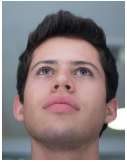
$\Psi\Phi$-Learning: Reinforcement Learning with Demonstrations using Successor Features and Inverse Temporal Difference Learning

Angelos Filos[ϕ]

Clare Lyle[ϕ]



Yarin Gal[ϕ]

Sergey Levine[ψγ]

Natasha Jaques[*ψγ]

Gregory Farquhar[*δ]

[*] equal senior contribution [ϕ]University of Oxford [ψ]University of California, Berkeley [γ]Google Brain Team [δ]DeepMind

Social Reinforcement Learning [α. Problem Setting]

Social Reinforcement Learning [α. Problem Setting]

ENV

Control Markov Process

$\boxed{\mathcal{C}}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

Social Reinforcement Learning [α. Problem Setting]

ENV

EGO

Control Markov Process

Experience

$$\boxed{\mathcal{C}} \xRightarrow{R^{\text{ego}}} \boxed{\mathcal{B}}$$

$$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$$

$$\mathcal{B} = \left\{ \left(\mathsf{s}, \mathsf{a}, r^{\text{ego}}, \mathsf{s}' \right) \right\}$$

Social Reinforcement Learning [α. Problem Setting]

OTHERS

ENV

EGO

No-Reward Demonstrations

Control Markov Process

Experience

$$\mathcal{D} \xleftarrow{\{R^k\}_{k=1}^K} \mathcal{C} \xrightarrow{R^{\text{ego}}} \mathcal{B}$$

$\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

$\mathcal{B} = \{\left(\text{s}, \text{a}, r^{\text{ego}}, \text{s}'\right)\}$

Social Reinforcement Learning [α. Problem Setting]

OTHERS

No-Reward Demonstrations

$\mathcal{D}$

$\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$

ENV

Control Markov Process

$\mathcal{D} \xleftarrow{\{R^k\}_{k=1}^K} \mathcal{C} \xrightarrow{R^{\text{ego}}} \mathcal{B}$

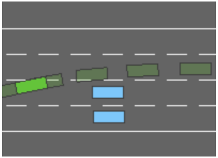
$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

EGO

Experience

$\mathcal{B}$

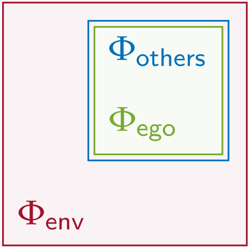
$\mathcal{B} = \left\{ \left(\mathsf{s}, \mathsf{a}, r^{\text{ego}}, \mathsf{s}' \right) \right\}$



Working Example

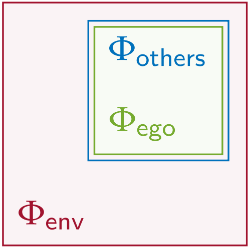
Social Reinforcement Learning [β. Taxonomy of Tasks]

Social Reinforcement Learning [β. Taxonomy of Tasks]

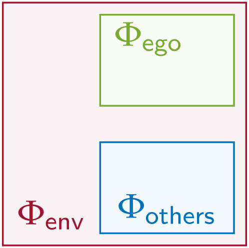


Single-Task
$R^{\text{ego}} = \{R^k\}$

Social Reinforcement Learning [β. Taxonomy of Tasks]

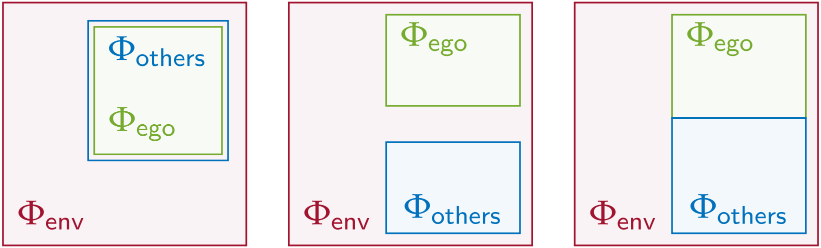


Single-Task
$R^{\text{ego}} = \{R^k\}$



Adversarial Task
$\Phi(R^{\text{ego}}) \cap \Phi(\{R^k\}) = \varnothing$

Social Reinforcement Learning [β. Taxonomy of Tasks]

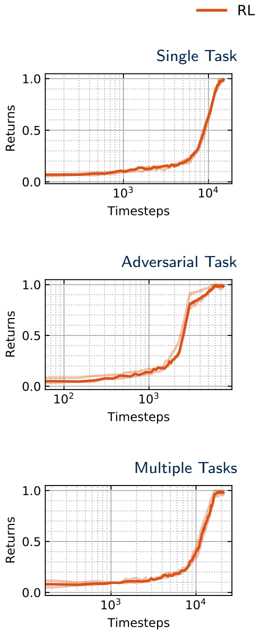


Single-Task
$R^{\text{ego}} = \{R^k\}$

Adversarial Task
$\Phi(R^{\text{ego}}) \cap \Phi(\{R^k\}) = \varnothing$

Multi-Task
$\Phi(R^{\text{ego}}) \cap \Phi(\{R^k\}) \neq \varnothing$

Baselines [α. Reinforcement Learning]



Baselines [α. Reinforcement Learning]

ENV

Control Markov Process

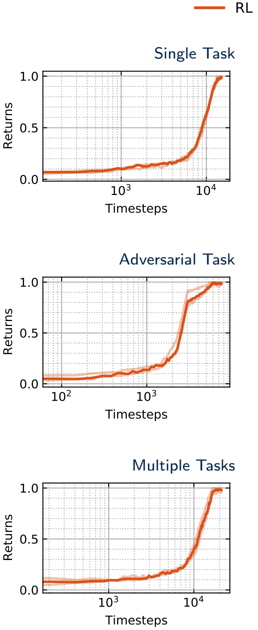
$\boxed{\mathcal{C}} \xrightarrow{R^{\text{ego}}} \boxed{\mathcal{B}}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

EGO

Experience

$\mathcal{B} = \left\{ \left(\mathsf{s}, \mathsf{a}, r^{\text{ego}}, \mathsf{s}' \right) \right\}$



Baselines [α. Reinforcement Learning]

ENV

EGO

Control Markov Process

Experience

$$\boxed{\mathcal{C}} \xRightarrow{R^{\text{ego}}} \boxed{\mathcal{B}}$$

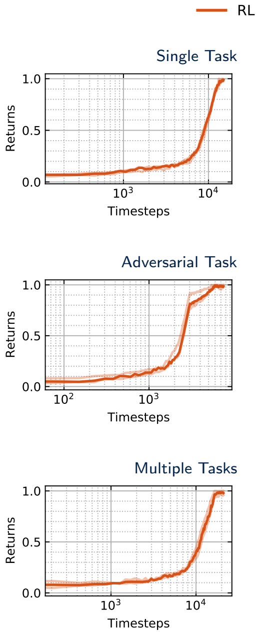
$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

$\mathcal{B} = \left\{ \left(s, a, r^{\text{ego}}, s' \right) \right\}$

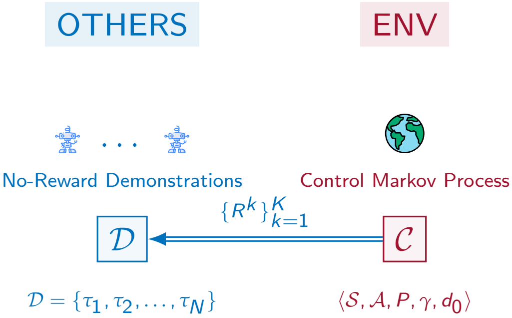
ignores OTHERS' decisions

$\Downarrow \Downarrow \Downarrow$

fails to accelerate its learning from demonstrations



Baselines [β. Learning From Demonstrations]



Baselines [β. Learning From Demonstrations]

OTHERS

ENV

. . .

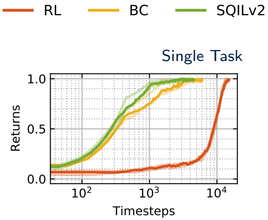
No-Reward Demonstrations

Control Markov Process

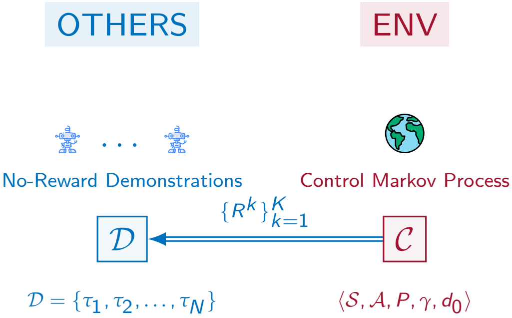
$$\mathcal{D} \xleftarrow{\{R^k\}_{k=1}^{K}} \mathcal{C}$$

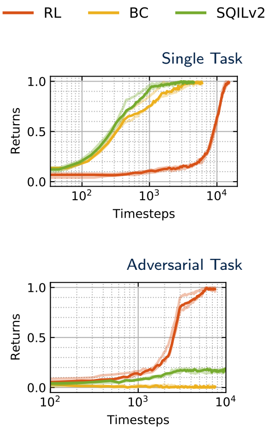
$\mathcal{D} = \{\tau_1, \tau_2, \dots, \tau_N\}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$



Baselines [β. Learning From Demonstrations]





Baselines [β. Learning From Demonstrations]

OTHERS

ENV

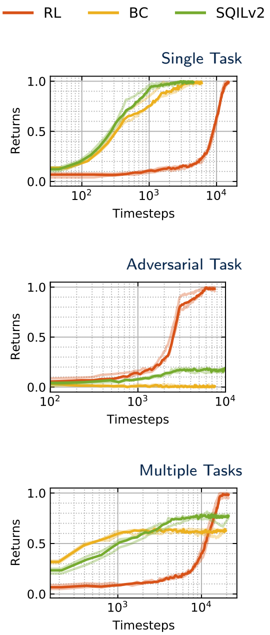
No-Reward Demonstrations

Control Markov Process

$$\mathcal{D} \xleftarrow{\{R^k\}_{k=1}^{K}} \mathcal{C}$$

$\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$



Baselines [β. Learning From Demonstrations]

OTHERS

ENV

No-Reward Demonstrations

Control Markov Process

$$\mathcal{D} \xleftarrow{\{R^k\}_{k=1}^K} \mathcal{C}$$

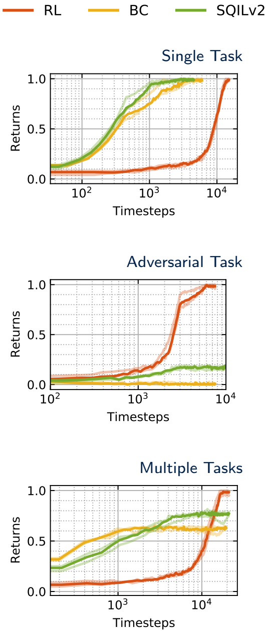
$\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$

$\langle \mathcal{S}, \mathcal{A}, P, \gamma, d_0 \rangle$

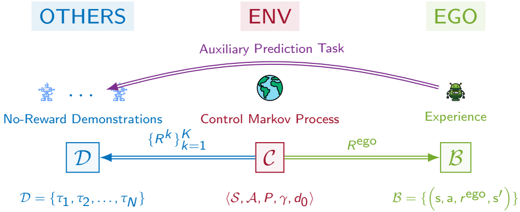
ignores EGO experience

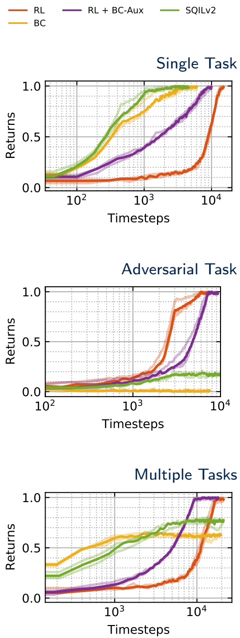
$\Downarrow \Downarrow \Downarrow$

bottlenecked by the quality/relevance of the demonstrations

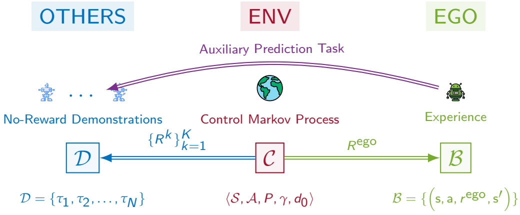


Baselines [γ. Behavioural Cloning as an Auxiliary Task]





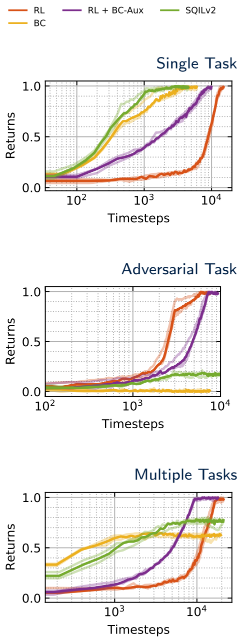
Baselines [γ. Behavioural Cloning as an Auxiliary Task]



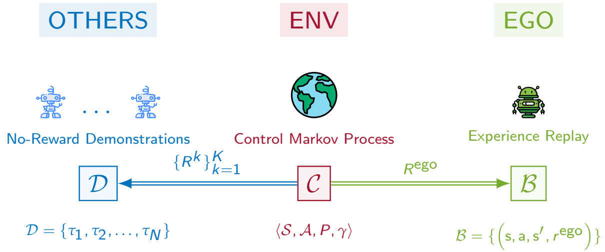
ignores structure in demonstrations

$\Downarrow \Downarrow \Downarrow$

helps only indirectly via representation learning



Desiderata for Social Reinforcement Learners



α. Preserse unbiased asymptotic performance of RL.

- RL, BC, IRL, RL+BC-Aux

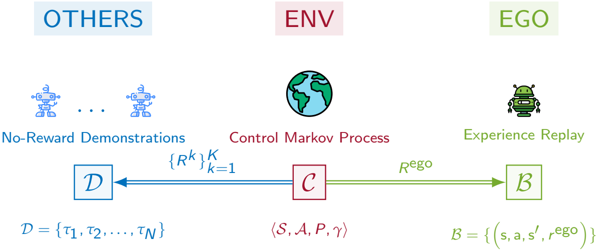
β. Accelerate RL from (no-reward) demonstrations.

- RL, BC, IRL, RL+BC-Aux

γ. Focus on *actionable representations* that inform action selection.

- RL, BC, IRL, RL+BC-Aux

Desiderata for Social Reinforcement Learners



α. Preserse unbiased asymptotic performance of RL.

- RL, BC, IRL, RL+BC-Aux

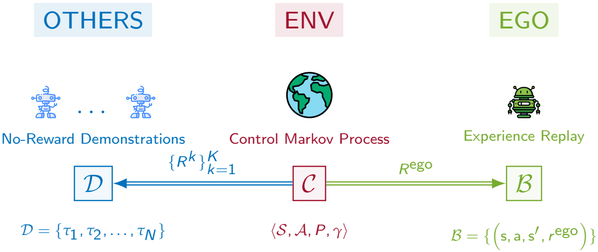
β. Accelerate RL from (no-reward) demonstrations.

- RL, BC, IRL, RL+BC-Aux

γ. Focus on *actionable representations* that inform action selection.

- RL, BC, IRL, RL+BC-Aux

Desiderata for Social Reinforcement Learners



α. Preserse unbiased asymptotic performance of RL.

- RL, BC, IRL, RL+BC-Aux

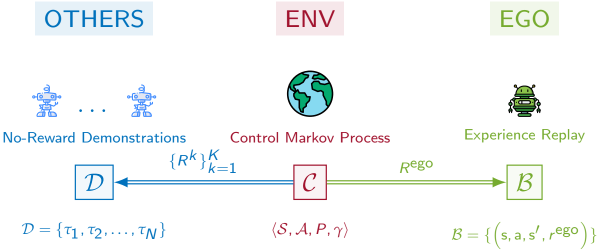
β. Accelerate RL from (no-reward) demonstrations.

- RL, BC, IRL, RL+BC-Aux

γ. Focus on *actionable representations* that inform action selection.

- RL, BC, IRL, RL+BC-Aux

Desiderata for Social Reinforcement Learners



α. Preserse unbiased asymptotic performance of RL.

- RL, BC, IRL, RL+BC-Aux

β. Accelerate RL from (no-reward) demonstrations.

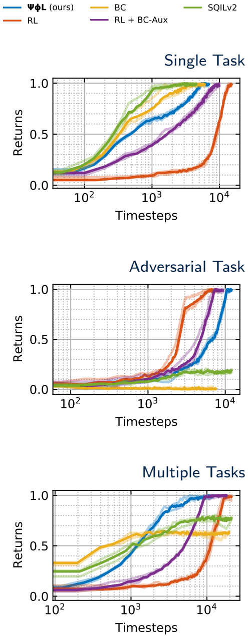
- RL, BC, IRL, RL+BC-Aux

γ. Focus on *actionable representations* that inform action selection.

- RL, BC, IRL, RL+BC-Aux

$\Psi\Phi$-Learning [α. Principle & Results]

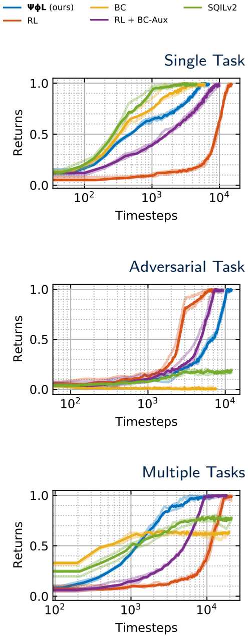
ΨΦ-Learning [α. Principle & Results]



$\Psi\Phi$-Learning [α. Principle & Results]

Modelling Principle

The OTHER agents are goal-directed and optimal for **some** *task. Their behaviour should be integrated to the EGO agent's policy improvement only when it is relevant to its task.*



ΨΦ-Learning [α. Principle & Results]

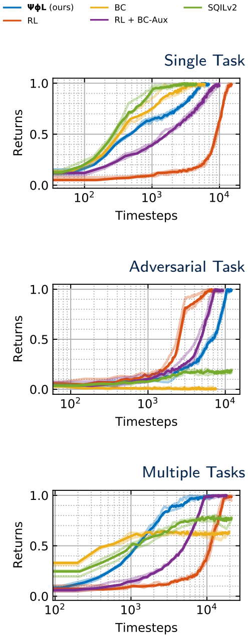
Modelling Principle

The OTHER agents are goal-directed and optimal for **some** *task. Their behaviour should be integrated to the EGO agent's policy improvement only when it is relevant to its task.*

Reasoning about Tasks

α. Task space.

β. Task inference.



ΨΦ-Learning [α. Principle & Results]

Modelling Principle

*The OTHER agents are goal-directed and optimal for **some** task. Their behaviour should be integrated to the EGO agent's policy improvement only when it is relevant to its task.*

Reasoning about Tasks

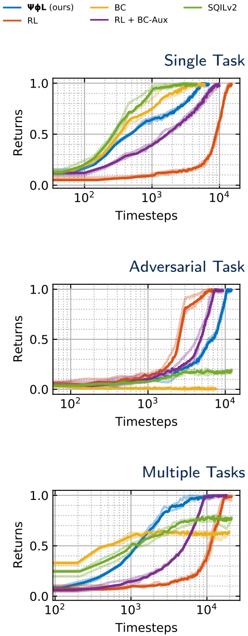
α. Task space.

β. Task inference.

Task-Aware Policy Ops

α. Fast policy evaluation.

β. Fast policy improvement.



$Q(\mathsf{s}, \mathsf{a})$

$\Psi\Phi$-Learning [β. Successor Features Reparametrisation]

$Q^{\pi,\mathbf{w}}(\mathsf{s}, \mathsf{a})$

$\Psi\Phi$-Learning [β. Successor Features Reparametrisation]

$$Q^{\pi,\mathbf{w}}(\mathsf{s},\mathsf{a}) \triangleq \Psi^{\pi}(\mathsf{s},\mathsf{a})^{\top}\mathbf{w}$$

$\Psi\Phi$-Learning [β. Successor Features Reparametrisation]

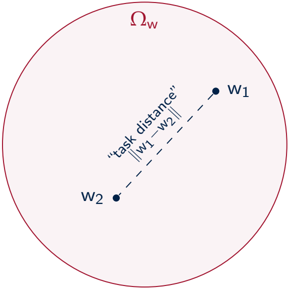
$$Q^{\pi,\mathbf{w}}(\mathbf{s},\mathbf{a}) \triangleq \Psi^{\pi}(\mathbf{s},\mathbf{a})^{\top}\mathbf{w}$$

task-specifig
"preference vector"

$\Psi\Phi$-Learning [β. Successor Features Reparametrisation]

$$Q^{\pi,\mathrm{w}}(\mathrm{s},\mathrm{a}) \triangleq \underbrace{\Psi^{\pi}(\mathrm{s},\mathrm{a})^{\top}\mathrm{w}}_{}$$

task-specifig
"preference vector"



Task space.

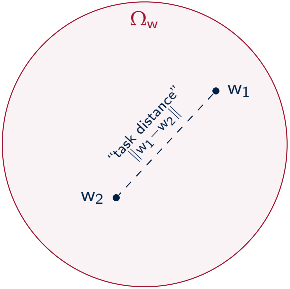
ΨΦ-Learning [β. Successor Features Reparametrisation]

policy-dependent
task-agnostic
"cumulants"

$$\mathbb{E}^{\pi} \sum_{t=0}^{\infty} \overline{\Phi(s_t, a_t)}$$

$$Q^{\pi,w}(s, a) \triangleq \Psi^{\pi}(s, a)^{\top} w$$

task-specifig
"preference vector"



Task space.

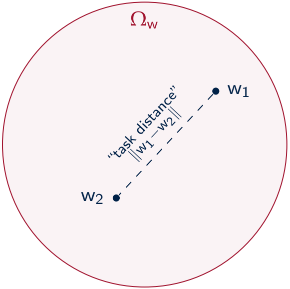
ΨΦ-Learning [β. Successor Features Reparametrisation]

policy-dependent
task-agnostic
"cumulants"

$$\mathbb{E}^{\pi} \sum_{t=0}^{\infty} \overbrace{\Phi(s_t, a_t)}$$

$$Q^{\pi, \mathrm{w}}(\mathrm{s}, \mathrm{a}) \triangleq \Psi^{\pi}(\mathrm{s}, \mathrm{a})^{\top} \mathrm{w}$$

task-specifig
"preference vector"



Task space.

Generalised Policy Improvement (GPI, Barreto et al., 2017)

Given the EGO agent's preference vector $\mathrm{w}^{\mathrm{ego}}$ and the OTHER agents' successor features $\{\Psi^k\}_{k=1}^K$, we can improve the EGO agent's policy by acting according to:

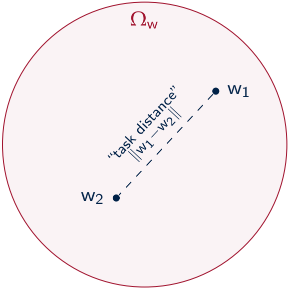
$$\pi'(\mathrm{s}) = \arg\max_{a} \max_{i=[K],\mathrm{ego}} \underbrace{\Psi^i(\mathrm{s}, a)^{\top} \mathrm{w}^{\mathrm{ego}}}_{\substack{\text{evaluate policy } i \\ \text{under task } \mathrm{w}^{\mathrm{ego}}}} \succeq \pi^{\mathrm{ego}}.$$

ΨΦ-Learning [β. Successor Features Reparametrisation]

$$\mathbb{E}^{\pi} \sum_{t=0}^{\infty} \overbrace{\Phi(s_t, a_t)}^{\text{policy-dependent task-agnostic ``cumulants''}}$$

$$Q^{\pi,\mathrm{w}}(\mathrm{s}, \mathrm{a}) \triangleq \Psi^{\pi}(\mathrm{s}, \mathrm{a})^{\top} \mathrm{w}$$

task-specifig
“preference vector”



Task space.

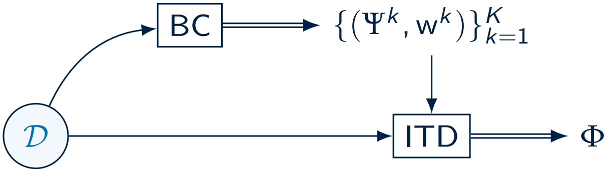
Generalised Policy Improvement (GPI, Barreto et al., 2017)

Given the EGO agent’s preference vector w^{ego} and the OTHER agents’ successor features $\{\Psi^k\}_{k=1}^{K}$, we can improve the EGO agent’s policy by acting according to:

$$\pi'(\mathrm{s}) = \arg\max_{a} \max_{i=[K],\text{ego}} \underbrace{\Psi^{i}(\mathrm{s}, a)^{\top} \mathrm{w}^{\text{ego}}}_{\text{evaluate policy } i \text{ under task } \mathrm{w}^{\text{ego}}} \succeq \pi^{\text{ego}}.$$

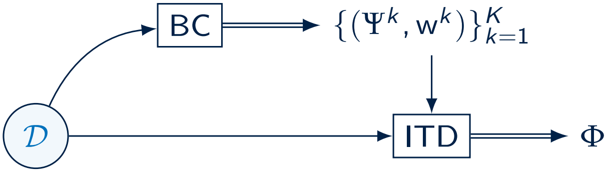
Where do cumulants Φ and preference vectors w come from?

$\Psi\Phi$-Learning [γ. Inverse Temporal Difference Learning]



I/O schematic for inverse TD learning algorithm.

$\Psi\Phi$-Learning [γ. Inverse Temporal Difference Learning]



I/O schematic for inverse TD learning algorithm.

Successor Features & Bellman Consistency $\Rightarrow$ Cumulants

$$\mathcal{L}_{\text{BC-}Q}(\boldsymbol{\theta}_{\Psi^k}, \mathrm{w}^k) \triangleq -\mathbb{E} \log \frac{\exp(\Psi(\mathrm{s}_t, \mathrm{a}_t; \boldsymbol{\theta}_{\Psi^k})^\top \mathrm{w}^k)}{\sum_a \exp(\Psi(\mathrm{s}_t, a; \boldsymbol{\theta}_{\Psi^k})^\top \mathrm{w}^k)}$$

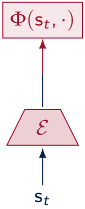
$$\mathcal{L}_{\text{ITD}}(\boldsymbol{\theta}_\Phi, \boldsymbol{\theta}_{\Psi^k}) \triangleq \mathbb{E} \left\| \Psi(\mathrm{s}_t, \mathrm{a}_t; \boldsymbol{\theta}_{\Psi^k}) - \Phi(\mathrm{s}_t, \mathrm{a}_t; \boldsymbol{\theta}_\Phi) - \gamma \Psi(\mathrm{s}_{t+1}, \mathrm{a}_{t+1}; \tilde{\boldsymbol{\theta}}_{\Psi^k}) \right\|.$$

ENV

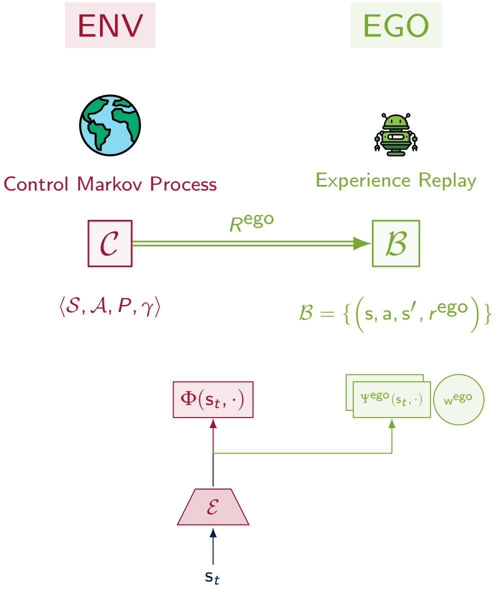
Control Markov Process

$\boxed{\mathcal{C}}$

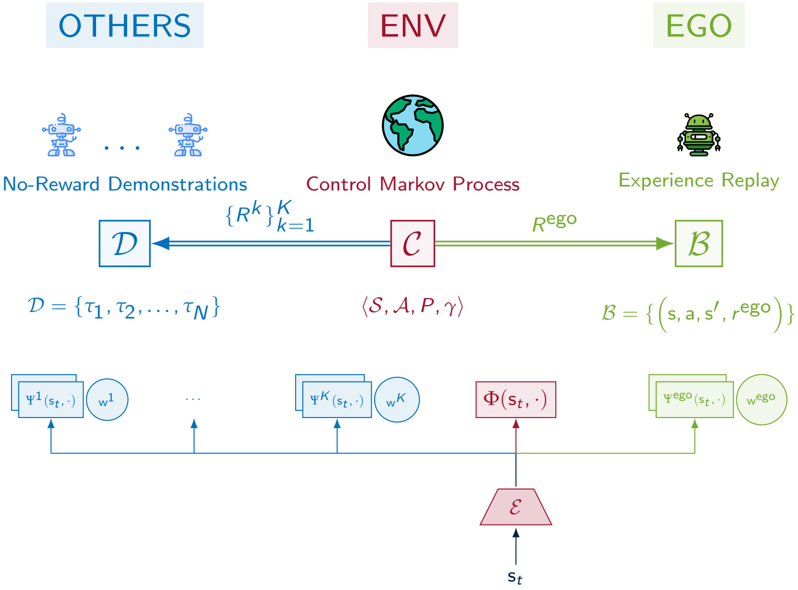
$\langle \mathcal{S}, \mathcal{A}, P, \gamma \rangle$



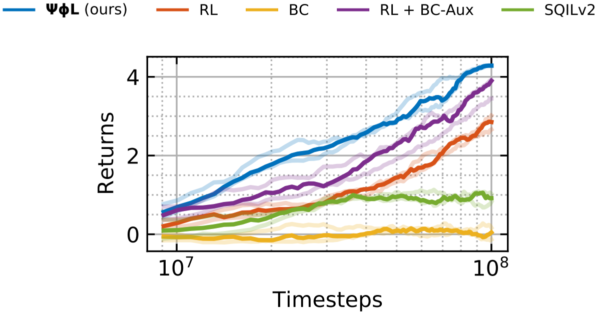
$\Psi\Phi$-Learning [δ. Implementation]

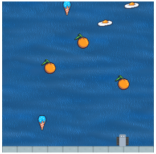


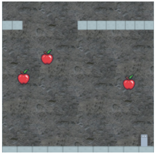
ΨΦ-Learning [δ. Implementation]



Experiments [α. Deep RL from RGB Observations]



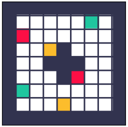




Experiments [β. Few-Shot Transfer to New Reward Functions]

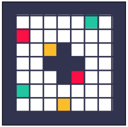
Experiments [β. Few-Shot Transfer to New Reward Functions]

	0-shot		1-shot		100-shot	
Methods	`R-G`	`-R-G`	`R-G`	`-R-G`	`R-G`	`-R-G`
SQILv2♣ (Reddy *et al.*, 2019)	0.0±0.0	−1.0±0.0	0.0±0.0	−1.0±0.0	1.0±0.0	1.0±0.0
ΨΦ-learning ♢	0.2±0.1	−0.4±0.2	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0



Experiments [β. Few-Shot Transfer to New Reward Functions]

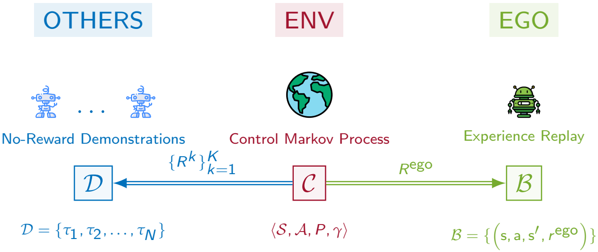
	0-shot		1-shot		100-shot	
Methods	`R-G`	`-R-G`	`R-G`	`-R-G`	`R-G`	`-R-G`
SQILv2♣ (Reddy *et al.*, 2019)	0.0±0.0	−1.0±0.0	0.0±0.0	−1.0±0.0	1.0±0.0	1.0±0.0
ΨΦ-learning ♢	0.2±0.1	−0.4±0.2	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0



Zero-Shot Generalisation Bound of ΨΦ-Learning)

$$\underbrace{Q^*(s,a) - Q^\pi(s,a)}_{\text{sub-optimality gap}} \leq \frac{2}{1-\gamma}\Bigg[\underbrace{(\phi_{\max}\|w_j - w'\|}_{\text{relevance of demonstrations}} \underbrace{+ 2\delta_r) + \|w'\|\delta_\Psi + \frac{1}{(1-\gamma)}\delta_r}_{\text{approximation error}}\Bigg]$$

Summary



ΨΦ-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

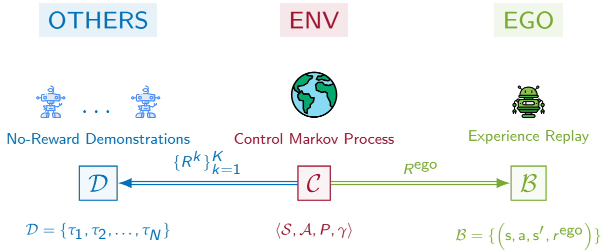
β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of "poor" quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.

Summary



ΨΦ-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

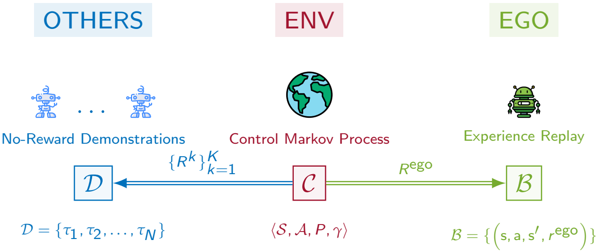
β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of “poor” quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.

Summary



ΨΦ-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

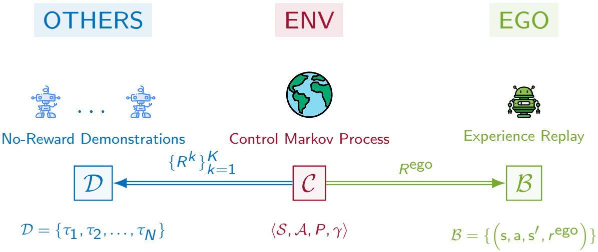
β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of "poor" quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.

Summary



$\Psi\Phi$-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

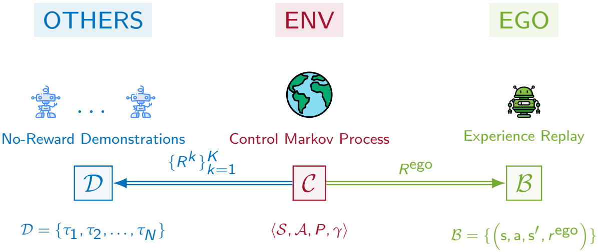
β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of “poor” quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.

Summary



ΨΦ-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

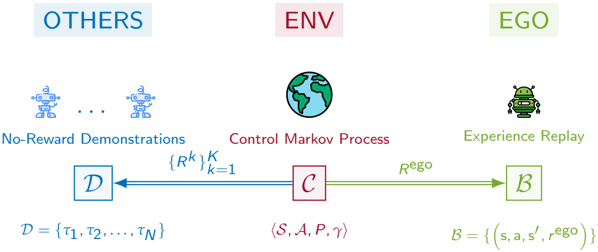
β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of “poor” quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.

Summary



$\Psi\Phi$-Learning

α. Combination of our **offline** IRL method with generalised policy improvement.

β. Utilisation of no-reward demonstrations for accelerating RL.

γ. Graceful fallback to solitary RL when demos are of “poor” quality (theorem).

δ. Few-shot transfer to new reward functions.

ϵ. Scalable for deep RL, RGB observations.