模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Active Exploration for Inverse Reinforcement Learning

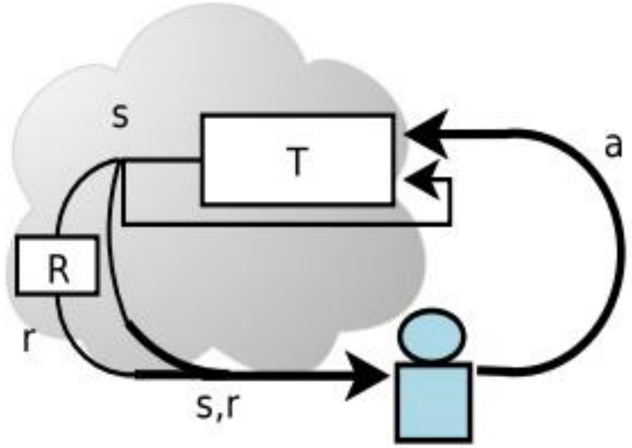
David Lindner
Department of Computer Science
ETH Zurich
david.lindner@inf.ethz.ch

Andreas Krause
Department of Computer Science
ETH Zurich
krausea@ethz.ch

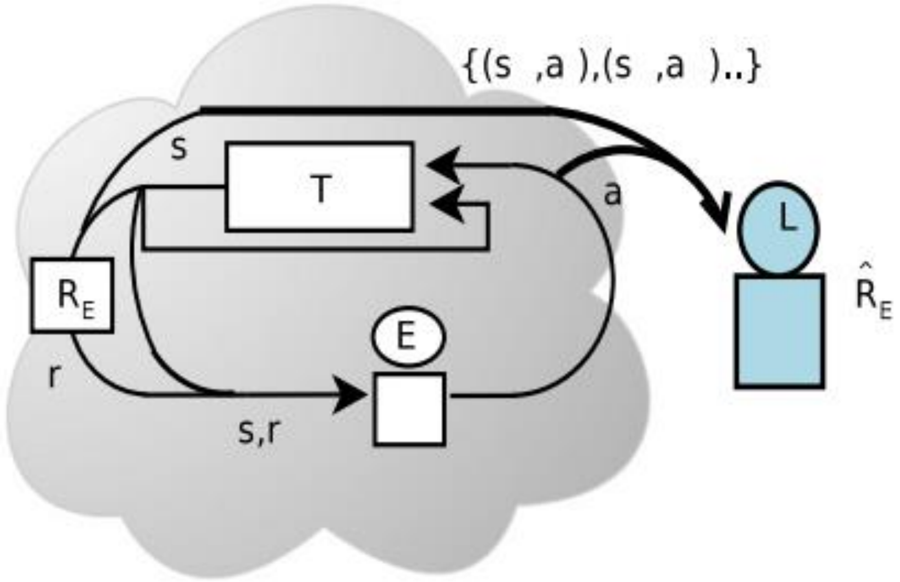
Giorgia Ramponi
ETH AI Center
giorgia.ramponi@ai.ethz.ch

Preliminaries

IRL recovers the feasible reward function in which the optimal policy is the expert policy.



RL paradigm



IRL paradigm

Preliminaries

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

MDP: [MDP] An MDP $\mathcal{M} := \langle S, A, T, R, \gamma \rangle$

Transition: $T : S \times A \to \text{Prob}(S)$

Reward: $R : S \times A \to \mathbb{R}$

Value: $$V^{\pi}(s_0) = E_{s,\pi(s)}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) | s_0\right]$$

MDP goal: $V^{\pi^*}(s) = V^*(s) = \sup_{\pi} V^{\pi}(s)$

Visit frequency: $$\psi^{\pi}(s) = \psi^0(s) + \gamma \sum_{s' \in S} T(s, \pi(s), s') \; \psi^{\pi}(s')$$

Value: $$V^*(s) = \sup_{\pi} \sum_{s \in S} \psi_*^{\pi}(s) \; R(s, \pi(s))$$

Preliminaries

MDP without reward: $\mathcal{M} \backslash_R$

Feature reward: $$R(s,a) = w_1\phi_1(s,a) + w_2\phi_2(s,a) + \ldots + w_k\phi_k(s,a)$$
$$= \boldsymbol{w}^T\boldsymbol{\phi}(s,a).$$

Feature reward count: $$\mu^{\phi_k}(\pi) = \sum_{t=0}^{\infty} \psi^{\pi}(s_t)\ \phi_k(s_t, \pi(s_t))$$

Empirical feature count: $$\hat{\mu}^{\phi_k}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\gamma^t\phi_k(s_t, a_t).$$

Value: $$V^{\pi} = \boldsymbol{w}^T\boldsymbol{\mu}^{\phi}(\pi) = \sum_{s,a}\psi^{\pi}(s)\ \boldsymbol{w}^T\boldsymbol{\phi}(s,a)$$
$$= \sum_{s,a}\psi^{\pi}(s)\ R(s,a).$$

Current Methods

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

Margin Optimization：

$$\sum_{s \in S} Q^{\pi}(s, a^*) - \max_{a \in A \setminus \{a^*\}} Q^{\pi}(s, a)$$

$$|\boldsymbol{\mu}^{\phi}(\pi) - \hat{\boldsymbol{\mu}}^{\phi}(\mathcal{D})|.$$

$$\hat{\pi}_E(a|s) - \pi_E(a|s)$$

Entropy Optimization：

$$\max_{\Delta} \; - \sum_{\tau \in (S \times A)^l} Pr(\tau) \log Pr(\tau)$$

$$\min_{P \in \Delta} \sum_{\tau \in (S \times A)^l} P(\tau) \log \frac{P(\tau)}{Q(\tau)}.$$

Bayesian Update：

$$P(\langle s, a \rangle | \hat{R}_E) \propto e^{\left(\frac{Q^*(s,a;\hat{R}_E)}{\beta}\right)}$$

Current Methods

Method	$\hat{R}_E$ params	Optimization objective	Notable aspect
Max margin methods - maximize the margin between value of observed behavior and the hypothesis			
MMP	w	value of obs. τ - max of values from all other τ (Eq. 8)	provable convergence
MAX-MARGIN		feature exp. of policy - empirical feature exp. (Eq. 9)	sample bounds
MWAL		min diff. in value of policy and observed τ across features	first bound on iteration complexity
HYBRID-IRL		empirical stochastic policy - computed policy of expert (Eq. 10)	natural gradients and efficient optimization
LEARCH	$R(\phi)$	value of obs. τ - max of values from all other τ (Eq. 8)	nonlinear reward with suboptimal input
Silver et al. [26]			normalization of outlier inputs
Max entropy methods - maximize the entropy of the distribution over behaviors			
MAXENTIRL	w	entropy of distribution over trajectories (Eq. 11)	low learning bias
STRUCTURED APPRENTICESHIP		entropy of distribution over policies (Eq. 12)	efficient optimization
DEEP MAXENTIRL		gradient of likelihood equivalent of MaxEnt (Eq. 13)	nonlinear reward
PI-IRL			continuous state-action spaces
REIRL		relative entropy of distribution from baseline policy (Eq. 14)	suboptimal input and unknown dynamics
Bayesian learning methods - learn posterior over hypothesis space using Bayes rule			
BIRL	$R(s)$	posterior with Boltzmann data likelihood (Eq. 16)	first Bayesian IRL formulation
Lopes et al. [38]		entropy of multinomial$(p_1(s), p_2(s), \ldots, p_{\lvert A\rvert-1}(s))$ derived from posterior	active learning
GP-IRL	$f(\boldsymbol{r}, \boldsymbol{\theta})$	Gaussian process posterior	nonlinear reward
MLIRL	w	differentiable likelihood with Boltzmann policy (Eq. 16)	first ML approach
Classification and regression - learn a prediction model that imitates observed behavior			
SCIRL	w	Q-function as classifier scoring function	actions as state labels provable convergence
CSI			unknown dynamics
FIRL	regression tree	norm of ($\hat{R}_E$ - projection of $\hat{R}_E$)	avoids manual feature engineering

Definitions

MDP\R $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, H, s_0)$

R(s, a, h) $\mathcal{S} \times \mathcal{A} \times [H] \to [0, R_{\max}]$

Q(s, a) $Q^{\pi,h}_{\mathcal{M}\cup r}(s,a) = r_h(s,a) + \sum_{s',a'} \pi_{h+1}(a'|s') P(s'|s,a) Q^{\pi,h+1}_{\mathcal{M}\cup r}(s',a')$

V(s) $V^{\pi,h}_{\mathcal{M}\cup r}(s) = \sum_a \pi_h(a|s) Q^{\pi,h}_{\mathcal{M}\cup r}(s,a)$

A(s, a) $A^{\pi,h}_{\mathcal{M}\cup r}(s,a) = Q^{\pi,h}_{\mathcal{M}\cup r}(s,a) - V^{\pi,h}_{\mathcal{M}\cup r}(s)$

Feasible policy $\Pi^*_{\mathcal{M}\cup r}$

Frequency $\eta^{h,h}_{\mathcal{M},\pi}(s'|s) := \mathbb{1}_{\{s'=s\}}$ and $\eta^{h,h'+1}_{\mathcal{M},\pi}(s'|s) := \sum_{s'',\tilde{a}} P(s'|s'',\tilde{a}) \pi_{h'}(\tilde{a}|s'') \eta^{h,h'}_{\mathcal{M},\pi}(s''|s)$.

Definition: We can define a policy is optimal iff the A≤0 for any states and actions.

Definitions

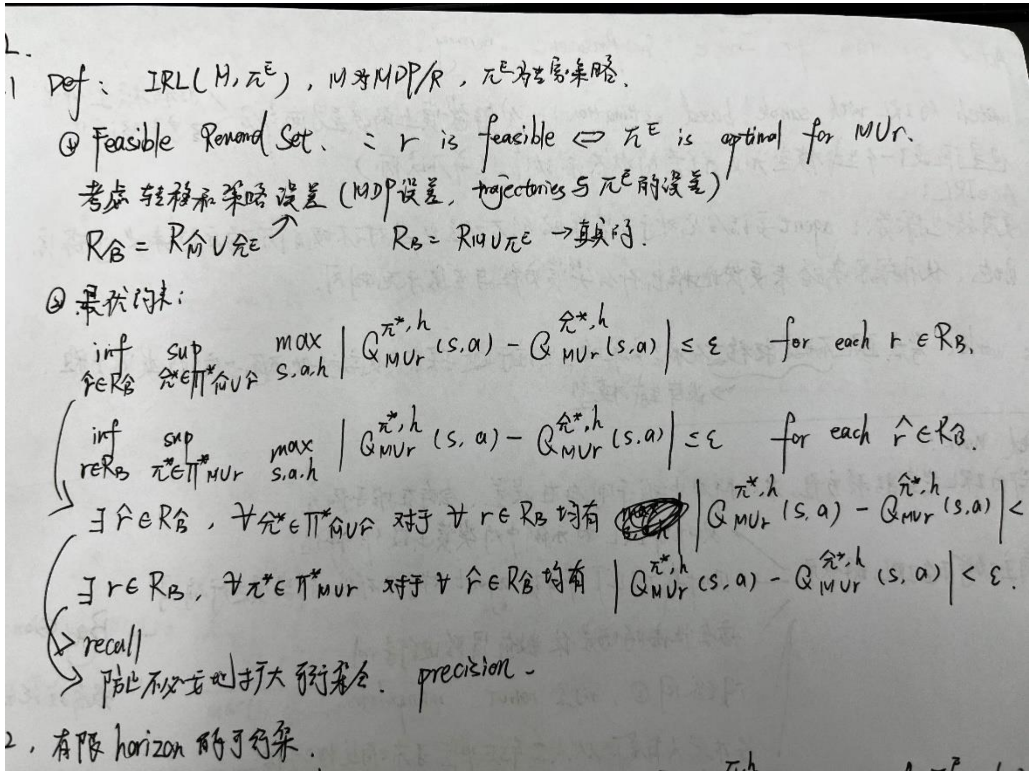
Definition 1 (Feasible Reward Set). *A reward function* r *is* feasible *for an IRL problem* $(\mathcal{M}, \pi^E)$, *if and only if the expert policy* π^E *is optimal in* $\mathcal{M} \cup r$. *We call the set of all feasible reward functions* $\mathcal{R}_{\mathcal{M}\cup\pi^E}$ *the* feasible reward set. *If we estimate the transition model and expert policy from samples, we refer to the* recovered feasible set $\mathcal{R}_{\widehat{\mathfrak{B}}} = \mathcal{R}_{\widehat{\mathcal{M}}\cup\hat{\pi}^E}$ *in contrast to the* exact feasible set $\mathcal{R}_{\mathfrak{B}} = \mathcal{R}_{\mathcal{M}\cup\pi^E}$.

Definition 2 (Optimality Criterion). *Let* $\mathcal{R}_{\mathfrak{B}}$ *be the exact feasible set and* $\mathcal{R}_{\widehat{\mathfrak{B}}}$ *be the feasible set recovered after observing* $n \geq 0$ *samples collected from* $\mathcal{M}$ *and* π^E. *We say that an algorithm for Active IRL is* (ϵ, δ, n)-correct *if after* n *iterations with probability at least* $1 - \delta$ *it holds that:*

$$\inf_{\hat{r}\in\mathcal{R}_{\widehat{\mathfrak{B}}}} \sup_{\hat{\pi}^*\in\Pi^*_{\widehat{\mathcal{M}}\cup\hat{r}}} \max_{s,a,h} \left| Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) \right| \leq \epsilon \quad \text{for each } r \in \mathcal{R}_{\mathfrak{B}},$$

$$\inf_{r\in\mathcal{R}_{\mathfrak{B}}} \sup_{\pi^*\in\Pi^*_{\mathcal{M}\cup r}} \max_{s,a,h} \left| Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) \right| \leq \epsilon \quad \text{for each } \hat{r} \in \mathcal{R}_{\widehat{\mathfrak{B}}},$$

where π^* *is an optimal policy in* $\mathcal{M} \cup r$ *and* $\hat{\pi}^*$ *is an optimal policy in* $\widehat{\mathcal{M}} \cup \hat{r}$.



1. Def: IRL(M, π^E), M为MDP\R, π^E为专家策略.

① Feasible Reward Set: r is feasible $\Leftrightarrow$ π^E is optimal for $M \cup r$.

考虑转移和策略误差 (MDP误差, trajectories与π^E的误差)

$R_{\hat{B}} = R_{\hat{M}\cup\hat{\pi}^E}$ $\quad R_B = R_{M\cup\pi^E}$ → 真实的.

② 最优约束:

$$\inf_{\hat{r}\in R_{\hat{B}}} \sup_{\hat{\pi}^*\in\Pi^*_{\hat{M}\cup\hat{r}}} \max_{s,a,h} \left| Q^{\pi^*,h}_{M\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{M\cup r}(s,a) \right| \leq \varepsilon \quad \text{for each } r \in R_B.$$

$$\inf_{r\in R_B} \sup_{\pi^*\in\Pi^*_{M\cup r}} \max_{s,a,h} \left| Q^{\pi^*,h}_{M\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{M\cup r}(s,a) \right| \leq \varepsilon \quad \text{for each } \hat{r} \in R_{\hat{B}}.$$

$\exists\, \hat{r}\in R_{\hat{B}}$, $\forall\, \hat{\pi}^*\in\Pi^*_{\hat{M}\cup\hat{r}}$ 对于 $\forall\, r\in R_B$ 均有 $\left| Q^{\pi^*,h}_{M\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{M\cup r}(s,a) \right| <$

$\exists\, r\in R_B$, $\forall\, \pi^*\in\Pi^*_{M\cup r}$ 对于 $\forall\, \hat{r}\in R_{\hat{B}}$ 均有 $\left| Q^{\pi^*,h}_{M\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{M\cup r}(s,a) \right| < \varepsilon$.

recall

防止不必要地扩大预测集. precision.

2. 有限 horizon 的可行集.

Recall:存在一个r_hat产生的策略能和RB里均有相近的表现，说明查全率高。
Precision：存在一个RB里的r能够和RB_hat中的所有r_hat都有相近的表现，存在限定了范围

Lemma

Lemma 3 (Feasible Reward Set Implicit). *A reward function r is feasible if and only if for all s, a, h it holds that: $A^{\pi,h}_{\mathcal{M}\cup r}(s,a) = 0$ if $\pi^E_h(a|s) \geq 0$ and $A^{\pi,h}_{\mathcal{M}\cup r}(s,a) \leq 0$ if $\pi^E_h(a|s) = 0$. Moreover, if the second inequality is strict, π^E is uniquely optimal, i.e., $\Pi^*_{\mathcal{M}\cup r} = \{\pi^E\}$.*

Lemma 4 (Feasible Reward Set Explicit). *A reward function r is feasible if and only if there exists an $\{A_h \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}_{\geq 0}\}_{h\in[H]}$ and $\{V_h \in \mathbb{R}^{\mathcal{S}}\}_{h\in[H]}$ such that for all s, a, h it holds that:*

$$r_h(s,a) = -A_h(s,a)\mathbb{1}_{\{\pi^E_h(a|s)=0\}} + V_h(s) + \sum_{s'} P(s'|s,a)V_{h+1}(s')$$

4.2，有限 horizon 的可行集.

Implicit: r is feasible $\Leftrightarrow$ $A^{\pi,h}_{MUr}(s,a)=0$ if $\pi^E_h(a|s)\geqslant 0$ 且 $A^{\pi,h}_{MUr}(s,a)\leq 0$ if $\pi^E_h(a|s)\geqslant 0$.

$$A^{\pi,h}_{MUr}(s,a) = r_h(s,a) + \sum_{s',a'} \pi_{h+1}(a'|s')P(s'|s,a)Q^{\pi,h+1}_{MUr}(s',a') - \sum_a \pi_h(a|s)Q^{\pi,h}_{MUr}(s,a)$$

$$= r_h(s,a) + \sum_{s',a'} P(s'|s,a)V^{\pi,h+1}_{MUr}(s') - V^{\pi,h}_{MUr}(s) \rightarrow \text{Sarsa?}$$

Explicit: r is feasible $\Leftrightarrow$ $\exists\{A_h\in R^{S\times A}_{\geqslant 0}\}_{h\in[H]}$ 且 $\{V_h\in R^S\}_{h\in[H]}$ 对 $\forall s,a,h$ 存在.

$$r_h(s,a) = \underbrace{-A_h(s,a)\mathbb{1}\{\pi^E_h(a|s)=0\}}_{} + V_h(s) + \sum_{s'}P(s'|s,a)V_{h+1}(s')$$

不太明白.

~~$A_h(s,a)<0$ 且 $\pi^E_h(a|s)\neq 0$~~

~~已选择了但优势较小~~

对专家选择的为0，专家没选择的为 $A_h(s,a)$

Error Propagation in generative model

Theorem 5 (Error Propagation). *Let* $(\mathcal{M}, \pi^E)$ *and* $(\widehat{\mathcal{M}}, \widehat{\pi}^E)$ *be two IRL problems. Then, for any* $r \in \mathcal{R}_{(\mathcal{M},\pi^E)}$ *there exists* $\widehat{r} \in \widehat{\mathcal{R}}_{(\widehat{\mathcal{M}},\widehat{\pi}^E)}$ *such that:*

$$|r_h(s,a) - \widehat{r}_h(s,a)| \leq \underbrace{\boxed{A_h(s,a)|\pi_h^E(a|s) - \widehat{\pi}_h^E(a|s)|}}_{\text{专家策略误差}} + \underbrace{\boxed{\sum_{s'} V_{h+1}(s')|P(s'|s,a) - \widehat{P}(s'|s,a)|}}_{\text{转移模型误差}}$$

and we can bound $V_h \leq (H-h)R_{\max}$ *and* $A_h \leq (H-h)R_{\max}$.

Lemma 6. *Let* $\mathcal{M}$ *be an MDP**R*, $r, \hat{r}$ *two reward functions with optimal policies* $\pi^*, \hat{\pi}^*$. *Then,*

$$Q^{\pi^*,h}_{\mathcal{M}\cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M}\cup r}(s,a) \leq \sum_{h'=h}^{H} \sum_{s',a'} \left(\eta^{h,h'}_{\mathcal{M},\pi^*}(s',a'|s,a) - \eta^{h,h'}_{\mathcal{M},\hat{\pi}^*}(s',a'|s,a)\right) \underbrace{\boxed{(r_{h'}(s',a') - \hat{r}_{h'}(s',a'))}}_{\text{可用Theorem5代入获得专家策略和转移模型各自的误差}}$$

Ace-IRL with generative model

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

$$\widehat{P}_k(s'|s,a) = \frac{\sum_{h=1}^{H} n_k^h(s,a,s')}{\max(1, \sum_{h=1}^{H} n_k^h(s,a))} \qquad \hat{\pi}_{k,h}^{E}(a|s) = \frac{n_k^h(s,a)}{\max(1, n_k^h(s))}.$$

In Appendix B.3 we derive Hoeffding's confidence intervals for the transition model and the expert policy. Combining these with Theorem 5, we can compute the uncertainty on the recovered reward as:

$$C_k^h(s,a) = (H-h)R_{\max} \min\left(1, 2\sqrt{\frac{2\ell_k^h(s,a)}{n_k^h(s,a)}}\right),$$

where $\ell_k^h(s,a) = \log\left(24SAH(n_k^h(s,a))^2/\delta\right)$. We can show that for any pair of reward functions $r \in \mathcal{R}_{\mathfrak{B}}$ and $\hat{r} \in \mathcal{R}_{\hat{\mathfrak{B}}}$, the difference $|r_h(s,a) - \hat{r}_{k,h}(s,a)| \leq C_k^h(s,a)$. This uncertainty estimate will be a key component in all of our theoretical analysis.

Theorem 7 (Sample Complexity of Uniform Sampling IRL). *The uniform sampling strategy fulfills Definition 2 with a number of samples upper bounded by:*

$$n \leq \tilde{\mathcal{O}}\left(H^5 R_{\max}^2 SA/\epsilon^2\right),$$

where $\tilde{\mathcal{O}}$ suppresses logarithmic terms.

Ace-IRL without generative model

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

Definition 2 (Optimality Criterion). *Let $\mathcal{R}_{\mathfrak{B}}$ be the exact feasible set and $\mathcal{R}_{\hat{\mathfrak{B}}}$ be the feasible set recovered after observing $n \geq 0$ samples collected from $\mathcal{M}$ and π^E. We say that an algorithm for Active IRL is* (ϵ, δ, n)-correct *if after n iterations with probability at least $1-\delta$ it holds that:*

$$\inf_{\hat{r} \in \mathcal{R}_{\hat{\mathfrak{B}}}} \sup_{\hat{\pi}^* \in \Pi^*_{\widehat{\mathcal{M}} \cup \hat{r}}} \max_{s,a,h} \left| Q^{\pi^*,h}_{\mathcal{M} \cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M} \cup r}(s,a) \right| \leq \epsilon \quad \text{for each } r \in \mathcal{R}_{\mathfrak{B}},$$

$$\inf_{r \in \mathcal{R}_{\mathfrak{B}}} \sup_{\pi^* \in \Pi^*_{\mathcal{M} \cup r}} \max_{s,a,h} \left| Q^{\pi^*,h}_{\mathcal{M} \cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M} \cup r}(s,a) \right| \leq \epsilon \quad \text{for each } \hat{r} \in \mathcal{R}_{\hat{\mathfrak{B}}},$$

where π^* is an optimal policy in $\mathcal{M} \cup r$ and $\hat{\pi}^*$ is an optimal policy in $\widehat{\mathcal{M}} \cup \hat{r}$.

$$\longrightarrow \quad \hat{e}^h_k(s,a;\pi^*,\hat{\pi}^*) = \left| Q^{\pi^*,h}_{\mathcal{M} \cup r}(s,a) - Q^{\hat{\pi}^*,h}_{\mathcal{M} \cup r}(s,a) \right|.$$

Error Setting

$$E^h_k(s,a) = \min\Big((H-h)R_{\max}, C^h_k(s,a) + \sum_{s'} \widehat{P}(s'|s,a) \max_{a' \in \mathcal{A}} E^{h+1}_k(s',a')\Big). \qquad \text{(EB1)}$$

$$4 \max_a E^0_k(s_0,a) \leq \epsilon.$$

②AcEIRL, without generative model. — 需要通过探险策略来获取关于环境的理想信息.

定义由 $\hat{r} \in R_{\hat{B}}$ 产生的最优策略 $\hat{\pi}^*$ 和由 $r \in R_B$ 产生的最优策略 π^* 的误差上界为

$$\hat{e}^h_k(s,a;\pi^*,\hat{\pi}^*) = \left| Q^{\pi^*,h}_{MUr}(s,a) - Q^{\hat{\pi}^*,h}_{MUr}(s,a) \right|.$$

同时，由 $|r_h(s,a) - \hat{r}_{k,h}(s,a)| \leq C^h_k(s,a)$ 可以推出：

$$E^h_k(s,a) = \min\Big((H-h)R_{max}, C^h_k(s,a) + \sum_{s'} \hat{P}(s'|s,a) \max_{a' \in A} E^{h+1}_k(s',a')\Big).$$

h之后的误差上界

显而易见，$\hat{e}^h_k(s,a;\pi^*,\hat{\pi}^*) \leq E^h_k(s,a)$. 从而令 $r = C^h_k$ 即可. → AceIRL Greedy

探索策略，在 $4\max_a E^0_k(s_0,a) \leq \varepsilon$ 时停止.

采样复杂度为 $\tilde{O}(H^5 R^2_{max} SA/\varepsilon^2)$

这可以理解为我在这个状态上学得不好，C^k_h 就大，所以我要去多学这个状态的 r.

在这里已经可以单纯的将误差上界作为奖赏去进行探索了，可以理解为误差上界越高代表当前的状态信息越少，越需要更多的信息多reward进行更好的近似。

这种探索策略称为AceIRL-Greedy。

Ace-IRL without generative model

Drawbacks of the AceIRL-greedy:

1. AceIRL-greedy doesn't reduce the future uncertainty.

AceIRL缺陷：①探索了高不确定性状态，但我们的目标在于在下一次迭代减少不确定性。

② AceIRL探索为减少所有策略的不确定性，但目标是减少合理的~~策略~~最优策略的不确定度。

所以作者提出我们需要一个在探索过程中减少不确定性的探索策略，即我们需要一个能够最小化 E_{k+1}^h 的探索策略。但 E_{k+1}^h 不好算，可以通过对当前转移模型的估计。

①

$$\hat{C}_{k+1}^h(s,a) = (H-h) R_{max} \min\left(1, 2\sqrt{\frac{2\ell_k^h(s,a)}{n_k^h(s,a)+\hat{n}_{\pi}^h(s,a)}}\right).$$

$$\hat{n}_{\pi}^h(s,a) = N_E \cdot \eta_{M,\pi}^{o,h}(s,a|s_0)$$

→ π在时间h访问(s, a)的期望数。

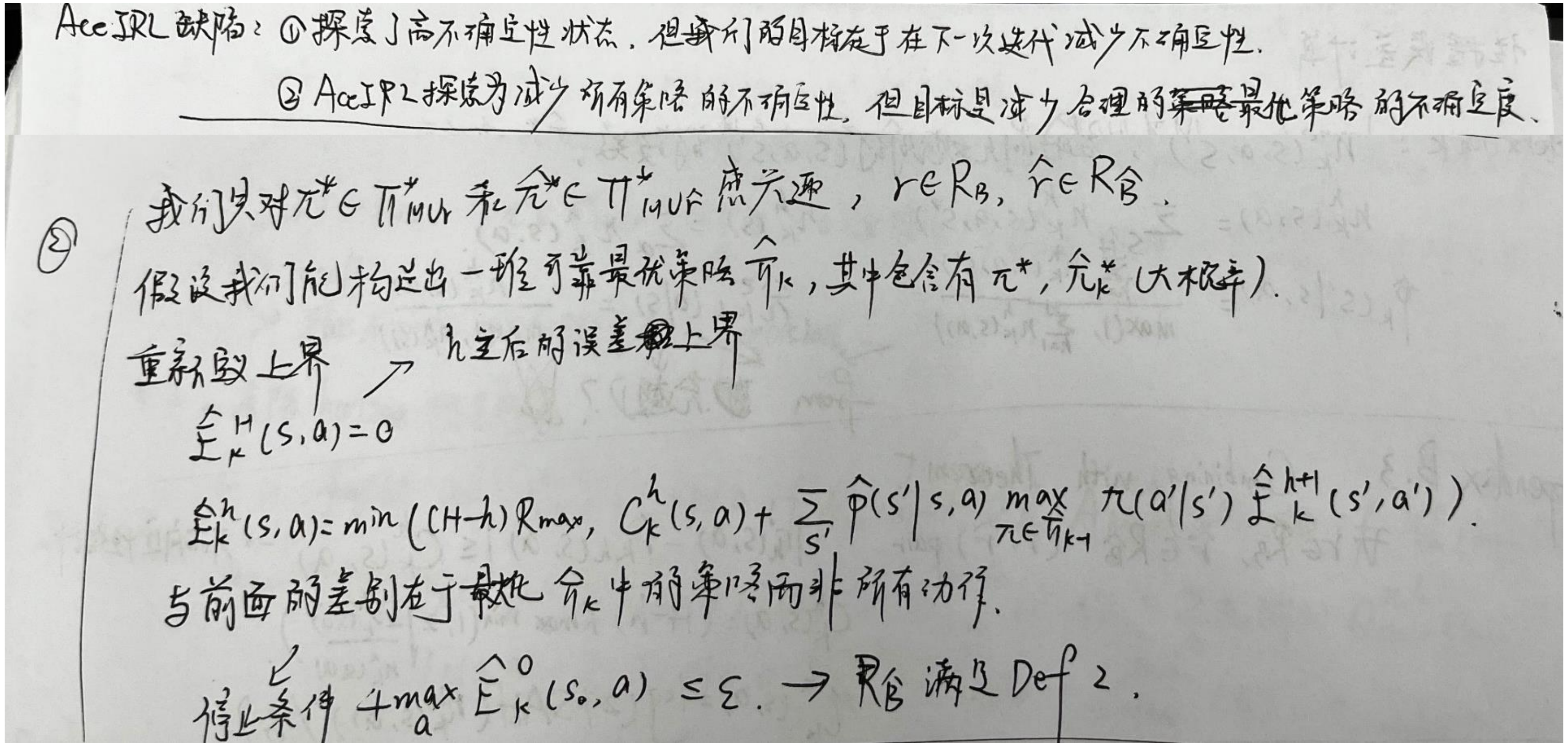
→ N_E 为使用π探索的次数

尽管我们目标是不确定性采样，但现在这个能更好的衡量选择一个探索策略的信息量。这个问题不基于IRL，并可以用于提升reward-free探索。

Ace-IRL without generative model

Drawbacks of the AceIRL-greedy：

2. AceIRL-greedy reduce all policies’ uncertainty.


AceIRL缺陷：①探索了高不确定性状态，但我们的目标在于在下一次迭代减少不确定性.

② AceIRL探索为减少所有策略的不确定性，但目标是减少合理的~~策略~~最优策略的不确定度.

② 我们只对 $\pi^* \in \Pi^*_{M\cup r}$ 和 $\hat{\pi}^* \in \Pi^*_{\hat{M}\cup \hat{r}}$ 感兴趣，$r \in R_B$，$\hat{r} \in R_{\hat{B}}$.

假设我们能构造出一组可靠最优策略 $\hat{\Pi}_k$，其中包含有 π^*，$\hat{\pi}_k^*$（大概率）.

重新设上界 ↗ h步后的误差上界

$$\hat{E}_k^H(s,a)=0$$

$$\hat{E}_k^h(s,a)=\min\left((H-h)R_{max},\ C_k^h(s,a)+\sum_{s'}\hat{P}(s'|s,a)\max_{\pi\in\hat{\Pi}_{k-1}}\pi(a'|s')\hat{E}_k^{h+1}(s',a')\right).$$

与前面的差别在于~~最大化~~ $\hat{\Pi}_k$ 中的策略而非所有动作.

停止条件 $4\max_a \hat{E}_k^0(s_0,a)\le \varepsilon$. → $R_{\hat{B}}$ 满足 Def 2.

Ace-IRL Sample Complexity

Theorem 8. *[AceIRL Sample Complexity] AceIRL returns a (ϵ, δ, n)-correct solution with*

$$n \leq \tilde{\mathcal{O}}\left(\min\left[\frac{H^5 R_{\max}^2 SA}{\epsilon^2}, \frac{H^4 R_{\max}^2 SA\epsilon_{\tau-1}^2}{\min_{s,a,h}(A^{*,h}_{\mathcal{M}\cup r}(s,a))^2\epsilon^2}\right]\right)$$

where $\epsilon_{\tau-1}$ depends on the choice of N_E, the number of episodes of exploration in each iteration. $A^{*,h}_{\mathcal{M}\cup r}(s,a)$ is the advantage function of $r \in \operatorname{argmin}_{r\in\mathcal{R}_{\mathfrak{B}}} \max_{h,s,a}(r_h(s,a) - \hat{r}_{k,h}(s,a))$, the reward function from the feasible set $\mathcal{R}_{\mathfrak{B}}$ closest to the estimated reward function $\hat{r}_k$.

This result is the minimum of two terms. The first term is problem independent and it is achieved both by AceIRL Greedy and the full AceIRL. This bound matches the bound we saw previously with a generative model. Hence, AceIRL achieves the same results without access to the generative model. Using (ACE) can yield a better sample complexity, represented by the second term in the minimum. This bound depends on two main components: the ratio $\epsilon_{\tau-1}/\epsilon$ and the advantage function $A^{*,h}_{\mathcal{M}\cup r}(s,a)$. The ratio depends on the choice of N_E, the number of exploration episodes per iteration. If N_E is small, then the ϵ-ratio will be also small. If N_E is large the algorithm will perform similarly to a uniform sampling strategy. Appendix B.5 provides the full proof of this theorem.

Ace-IRL

Implementing AceIRL. To implement the full algorithm, we need to solve an optimization problem:

$$\pi_k \in \operatorname*{argmin}_{\pi} \max_{\hat{\pi} \in \hat{\Pi}_{k-1}} \hat{E}^0_{k+1}(s_0, \hat{\pi}(s_0)) \tag{ACE}$$

Algorithm 1 AceIRL algorithm for IRL in an unknown environment.

Input: significance $\delta \in (0,1)$, target accuracy ϵ, IRL algorithm $\mathscr{A}$, number of episodes N_E
Initialize $k \leftarrow 0,\ \ \epsilon_0 \leftarrow H/10$
while $\epsilon_k > \epsilon/4$ **do**
    Solve (convex) optimization problem (ACE) to obtain π_k
    Explore with policy π_k for N_E episodes, observing transitions and expert actions
    $k \leftarrow k + 1$
    Update $\widehat{P}_k$, $\hat{\pi}_k$, C_k^h, and $\hat{r}_k \leftarrow \mathscr{A}(\mathcal{R}_{\hat{\mathfrak{B}}})$
    Update accuracy $\epsilon_k \leftarrow \max_a \hat{E}^0_k(s_0, a)$
end while
return Estimated reward function $\hat{r}_k$

Experiment

	Uniform sampling (gener. model)	TRAVEL (gener. model) (Metelli et al., 2021)	Random Exploration	AceIRL Greedy	AceIRL (Full)
Four Paths (Figure 1)	1900 ± 71		17840 ± 1886		
– $N_E = 50$		1560 ± 76		24180 ± 1747	$\mathbf{10780 \pm 1369}$
– $N_E = 100$		2000 ± 0		32760 ± 2172	14080 ± 1603
– $N_E = 200$		4000 ± 0		52000 ± 4057	16160 ± 2033
Double Chain (Kaufmann et al., 2021)	1980 ± 66		23640 ± 2195		
– $N_E = 50$		1120 ± 46		16240 ± 842	$\mathbf{11580 \pm 870}$
– $N_E = 100$		2000 ± 0		22200 ± 1329	15440 ± 1031
– $N_E = 200$		4000 ± 0		37200 ± 1664	20400 ± 1629
Metelli et al. (2021):					
Random MDPs ($N_E = 1$)	22 ± 1	27 ± 1	$\mathbf{22 \pm 1}$	23 ± 1	$\mathbf{21 \pm 1}$
Chain ($N_E = 1$)	78 ± 2	76 ± 4	161 ± 8	153 ± 8	$\mathbf{142 \pm 9}$
Gridworld ($N_E = 1$)	43 ± 2	35 ± 2	$\mathbf{45 \pm 2}$	$\mathbf{46 \pm 3}$	$\mathbf{48 \pm 2}$

Table 1: Sample complexity of AceIRL compared to random exploration and methods that use a generative model. We show the number of samples necessary to find a policy with normalized regret less than 0.4. We report means and standard errors computed over 50 random seeds each. For each environment, we highlight in **bold** the method that achieves the best performance without access to a generative model. If multiple methods are within one standard error distance, we highlight all of them. Overall, AceIRL is the most sample efficient method without a generative model if N_E is chosen small enough. In Appendix C.3, we show learning curves for all individual experiments.

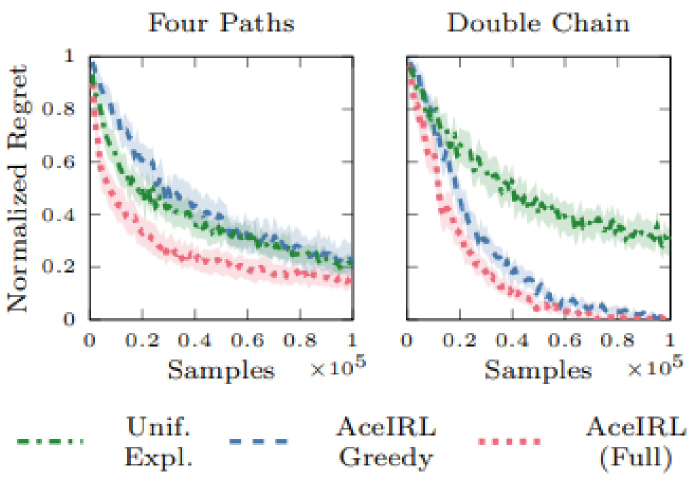


Figure 2: Normalized regret (lower is better) of the policy optimizing for the inferred reward in the estimated MDP as a function of the number of samples. The plots show the mean and 95% confidence intervals computed using 50 random seeds. We use $N_E = 50$.

Our main evaluation metric is a *normalized regret*: $\left(V^{\pi^*,0}_{\mathcal{M}\cup r}(s_0) - V^{\hat{\pi}^*,0}_{\mathcal{M}\cup r}(s_0)\right) / \left(V^{\pi^*,0}_{\mathcal{M}\cup r}(s_0) - V^{\bar{\pi}^*,0}_{\mathcal{M}\cup r}(s_0)\right)$, where π^* is the optimal policy for $\mathcal{M}\cup r$, $\hat{\pi}^*$ is the optimal policy for $\widehat{\mathcal{M}}\cup\hat{r}$, and $\bar{\pi}^*$ is the worst possible policy for r, i.e., the optimal policy for $\mathcal{M}\cup(-r)$.

Thanks