zeta e^{\theta^\top \mathbf{f}_\zeta}$ heavy to compute

Learning

Maximum likelihood over example trajectories:

$$\theta^* = \arg\max_\theta \; L(\theta) = \arg\max_\theta \sum_{\tilde{\zeta}_i} \log P(\tilde{\zeta}_i|\theta).$$

Learning

Maximum likelihood over example trajectories:

$$\theta^* = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \sum_{\tilde{\zeta}_i} \log P(\tilde{\zeta}_i | \theta).$$

Compute θ through gradient descent:

$$\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_{\zeta} P(\zeta|\theta) \mathbf{f}_{\zeta}.$$

Pipeline Summary

Reward model: $r(\mathbf{f}_\zeta) = \theta^\top \mathbf{f}_\zeta$

- Demonstrations $\tilde{\zeta}_i$

Pipeline Summary

Reward model: $r(\mathbf{f}_\zeta) = \theta^\top \mathbf{f}_\zeta$

- Demonstrations $\tilde{\zeta}_i$
- Feature count vector $\tilde{\mathbf{f}}$

Pipeline Summary

Reward model: $r(\mathbf{f}_\zeta) = \theta^\top \mathbf{f}_\zeta$

- Demonstrations $\tilde{\zeta}_i$
- Feature count vector $\tilde{\mathbf{f}}$
- From a given θ: $\forall \zeta, P(\zeta|\theta)$ given by MaxEnt

Pipeline Summary

Reward model: $r(\mathbf{f}_\zeta) = \theta^\top \mathbf{f}_\zeta$

- Demonstrations $\tilde{\zeta}_i$
- Feature count vector $\tilde{\mathbf{f}}$
- From a given θ: $\forall \zeta, P(\zeta|\theta)$ given by MaxEnt
- Gradient descent: $\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_\zeta P(\zeta|\theta)\mathbf{f}_\zeta$

Pipeline Summary

Reward model: $r(\mathbf{f}_\zeta) = \theta^\top \mathbf{f}_\zeta$

- Demonstrations $\tilde{\zeta}_i$
- Feature count vector $\tilde{\mathbf{f}}$
- From a given θ: $\forall \zeta, P(\zeta|\theta)$ given by MaxEnt
- Gradient descent: $\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_\zeta P(\zeta|\theta)\mathbf{f}_\zeta$
- "Optimal" θ^*

Today's Paper

Paper

Sergey Levine and Vladlen Koltun
Continuous Inverse Optimal Control with Locally Optimal Examples
Proceedings of the 29th International Conference on Machine Learning (2012)

Differences

- Dynamics function (known to the learner):

$$\mathbf{x}_t = \mathcal{F}(\mathbf{x}_{t-1}, \mathbf{u}_t)$$

- More general reward functions:

$$r(\mathbf{u}) = \sum_t r(\mathbf{x}_t, \mathbf{u}_t)$$

Contribution

Same model:

$$P(\mathbf{u}|\mathbf{x}_0) = \frac{1}{Z(\mathbf{u})} \exp\left(\sum_t r(\mathbf{x}_t, \mathbf{u}_t)\right)$$

Contribution

Same model:

$$P(\mathbf{u}|\mathbf{x}_0) = \frac{1}{Z(\mathbf{u})} \exp\left(\sum_t r(\mathbf{x}_t, \mathbf{u}_t)\right)$$

Algorithm

Faster computation of $Z(\mathbf{u})$: $O(T)$ instead of $O(T^3)$.

MaxEnt model:

$$P(\mathbf{u}|\mathbf{x}_0) = e^{r(\mathbf{u})} \left[\int e^{r(\tilde{\mathbf{u}})} \mathrm{d}\tilde{\mathbf{u}} \right]^{-1}$$

First-order approximation of $r(\tilde{\mathbf{u}})$:

$$r(\tilde{\mathbf{u}}) \approx r(\mathbf{u}) + (\tilde{\mathbf{u}} - \mathbf{u})^\top \mathbf{g} + \frac{1}{2}(\tilde{\mathbf{u}} - \mathbf{u})^\top \mathbf{H}(\tilde{\mathbf{u}} - \mathbf{u})$$

- inject into integral
- neglect (some) second-order variations: $\frac{\partial^2 \mathbf{x}}{\partial \mathbf{u}^2}$
- do the math...

Approximate log-likelihood:

$$2\mathcal{L} = \mathbf{g}^\top \mathbf{H}^{-1} \mathbf{g} + \log\det(-\mathbf{H}) - \dim(\mathbf{u}) \log(2\pi)$$

Convenient expression when *r* parametrized by θ:

$$\frac{\partial \mathcal{L}}{\partial \theta} = f_{\mathcal{F}}\left(\frac{\partial \mathbf{g}}{\partial \theta}, \frac{\partial \mathbf{H}}{\partial \theta} \right)$$

Pros and cons

Nice:

- Fast approximation of $P(\mathbf{u}|\mathbf{x}_0) \Rightarrow$ bigger domains
- Only needs *locally optimal* examples

Pros and cons

Nice:

- Fast approximation of $P(\mathbf{u}|\mathbf{x}_0) \Rightarrow$ bigger domains
- Only needs *locally optimal* examples

Not nice:

- Learner needs to know/model environment dynamics
- No measure of approximation errors

Other remarks

- Computation time cubic in $\dim(\mathbf{x})$, $\dim(\mathbf{u})$

Other remarks

- Computation time cubic in $\dim(\mathbf{x})$, $\dim(\mathbf{u})$
- High dimensional domains: what about exploration?

Thanks for your attention!