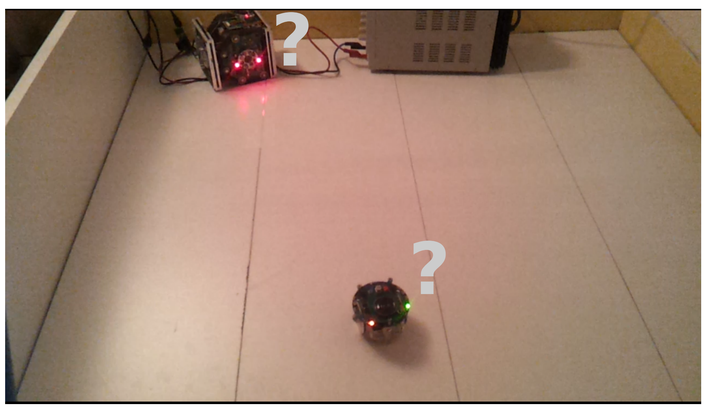
Interactive Robot Education

Riad Akrour, **Marc Schoenauer**, and **Michèle Sebag**
FirstName.LastName@inria.fr

Policy Learning

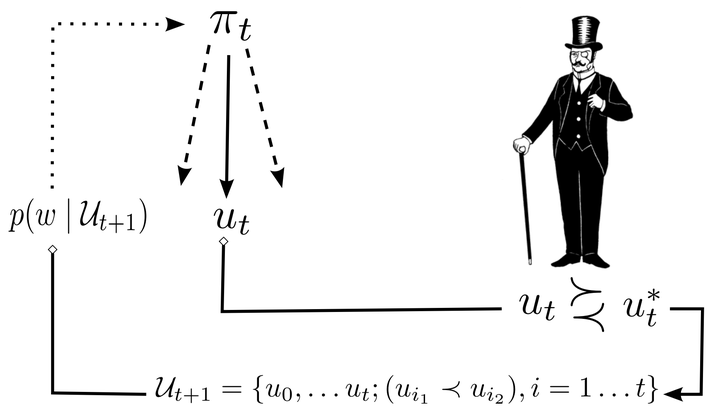


- Learn policy $\pi : \mathcal{S} \to \mathcal{A}$ solving a control problem

Control problem definition

- Three ways to leverage human expertise in the state of the art:
  - Define a reward function
    - Use reinforcement learning to find the policy maximizing it
  - Provide a set of target trajectories
    - Use learning by imitation to find the underlying policy
  - Compare two trajectories proposed by the learner [**?**, **?**]
    - Iterative process
    - **Least demanding alternative (requires a few bits of informations only)**

Interactive Robot Education



- **Iterate**
  - Expert: Express preferences over demonstrated policies
  - Agent: Update constraints archive and posterior over utilities
  - Agent: Find policy maximizing the Active Selection Criterion
  - Agent: Demonstrate selected policy to the expert

Response Model

- Let for a particular trajectory of a policy π and a feature function ψ, the vector $\mathbf{u} = \sum_{h=0}^{H} \gamma^h \psi(s_h)$

- Let W be the utility space, and $\mathbf{w}^*$ denoting the true (hidden) utility of the expert.

- For two trajectories $\mathbf{u}$ and $\mathbf{u}'$, noise scale parameter $\delta \in \mathbb{R}, \delta > 0$ and for $\mathbf{z} = \langle \mathbf{w}^*, (\mathbf{u} - \mathbf{u}') \rangle$ we define:
  - $P(\mathbf{u} \succ \mathbf{u}' \mid \mathbf{w}^*, \delta) = \frac{1}{2\delta}\mathbf{z} + \frac{1}{2}$, if $|\mathbf{z}| < \delta$ and $P(\mathbf{u} \succ \mathbf{u}' \mid \mathbf{w}^*, \delta) = 1$ (resp. 0) if $\mathbf{z} \geq \delta$ (resp. $\mathbf{z} \leq -\delta$)

- Let $\mathcal{U}_t = \{\mathbf{u}_0, \mathbf{u}_1, \dots; (\mathbf{u}_{i_1} \succ \mathbf{u}_{i_2}), i = 1 \dots t\}$ be the archive of demonstrated trajectories. Assuming uniform prior on W and uniform prior over noise parameters δ_i on interval $[0, M]$:
  - $p(\mathbf{w}; \mathcal{U}_t) \propto \prod_{i=1}^{t} \left(\frac{1}{2} + \frac{\mathbf{Z}_i(\mathbf{w})}{2M} \left(1 + \log \frac{M}{|\mathbf{Z}_i(\mathbf{w})|} \right) \right)$

Active Selection Criterion

- Let $\mathbf{u}^*$ expert's most preferred trajectory at time t

- **Question**: Given current archive $\mathcal{U}_t$ which policy to select for the next tentative demonstration? One maximizing *EUS* [**?**]

- $EUS(\{\mathbf{u}, \mathbf{u}^*\}; \mathcal{U}_t) = \int_W p(\mathbf{w} \mid \mathcal{U}_t)\ max(\langle \mathbf{w}, \mathbf{u} \rangle, \langle \mathbf{w}, \mathbf{u}^* \rangle)\ \mathrm{d}\mathbf{w}$

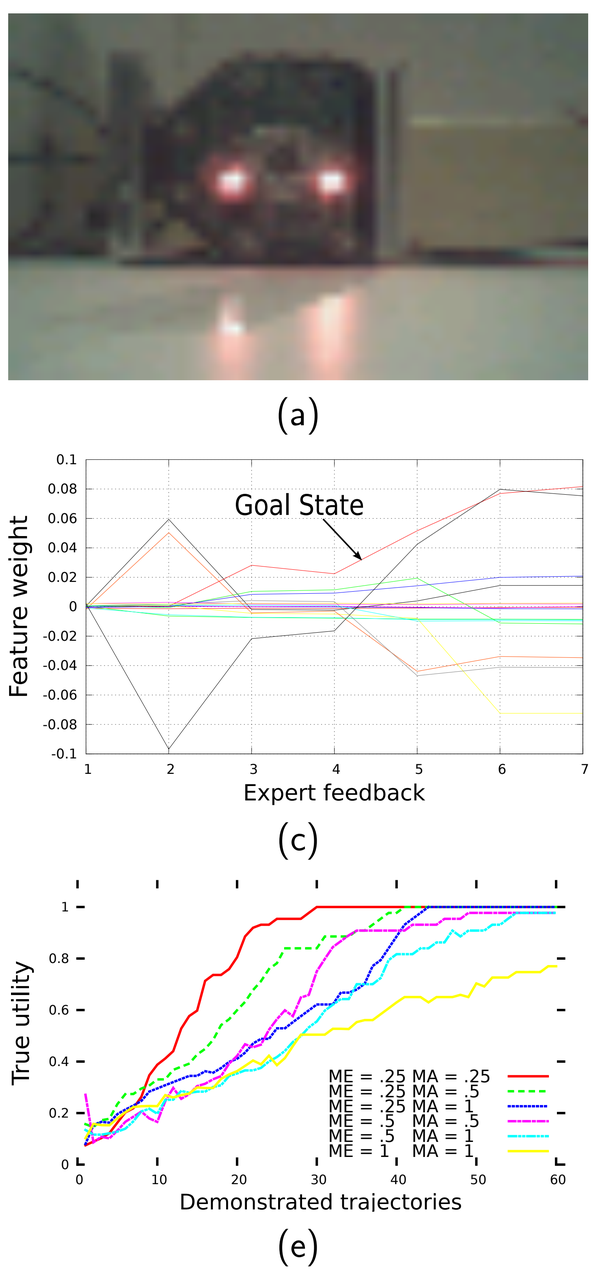
- Estimate the integral using importance sampling from a set n of particles sampled uniformly from W
  - Resample using MCMC when the number of effective particles $1/\sum_{i=1}^{n} \alpha_i^2$ (α_i the normalized importance weight) falls below a threshold

EUS Optimization

- Sample an initial utility $\mathbf{w}_0$ from the posterior $p(\mathbf{w}; \mathcal{U}_t)$
- At iteration k RL is used to find $\pi_k = \arg\max_\pi \mathbb{E}[\langle \mathbf{w}_k, \sum_{i=0}^{i=\infty} \gamma^i \phi(s_i) \mid s_{i+1} \sim p(s_i, \pi(s_i)) \rangle]$
- The average trajectory $\bar{\mathbf{u}}_k$ associated to policy π_k is determined
- A new utility function $\mathbf{w}_{k+1} = \int_W p(\mathbf{w} \mid \mathcal{U}_t) \mathbb{1}_{\{\langle w, \bar{u}_k - u_t \rangle > 0\}} \mathrm{d}\mathbf{w}$ is generated
- k is incremented until the *EUS* stops increasing.
- This process is prone to local optima and needs to be iterated

Experiments

- Reaching a target robot
  - Robot sees target and can estimate its distance using the camera
  - Expert wants the robot to come close enough to the target and stop
  - Expert preferences are transmitted using IR at the end of each demonstration
- Grid world
  - Expert answer emulated with $\mathbf{w}^*$ as in Fig.d
  - Varying in the experiment the two parameters:
    - ME, the noise scale of the emulated expert responses
    - MA, the noise scale of the agent model of expert responses



	⋯	1/4	1/2	1
			1/4	1/2
1/64				1/4
1/128	1/64			⋮
1/256	1/128	1/64		

(a) (b) (c) (d) (e) (f)

Figure: a) Start state from robot's perspective. b) Performance of the policy maximizing the average utility *vs* number of interactions with the expert, averaged out of 5 runs. c) Average utility weight vector in $\mathbb{R}^{16}$ *vs* number of interactions. d) The grid world: the agent initially in the center state must visit the upper rightmost goal state. The numbers represent the true hidden utility; e) Performance of the policy maximizing the average utility *vs* number of interactions with the expert, averaged out of 21 runs. f) Expert error rate up *vs* number of interactions.

Conclusion

- Only a small amount of domain knowledge necessary
- When the transition model is provided, learning time almost exclusively consumed by displaying trajectories to the expert
- There is an intricate relation between expert and agent capabilities
  - An expert making less mistakes when ranking trajectories implies a faster convergence
  - An agent learning faster provides the expert easier to rank trajectories