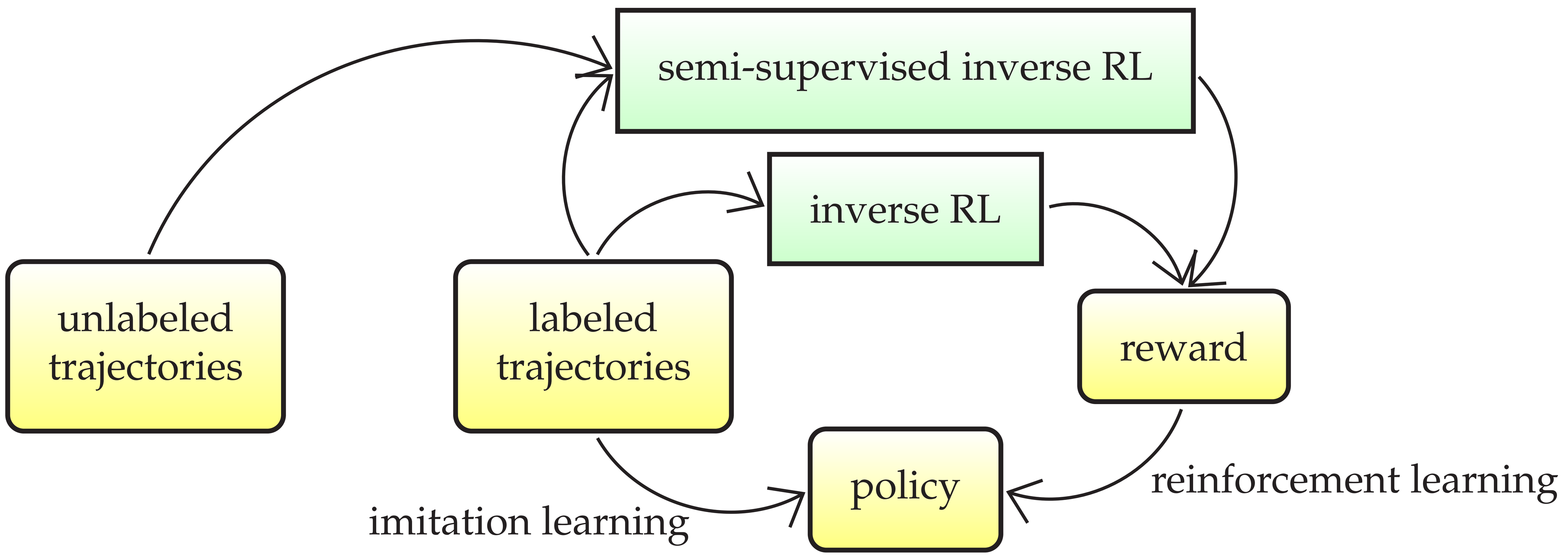
Semi-Supervised Inverse Reinforcement Learning

{ Michal.Valko, Mohammad.Ghavamzadeh and Alessandro.Lazaric }@inria.fr

informatics mathematics Inria

Problem: Apprenticeship Learning



- **Inverse reinforcement learning:** expert trajectories $\mapsto$ policy (via reward)
- **Problem:** expert trajectories are expensive to get or not available
- **Solution:** learn also from unlabeled trajectories and use the structure in the feature counts

Semi-Supervised Inverse Reinforcement Learning

- If we assume that the reward is linear in feature counts, $R^\star(s) = \boldsymbol{w}^\star \cdot \phi(s)$, then:

$$\mathbb{E}_{s_0\sim D}[V^\pi(s_0)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi\right] = \boldsymbol{w}\cdot\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\phi(s_t)|\pi\right] = \boldsymbol{w}\cdot\boldsymbol{\mu}(\pi).$$

- IRL of Abbeel and Ng [1] is based on **matching the feature counts** of the expert performer:

$$\left|\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)|\pi_E\right] - \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)|\tilde{\pi}\right]\right| = \left|\boldsymbol{w}^\intercal\boldsymbol{\mu}(\tilde{\pi}) - \boldsymbol{w}^\intercal\boldsymbol{\mu}_E\right| \le \|\boldsymbol{w}\|_2\,\|\boldsymbol{\mu}(\tilde{\pi}) - \boldsymbol{\mu}_E\|_2 \le \varepsilon$$

- **Semi-supervised learning** (SSL) makes distributional assumptions such compactness (gap, null-category) or smoothness (manifold). We choose to use the **gap assumption** and the related semi-supervised support vector machines (SVMs).
- Semi-supervised SVMs use besides the standard *hinge loss* $V(f,\boldsymbol{x}_i,y_i) = \max\{1-y\,|f(\boldsymbol{x})|\,,0\}$, also the *hat loss* $\widehat{V}(f,\boldsymbol{x}) = \max\{1-|f(\boldsymbol{x})|\,,0\}$ on unlabeled data [2] to compute max-margin decision boundary $\widehat{f}$ that **avoids dense regions** of data:

$$\widehat{f} = \min_f \sum_{i\in L} V(f,\boldsymbol{x}_i,y_i) + \gamma_l\,\|f\|^2 + \gamma_u \sum_{i\in U}\widehat{V}(f,\boldsymbol{x}_i),$$

- In semi-supervised IRL (**SSIRL**) we penalize the decision boundary $\boldsymbol{w}$ that crosses the empirical feature counts from unlabeled trajectories:

$$\min_{\boldsymbol{w}}\left(\max\left\{1-\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}_E,0\right\} + \gamma_l\,\|\boldsymbol{w}\|_2 + \sum_{j<i}\max\left\{1+\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}^{(j)},0\right\} + \gamma_u\sum_{u\in U}\max\left\{1-|\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}_u|\,,0\right\}\right)$$

Discussion

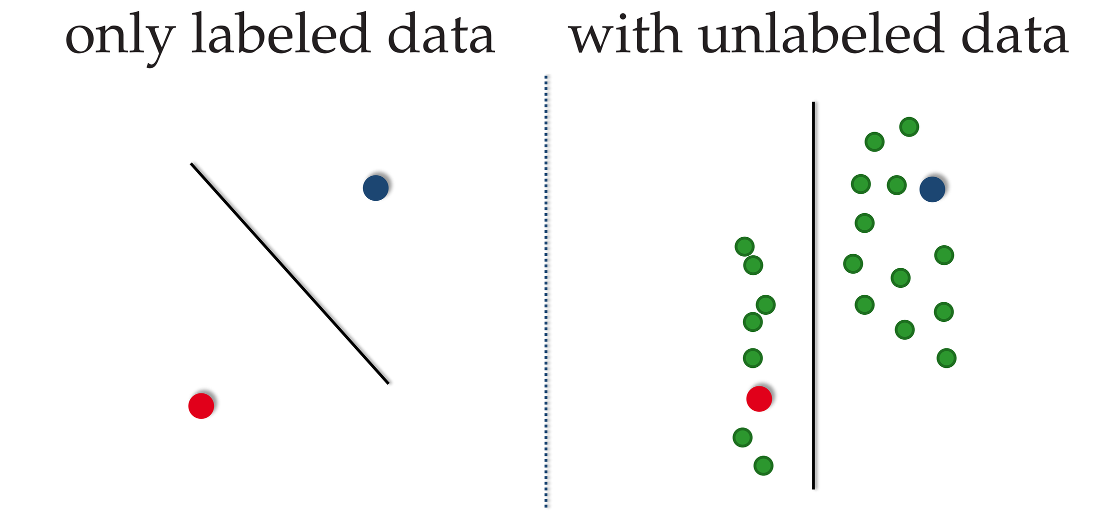
- **Contributions:**
  - **first** IRL method to take advantage of the **unlabeled** trajectories
  - assuming **clustered** feature counts can learn a better performing policy
- **Disadvantages:**
  - similar to [1] only outputs a **mixture** policy
  - stopping criterion is needed, because the method **converges to IRL** [1]
- **Future directions:**
  - enhance other inverse RL methods (MaxEnt IRL, MMP) with unlabeled trajectories
  - investigate manifold assumption for inverse RL

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 1—, New York, NY, USA, 2004. ACM.

[2] Kristin Bennett and Ayhan Demiriz. Semi-Supervised Support Vector Machines. In *Advances in Neural Information Processing Systems 11*, pages 368–374, 1999.

SSL: Cluster Assumption



SSIRL Algorithm

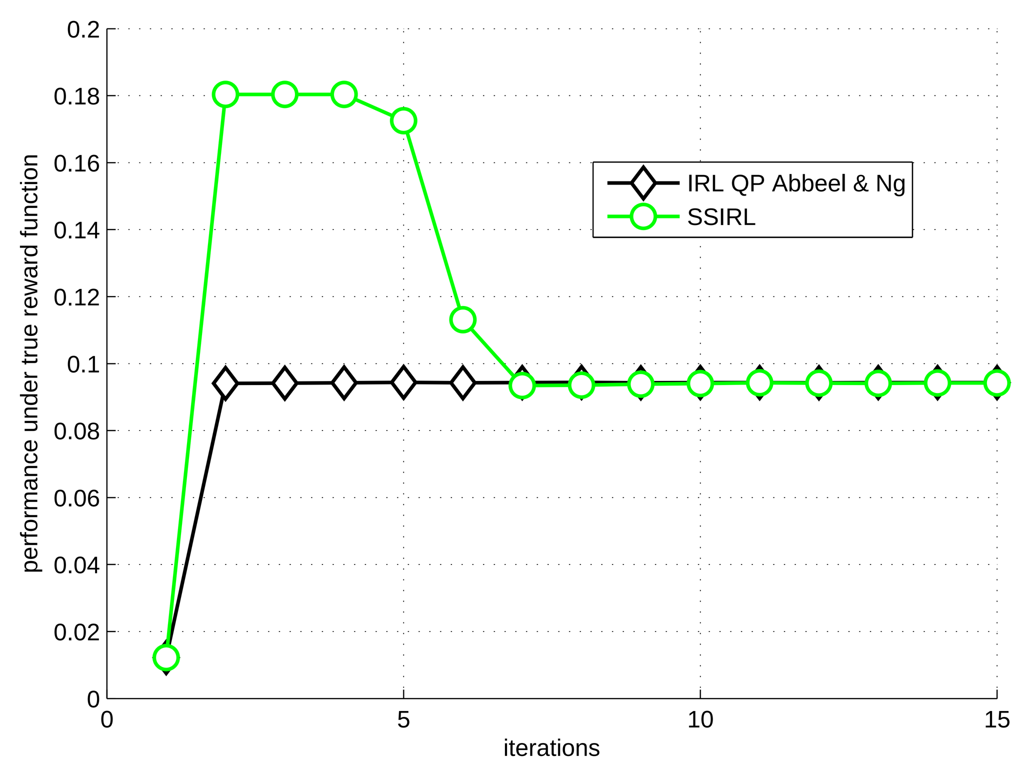
Input: ε, γ_l, γ_u
expert trajectories $\{s_{E,t}^{(i)}\}$
unlabeled trajectories
from U performers $\{s_{u,t}^{(i)}\}$
estimate $\widehat{\boldsymbol{\mu}}_E \leftarrow \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{\infty}\gamma_l^t\phi(s_{E,t}^{(i)})$
for $u = 1$ **to** U **do**
estimate $\widehat{\boldsymbol{\mu}}_u \leftarrow \frac{1}{m_u}\sum_{i=1}^{m_u}\sum_{t=0}^{\infty}\gamma^t\phi(s_{u,t}^{(i)})$
end for
randomly pick $\pi^{(0)}$ and set $i \leftarrow 1$
repeat

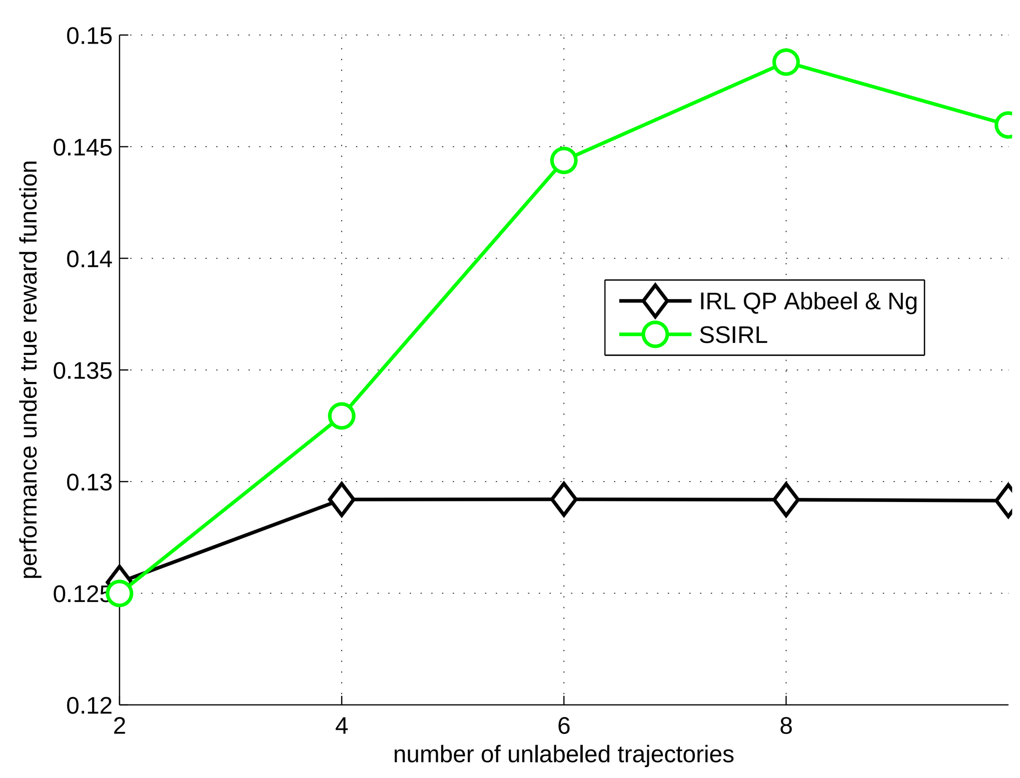
$$\boldsymbol{w}^{(i)} \leftarrow \min_{\boldsymbol{w}}\Big(\max\{1-\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}_E,0\} + \gamma_l\,\|\boldsymbol{w}\|_2 + \sum_{j<i}\max\{1+\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}^{(j)},0\} + \gamma_u\sum_{u\in U}\max\{1-|\boldsymbol{w}^\intercal\widehat{\boldsymbol{\mu}}_u|\,,0\}\Big)$$

$\boldsymbol{w}^{(i)} \leftarrow \boldsymbol{w}^{(i)}/\left\|\boldsymbol{w}^{(i)}\right\|_2$
$\pi^{(i)} \leftarrow \mathrm{MDP}(R = (\boldsymbol{w}^{(i)})^\intercal\boldsymbol{\phi})$
estimate $\widehat{\boldsymbol{\mu}}^{(i)} \leftarrow \boldsymbol{\mu}(\pi^{(i)})$
$t^{(i)} \leftarrow \min_i \boldsymbol{w}^\intercal(\widehat{\boldsymbol{\mu}}_E - \widehat{\boldsymbol{\mu}}^{(i)})$
$i \leftarrow i+1$
until $t^{(i)} \le \varepsilon$

Results: SSIRL vs. IRL

Gridworlds





Performance of the final mixture policies under the true reward (unknown to both algorithms).