PREFERENCE ELICITATION AND INVERSE REINFORCEMENT LEARNING

CONSTANTIN A. ROTHKOPF & CHRISTOS DIMITRAKAKIS

Frankfurt Institute for Advanced Studies
Frankfurt, Germany
rothkfopf@fias.uni-frankfurt.de

École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
christos.dimitrakakis@epfl.ch

INVERSE REINFORCEMENT LEARNING PROBLEM

PREFERENCE ELICITATION

- **Determine, whether a given decision maker prefers some events to other events, and if so, by how much.**
  (Friedman & Saveage, 1952)

  - **Assumption 1**:
    There exists a **partial ordering** among events, indicating **relative preferences**.
  - **Assumption 2**, *the expected utility hypothesis*:
    We can assign a **numerical utility to each event**, such that events with larger utilities are preferred, then the decision maker's preferred choice from a set of possible gambles will be the gamble with the highest expected utility.

- Then, the corresponding problem is to **determine the numerical utilities for a decision maker**.

INVERSE REINFORCEMENT LEARNING

- **Determine which task a reinforcement learning agent is carrying out in an environment.**
  (Ng & Russell, 2001)

  - **Assumption 1**:
    The transition function describing the environment is given.

  - **Assumption 2**:
    The agent is following a policy which is maximizing the cumulative total discounted reward.

- Then, the corresponding problem is to **determine the reward function** being maximized by the reinforcement learning agent.

FORMALIZATION

- Controlled Markov Process $\nu = (\mathcal{S}, \mathcal{A}, \mathcal{T})$ with state space $\mathcal{S}$, action space $\mathcal{A}$, and **known** transitions:

$$\mathbb{P}_\nu(s_{t+1} \in S \mid s^t, a^t) = \tau(S \mid s_t, a_t), \qquad S \subset \mathcal{S}$$

- Demonstrations consist of states and actions:

$$D \triangleq (a^T, s^T) \qquad s^T \equiv s_1, \ldots, s_T \qquad a^T \equiv a_1, \ldots, a_T$$

- The functional form of the utilities is determined by the definition of the total return in reinforcement learning:

$$U_t \triangleq \sum_{k=t}^{\infty} \gamma^k r_k$$

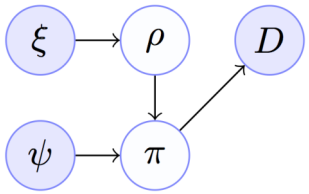
where we consider stochastic reward functions ρ

$$r_t \mid s_t = s, a_t = a \sim \rho(\cdot \mid s, a) \qquad (s, a) \in \mathcal{S} \times \mathcal{A}$$

with discount factor $\gamma \in [0, 1]$

- Note that we could have defined the utilities in other ways:
log(# of coins collected), hyperbolically discounted expected return, …

STATISTICAL MODEL



- Let $\mathcal{R}$ be the space of reward functions and $\mathcal{P}$ the space of policies. Define prior distributions over the space of reward functions ρ and over the space of policies π :

$$\rho \sim \xi(\cdot \mid \nu) \qquad \pi \mid \rho_{\mathrm{a}} = \rho \sim \psi(\cdot \mid \rho, \nu)$$

leading to a joint distribution over reward functions and policies:

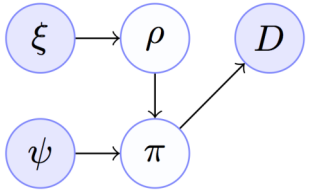
$$\phi(P, R \mid \nu) \triangleq \int_R \psi(P \mid \rho, \nu) \, \mathrm{d}\xi(\rho \mid \nu)$$

- As we can choose the policy prior, a reasonable choice to connect to previous IRL work is to consider stationary softmax policies:

$$\pi_\eta(a_t \mid s_t) = \frac{\exp(\eta Q^*_\mu(s_t, a_t))}{\sum_a \exp(\eta Q^*_\mu(s_t, a))}$$

- Note that we can consider different policy priors.

INFERENCE



- **Lemma 1**: For the above priors, a given controlled Markov process ν and observed state and action sequences $s^T a^T$ where the actions are drawn from a reactive policy π , the posterior measure on reward functions is:

$$\xi(B|s^T, a^T, \nu) = \frac{\int_B \int_{\mathcal{P}} \pi(a^T|s^T)\, \mathrm{d}\psi(\pi|\rho,\nu)\, \mathrm{d}\xi(\rho|\nu)}{\int_{\mathcal{R}} \int_{\mathcal{P}} \pi(a^T|s^T)\, \mathrm{d}\psi(\pi|\rho,\nu)\, \mathrm{d}\xi(\rho|\nu)} \qquad B \subset \mathcal{R}$$

  where $\pi(a^T \mid s^T) = \prod_{t=1}^{T} \pi(a_t|s_t)$

- Note that the state transitions $\tau(s_t|a_{t-1}, s_{t-1})$ do not appear in this expression.

AN MCMC PROCEDURE FOR INFERENCE

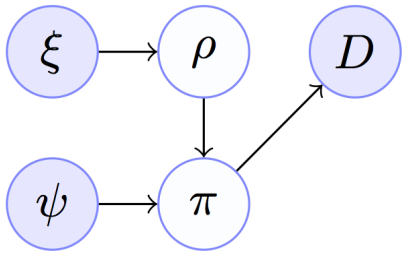
- Performing Metropolis-Hastings to sample from some distribution with density $f(x)$ using a proposal distribution with conditional density $g(\tilde{x} \mid x)$ has the form:

$$x_{(k+1)} = \begin{cases} \tilde{x}, & \text{w.p. } \min\left\{1, \frac{f(\tilde{x})/g(\tilde{x}|x_{(k)})}{f(x_{(k)})/g(x_{(k)}|\tilde{x})}\right\} \\ x_{(k)}, & \text{otherwise.} \end{cases}$$

- For the problem at hand this leads to $x = (\rho, \pi)$, $f(x) = \phi(\rho, \pi \mid s^T, a^T, \nu)$ and using independent proposal distributions $g(x) = \phi(\rho, \pi|\nu)$

$$\frac{\phi(\tilde{\rho}, \tilde{\pi} \mid s^T, a^T, \nu)}{\phi(\rho, \pi \mid s^T, a^T, \nu)} = \frac{\mathbb{P}_{\nu,\tilde{\pi}}(s^T, a^T)\phi(\tilde{\rho}, \tilde{\pi} \mid \nu)}{\mathbb{P}_{\nu,\pi_{(k)}}(s^T, a^T)\phi(\rho_{(k)}, \pi_{(k)} \mid \nu)}$$

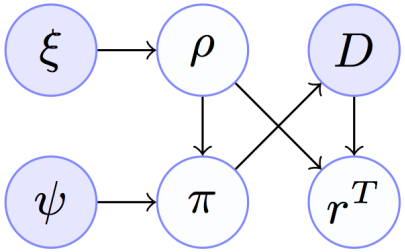
MH sampler



Algorithm 1 MH: Direct Metropolis-Hastings sampling from the joint distribution $\phi(\pi, \rho \mid a^T, s^T)$.

for $k = 1, \ldots$ **do**
  $\tilde{\rho} \sim \xi(\rho \mid \nu)$.
  $\tilde{\eta} \sim \mathit{Gamma}(\zeta, \theta)$
  $\tilde{\pi} = \mathit{Softmax}(\tilde{\rho}, \tilde{\eta}, \tau)$
  $\tilde{p} = \mathbb{P}_{\nu,\tilde{\pi}}(s^T, a^T)/[\xi(\rho \mid \nu) f_{\mathit{Gamma}}(\tilde{\eta}; \zeta, \theta)]$.
  w.p. $\min\left\{1, \tilde{p}/p_{(k-1)}\right\}$ **do**
    $\pi_{(k)} = \tilde{\pi}$, $\eta_{(k)} = \tilde{\eta}$, $\rho_{(k)} = \tilde{\rho}$, $p_{(k)} = \tilde{p}$.
  else
    $\pi_{(k)} = \pi_{(k-1)}$, $\eta_{(k)} = \eta_{(k-1)}$, $\rho_{(k)} = \rho_{(k-1)}$, $p_{(k)} = p_{(k-1)}$.
  done
end for

Gibbs sampler



Algorithm 2 G-MH: Two stage Gibbs sampler with an MH step

for $k = 1, \ldots$ **do**
  $\tilde{\rho} \sim \xi(\rho \mid r^T_{(k-1)}, \nu)$.
  $\tilde{\eta} \sim \mathit{Gamma}(\zeta, \theta)$
  $\tilde{\pi} = \mathit{Softmax}(\tilde{\rho}, \tilde{\epsilon}, \tau)$
  $\tilde{p} = \mathbb{P}_{\nu,\tilde{\pi}}(s^T, a^T)/[\xi(\rho \mid \nu) f_{\mathit{Gamma}}(\tilde{\eta}; \zeta, \theta)]$.
  w.p. $\min\left\{1, \tilde{p}/p_{(k-1)}\right\}$ **do**
    $\pi_{(k)} = \tilde{\pi}$, $\eta_{(k)} = \tilde{\eta}$, $\rho_{(k)} = \tilde{\rho}$, $p_{(k)} = \tilde{p}$.
  else
    $\pi_{(k)} = \pi_{(k-1)}$, $\eta_{(k)} = \eta_{(k-1)}$, $\rho_{(k)} = \rho_{(k-1)}$, $p_{(k)} = p_{(k-1)}$.
  done
  $r^T_{(k)} \mid s^T, a^T \sim \rho^T_{(k)}(s^T, a^T)$
end for

EMPIRICAL EVALUATION

- Experiments in two RL domains:
  - random MDPs of different sizes
  - random Maze tasks of different sizes
- Comparison of the performance via the L1 loss:

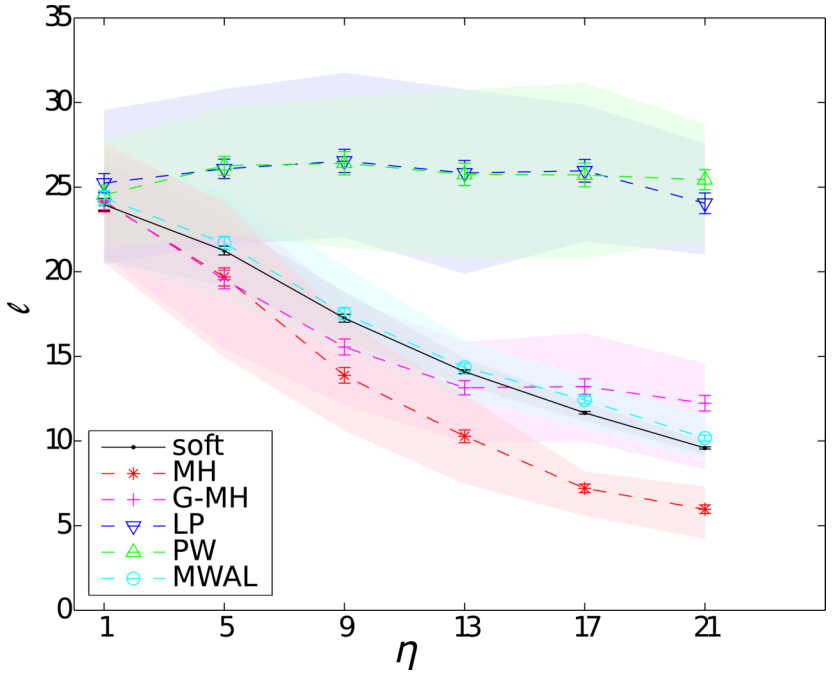
$$\ell(\pi) \triangleq \sum_{s \in \mathcal{S}} V_\mu^*(s) - V_\mu^\pi(s)$$

where

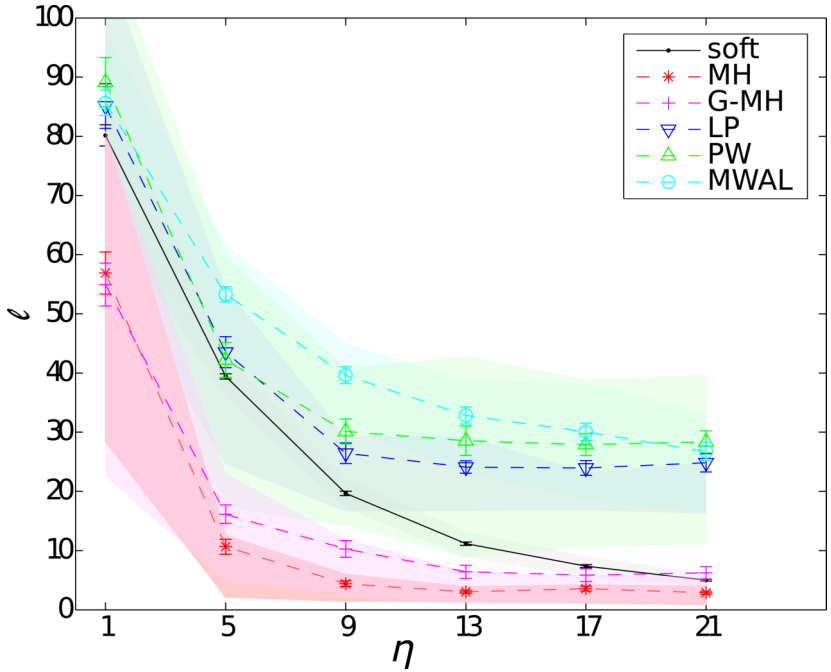
$$V_\mu^*(s) \triangleq \max_a Q_\mu^*(s,a) \qquad V_\mu^\pi(s) \triangleq \mathbb{E}_\pi Q_\mu^\pi(s,a)$$

- Comparison to other IRL methods:
  - linear program (Ng&Russell, 2000)
  - 'Bayesian inverse reinforcement learning' (Ramachandran&Amir, 2007)
  - MWAL (Syed&Shapire, 2010)

EXPERIMENTAL RESULTS

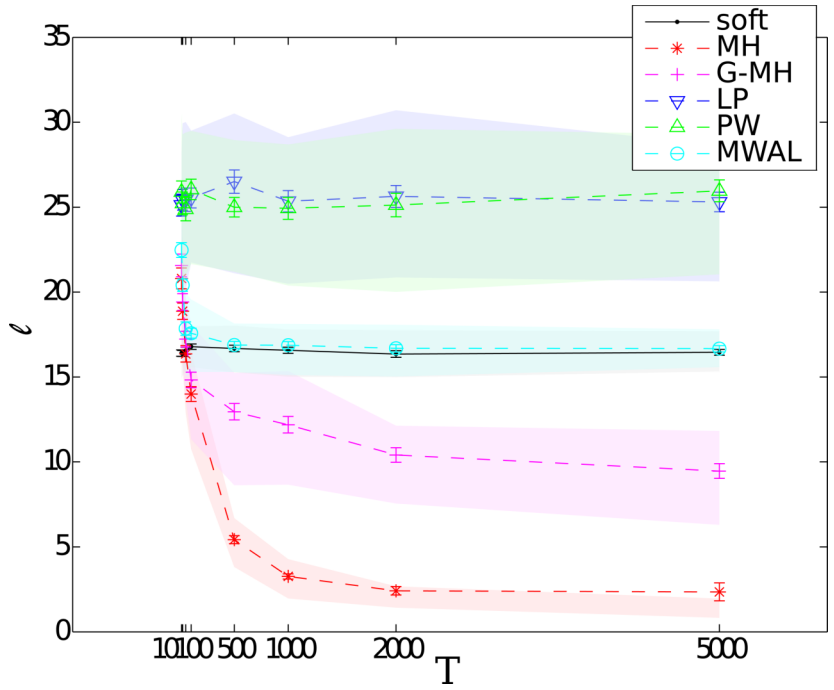


Total loss ℓ with respect to the optimal policy, as a function of the inverse temperature η of the softmax policy of the demonstrator for the Random MDP tasks, averaged over 100 runs. The shaded areas indicate the 80% percentile region, while the error bars the standard error.

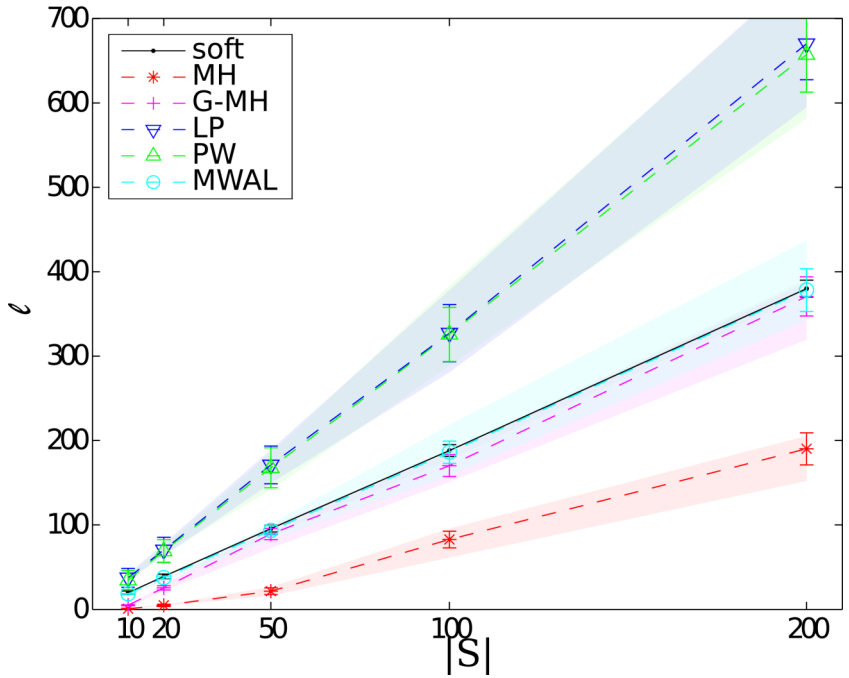


Total loss ℓ with respect to the optimal policy, as a function of the inverse temperature η of the softmax policy of the demonstrator for the Random Maze tasks, averaged over 100 runs. The shaded areas indicate the 80% percentile region, while the error bars the standard error.

EXPERIMENTAL RESULTS



Total loss ℓ with respect to the optimal policy, in the Random MDP task. The figure shows how performance improves as a function of the length T of the demonstrated sequence. All quantities are averaged over 100 runs. The shaded areas indicate the 80% percentile region, while the error bars the standard error.



Total loss ℓ with respect to the optimal policy, in the Random MDP task. The figure shows the effect of the number of states |S| of the underlying MDP. All quantities are averaged over 100 runs. The shaded areas indicate the 80% percentile region, while the error bars the standard error.

CONCLUSION

- **Unified framework of preference elicitation and inverse reinforcement learning**
  - two statistical inference models
  - two corresponding sampling procedures for inference

- **Formulation is general**
  - alternative priors for rewards and policies
  - alternative form of agent's preferences

- In experiments, we showed that for a particular choice of policy prior, closely corresponding to previous approaches, our samplers can outperform not only other well-known inverse reinforcement learning algorithms, but the demonstrating agent as well.

THANKS!

See also our contribution at **EWRL**:
Christos Dimitrakakis, Constantin Rothkopf:
Bayesian multitask inverse reinforcement learning