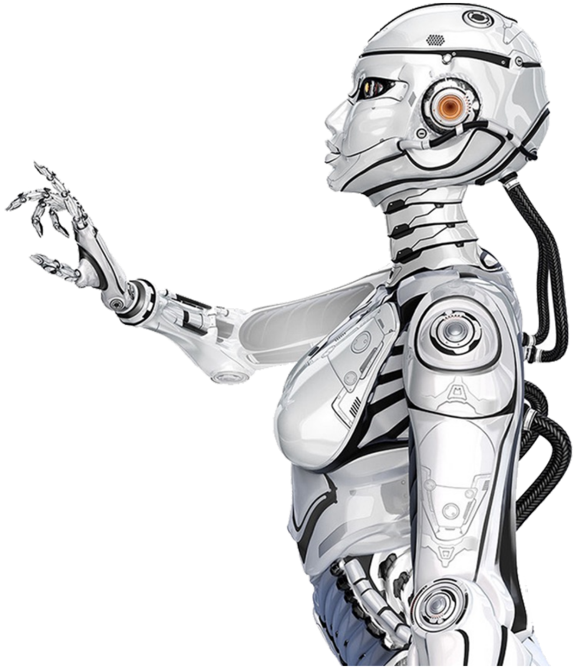
Reinforcement Learning with Multiple Experts: A Bayesian Model Combination Approach

(NIPS 2018)

Method : Replication

Team :王如均, 黃立鈞, 潘柏瑋, 廖晊程

Date : 2022/6/13



Outline

01. Introduction

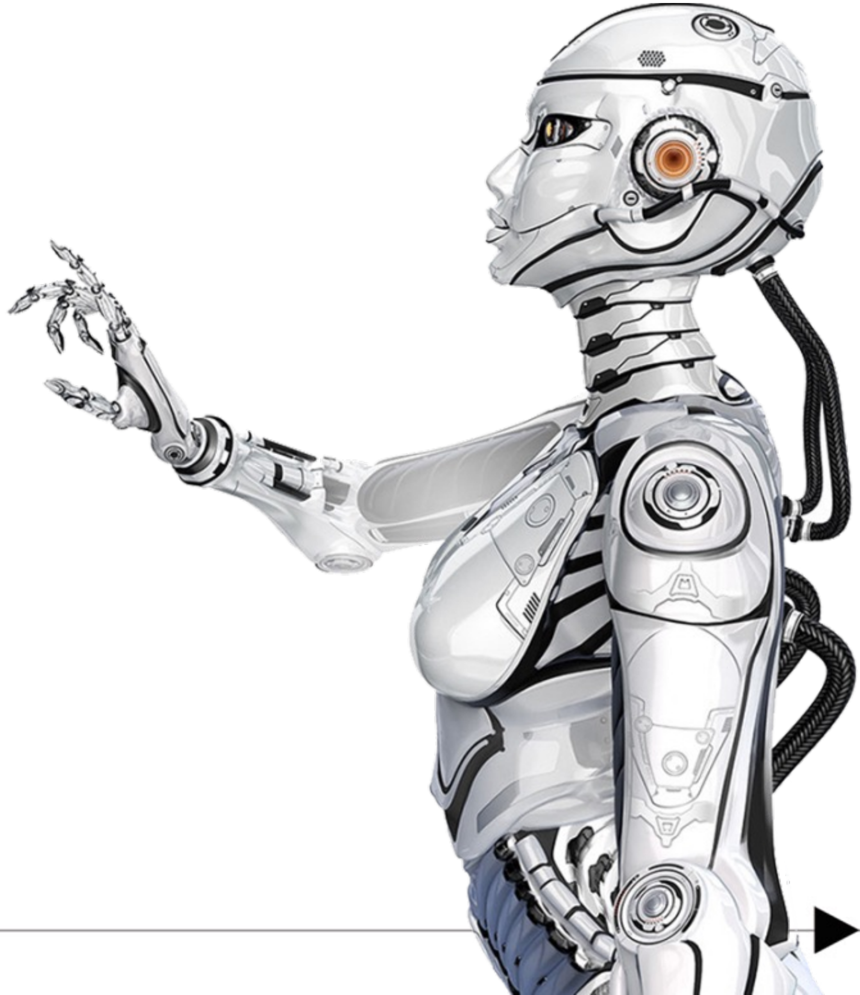
02. Related work

03. Bayesian Reward Shaping

04. Experimental Results

05. Conclusion

01. Introduction



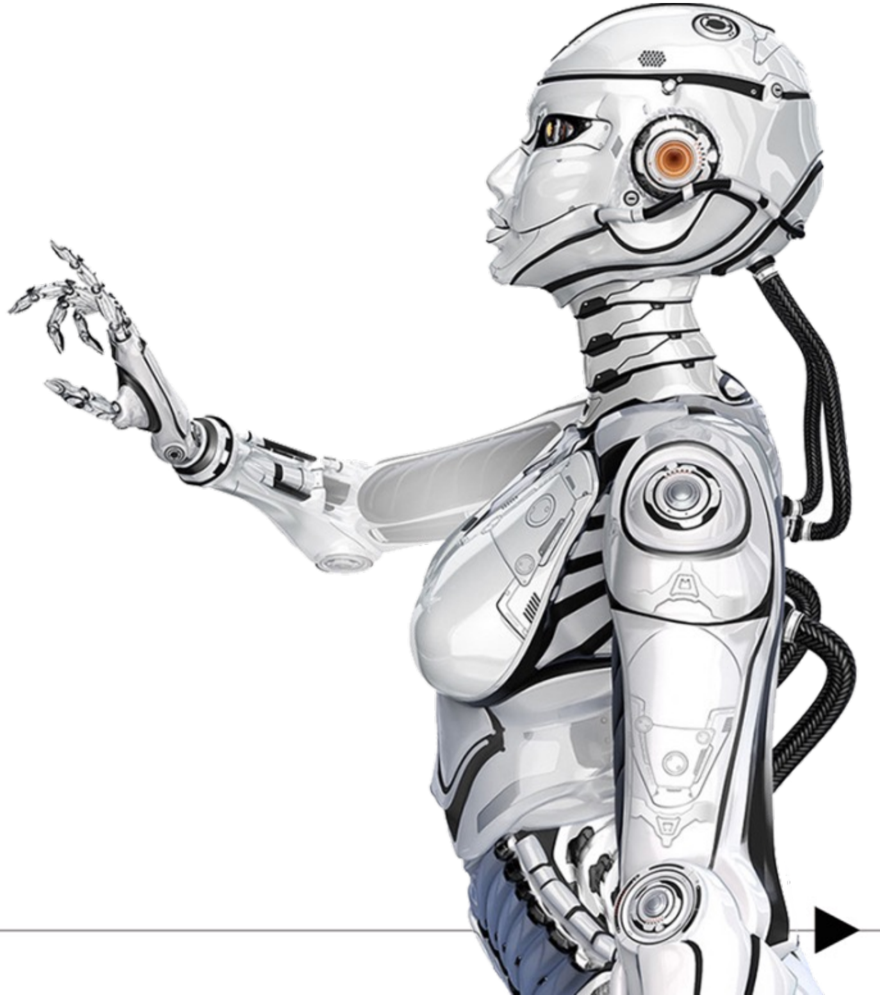
Problem

- Potential based reward shaping is a powerful technique for accelerating convergence of reinforcement learning algorithms. Such information includes an estimate of the optimal value function. However, this information is often biased or inaccurate and can mislead many reinforcement learning algorithms.

Contribution

- This paper apply Bayesian Model Combination with multiple experts in a way that learns to trust a good combination of experts as training progresses. It is both computationally efficient and general, and is shown numerically to improve convergence across discrete and continuous domains.

02. Related work



The approaches try to perturb the value functions to encourage more guided exploration of the state space.

- *Transfer learning :* Learned knowledge from one environment to another by fine-tuning the value function.
- *Inverse reinforcement learning :* The agent recovers an unknown reward function that can then be used for shaping
- Human expert directly provide the learning agent

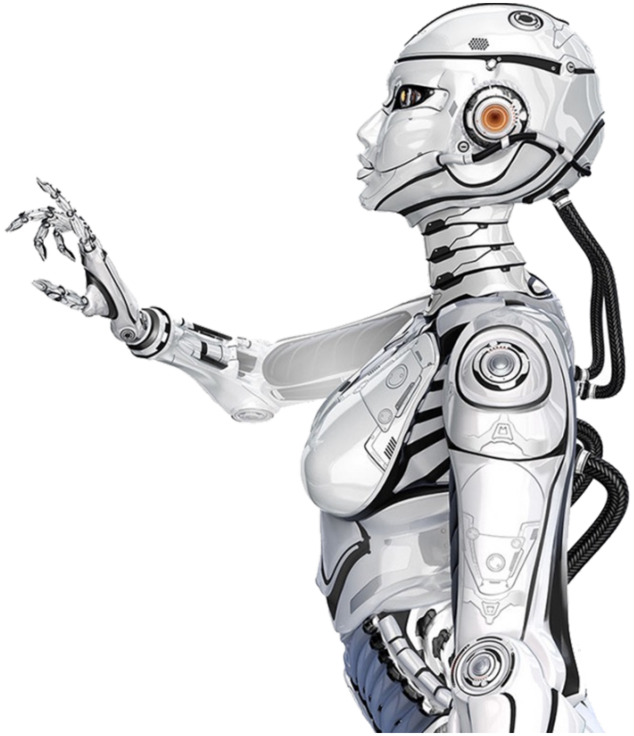
Policy shaping

- Reshapes the learned policies by on-line methods to learn a reward shaping function, but only for model-based learning using R-Max or model-free learning with multi-grid discretization

Combining multiple models/experts

- *Ensemble learning*
- *Kernel regression*
- *Contextual bandits*
- *Bayesian model averaging*

03. Bayesian Reward Shaping



Potential-Based Reward Shaping

The idea of *reward shaping* is to incorporate prior knowledge about the domain in the form of additional rewards during training to speed up convergence towards the optimal policy

- Form of additional rewards

$$R'(s,a,s') = R(s,a,s') + F(s,a,s')$$

- *Potential-based reward shaping*

$$F(s,a,s') = \gamma\Phi(s') - \Phi(s)$$

- Non-stationary time-dependent potential functions

$$F(s,a,t,s',t') = \gamma\Phi(s',t') - \Phi(s,t)$$

Bayesian Reward Shaping

Bayesian model averaging (BMA)	✓ ***Bayesian model combination (BMC)***
BMA converges asymptotically toward the optimal *hypothesis*	BMC converges toward the optimal *ensemble*

The BMC approach has two clear advantages over BMA:

(1) when two or more potential functions are optimal, it will converge to a linear combination of them

(2) It provides an estimator with reduced variance.

This paper shows how BMC can be used to incorporate imperfect advice from multiple experts into RL problems, all with the same space and time complexity as TD-learning.

Bayesian Model Combination

- w (weight vectors) : Interpreted as categorical distributions over experts and can learn the optimal distribution over experts

$$\mathcal{S}^{N-1} = \left\{\mathbf{w} \in \mathbb{R}^N : \textstyle\sum_{i=1}^{N} w_i = 1,\ w_i \geq 0\right\}$$

$$\begin{aligned}
\rho_t(s,a) = \mathbb{E}\left[q_{s,a}|\mathcal{D}\right] &= \int_{\mathbb{R}} q\,\mathbb{P}\left(q|\mathcal{D}\right)\mathrm{d}q \\
&= \int_{\mathbb{R}} q \int_{\mathcal{S}^{N-1}} \mathbb{P}\left(q|\mathcal{D},\mathbf{w}\right)\mathbb{P}\left(\mathbf{w}|\mathcal{D}\right)\mathrm{d}\mathbf{w}\,\mathrm{d}q = \int_{\mathbb{R}} q \int_{\mathcal{S}^{N-1}} \sum_{i=1}^{N} \mathbb{P}\left(q|i\right) w_i \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w}\,\mathrm{d}q \\
&= \sum_{i=1}^{N} \int_{\mathbb{R}} q\,\mathbb{P}\left(q|i\right) \int_{\mathcal{S}^{N-1}} w_i \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w}\,\mathrm{d}q = \sum_{i=1}^{N} \int_{\mathbb{R}} q\,\mathbb{P}\left(q|i\right)\mathbb{E}_{\pi_t}\left[w_i\right]\mathrm{d}q \\
&= \sum_{i=1}^{N} \mathbb{E}_{\pi_t}\left[w_i\right] \int_{\mathbb{R}} q\,\mathbb{P}\left(q|i\right)\mathrm{d}q = \sum_{i=1}^{N} \mathbb{E}_{\pi_t}\left[w_i\right]\mathbb{E}\left[q_{s,a}|i\right],
\end{aligned}$$

The total return can be written as a linear combination of individual return “contributions" from each expert model, weighted by the expected posterior belief that the expert is correct .

Posterior Approximation using Moment Matching

Bayes' theorem

$$\pi_{t+1}(\mathbf{w}) = \mathbb{P}\left(\mathbf{w}|\mathcal{D}, d\right) \propto \mathbb{P}\left(d|\mathbf{w}\right)\pi_t(\mathbf{w}) \propto \sum_{i=1}^{N} \mathbb{P}\left(d|i\right)\mathbb{P}\left(i|\mathbf{w}\right)\pi_t(\mathbf{w}) = \frac{1}{C_{t+1}}\sum_{i=1}^{N} e_i w_i \pi_t(\mathbf{w}),$$

$$C_{t+1} = \int_{\mathcal{S}^{N-1}} \sum_{i=1}^{N} e_i w_i \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w} = \sum_{i=1}^{N} e_i \int_{\mathcal{S}^{N-1}} w_i \pi_t(\mathbf{w})\,\mathrm{d}\mathbf{w} = \sum_{i=1}^{N} e_i\,\mathbb{E}_{\pi_t}\left[w_i\right]$$

Multivariate *Dirichlet distribution with parameters α*

- Generalized moments

$$\mathbb{E}_f\left[\prod_{i=1}^{N} w_i^{n_i}\right] = \frac{\Gamma\left(\sum_{i=1}^{N}\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{N}(\alpha_i+n_i)\right)}\prod_{i=1}^{N}\frac{\Gamma(\alpha_i+n_i)}{\Gamma(\alpha_i)},\ n_i \geq 0.$$

- Density function

$$f(\mathbf{w};\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{N}\alpha_i\right)}{\prod_{i=1}^{N}\Gamma(\alpha_i)}\prod_{i=1}^{N} w_i^{\alpha_i-1},\ \mathbf{w}\in\mathcal{S}^{N-1}$$

Posterior Approximation using Moment Matching (Cont.)

- Approximate moment matching with proposal Dir (α) by solving the system of equations

$$m_i = \frac{\alpha_i}{\alpha_0},\ i = 1, 2 \ldots N-1$$

$$s_1 = \frac{\alpha_1(\alpha_1+1)}{\alpha_0(\alpha_0+1)}$$

Unique positive solution

$$\alpha_0 = \frac{m_1 - s_1}{s_1 - m_1^2}$$

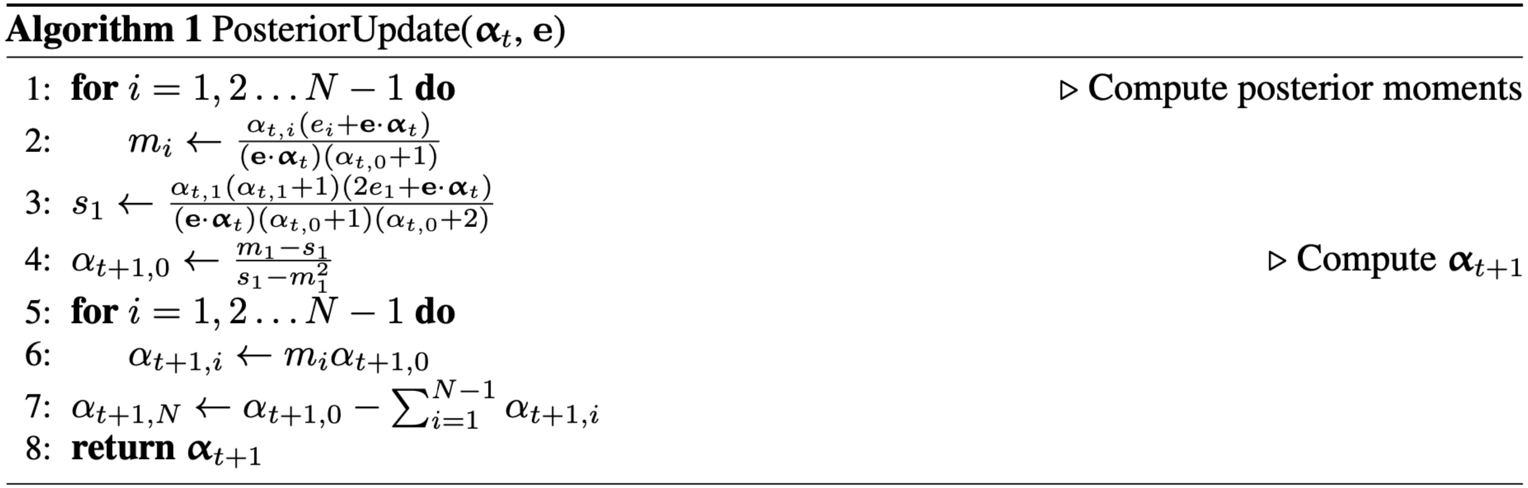
$$\alpha_i = m_i \alpha_0 = m_i \left(\frac{m_1 - s_1}{s_1 - m_1^2}\right),\ i = 1, 2 \ldots N-1$$

$$\begin{aligned} m_i &= \mathbb{E}_{\pi_{t+1}}\left[w_i\right] = \int_{\mathcal{S}^{N-1}} \frac{\alpha_{t,0}}{\mathbf{e}\cdot\boldsymbol{\alpha}_t} \sum_{j=1}^{N} e_j w_j w_i \pi_t(\mathbf{w})\, \mathrm{d}\mathbf{w} \\ &= \frac{\alpha_{t,0}}{\mathbf{e}\cdot\boldsymbol{\alpha}_t} \sum_{j=1}^{N} e_j \int_{\mathcal{S}^{N-1}} w_j w_i \pi_t(\mathbf{w})\, \mathrm{d}\mathbf{w} = \frac{\alpha_{t,0}}{\mathbf{e}\cdot\boldsymbol{\alpha}_t} \sum_{j=1}^{N} e_j\, \mathbb{E}_{\pi_t}\left[w_i w_j\right] \\ &= \frac{\alpha_{t,0}}{\mathbf{e}\cdot\boldsymbol{\alpha}_t} \left(e_i\, \mathbb{E}_{\pi_t}\left[w_i^2\right] + \sum_{j\neq i} e_j\, \mathbb{E}_{\pi_t}\left[w_i w_j\right] \right) \\ &= \frac{\alpha_{t,0}}{\mathbf{e}\cdot\boldsymbol{\alpha}_t} \left(e_i \frac{\alpha_{t,i}(\alpha_{t,i}+1)}{\alpha_{t,0}(\alpha_{t,0}+1)} + \sum_{j\neq i} e_j \frac{\alpha_{t,i}\alpha_{t,j}}{\alpha_{t,0}(\alpha_{t,0}+1)} \right) \\ &= \frac{\alpha_{t,i}(e_i + \mathbf{e}\cdot\boldsymbol{\alpha}_t)}{(\mathbf{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)}. \end{aligned}$$

$$s_1 = \frac{\alpha_{t,1}(\alpha_{t,1}+1)(2e_1 + \mathbf{e}\cdot\boldsymbol{\alpha}_t)}{(\mathbf{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)(\alpha_{t,0}+2)}$$

Algorithm

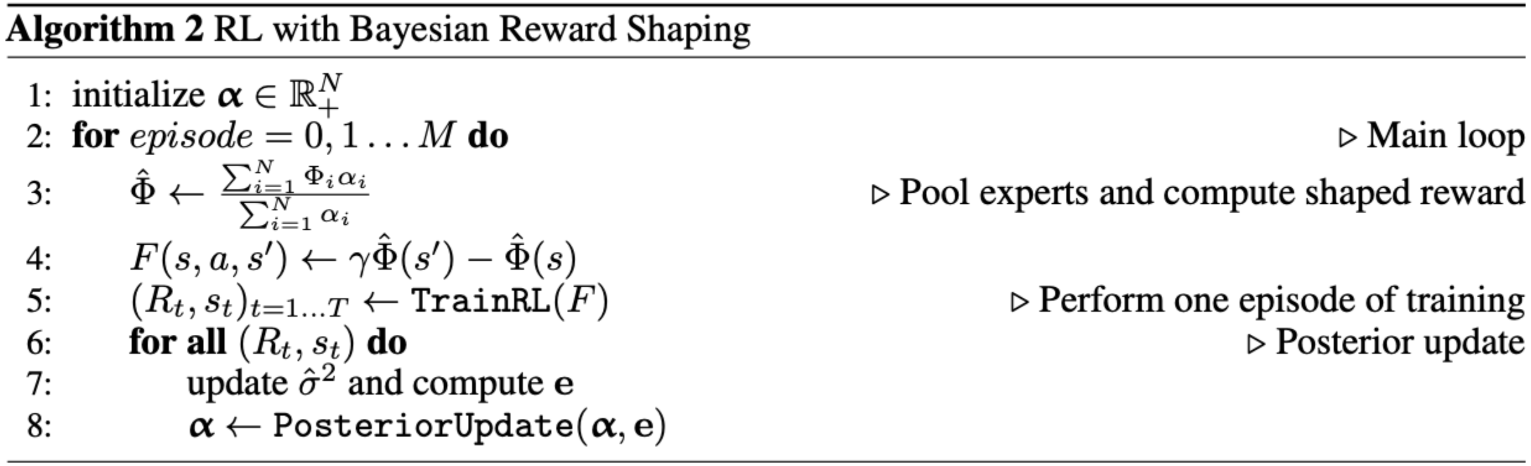
Algorithm1 : Posterior Update



Algorithm 1 PosteriorUpdate($\boldsymbol{\alpha}_t$, $\mathbf{e}$)

```
for $i = 1, 2 \ldots N-1$ do                                   ▷ Compute posterior moments
    $m_i \leftarrow \frac{\alpha_{t,i}(e_i + \mathbf{e}\cdot\boldsymbol{\alpha}_t)}{(\mathbf{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)}$
$s_1 \leftarrow \frac{\alpha_{t,1}(\alpha_{t,1}+1)(2e_1 + \mathbf{e}\cdot\boldsymbol{\alpha}_t)}{(\mathbf{e}\cdot\boldsymbol{\alpha}_t)(\alpha_{t,0}+1)(\alpha_{t,0}+2)}$
$\alpha_{t+1,0} \leftarrow \frac{m_1 - s_1}{s_1 - m_1^2}$                    ▷ Compute $\boldsymbol{\alpha}_{t+1}$
for $i = 1, 2 \ldots N-1$ do
    $\alpha_{t+1,i} \leftarrow m_i \alpha_{t+1,0}$
$\alpha_{t+1,N} \leftarrow \alpha_{t+1,0} - \sum_{i=1}^{N-1} \alpha_{t+1,i}$
return $\boldsymbol{\alpha}_{t+1}$
```

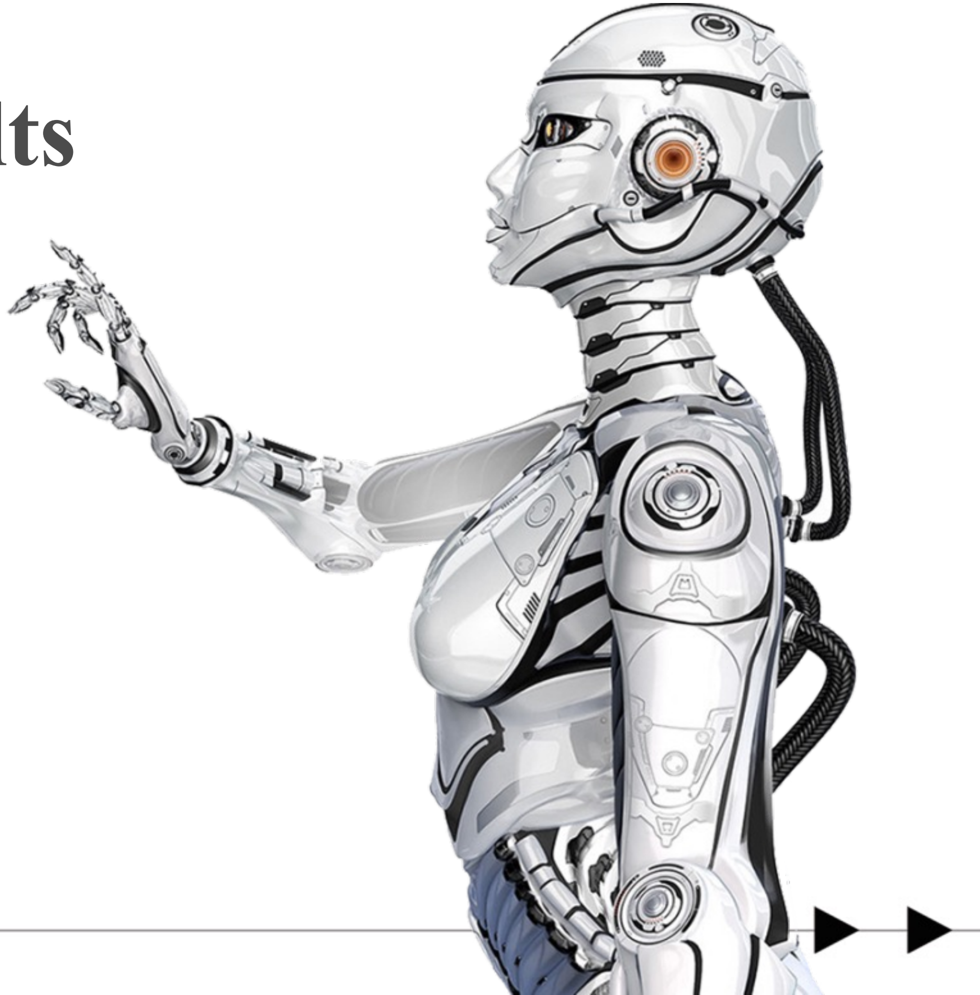
Algorithm 2 : RL with Bayesian Reward Shaping



Algorithm 2 RL with Bayesian Reward Shaping

```
initialize $\boldsymbol{\alpha} \in \mathbb{R}_+^N$
for $episode = 0, 1 \ldots M$ do                                ▷ Main loop
    $\hat{\Phi} \leftarrow \frac{\sum_{i=1}^{N} \Phi_i \alpha_i}{\sum_{i=1}^{N} \alpha_i}$          ▷ Pool experts and compute shaped reward
    $F(s, a, s') \leftarrow \gamma\hat{\Phi}(s') - \hat{\Phi}(s)$
    $(R_t, s_t)_{t=1\ldots T} \leftarrow \texttt{TrainRL}(F)$            ▷ Perform one episode of training
    for all $(R_t, s_t)$ do                                      ▷ Posterior update
        update $\hat{\sigma}^2$ and compute $\mathbf{e}$
        $\boldsymbol{\alpha} \leftarrow \texttt{PosteriorUpdate}(\boldsymbol{\alpha}, \mathbf{e})$
```

04. Experimental Results



CartPole

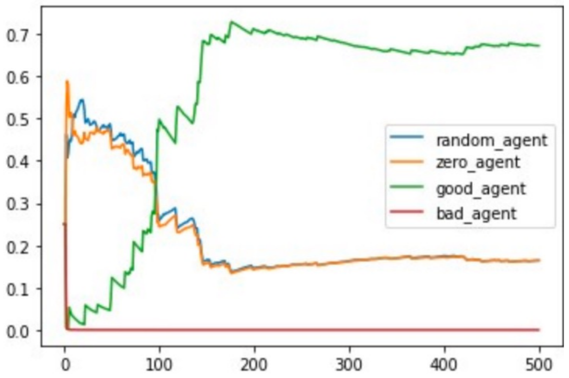
This is a classical control problem implemented in OpenAI Gym

Parameters	Value
γ	0.95
ε	1, Decay:0.98
α	0.5, Decay:0.99
Reward	+1 at every step as long as the pole is upright
Stoppage	Last 5 episodes is 500

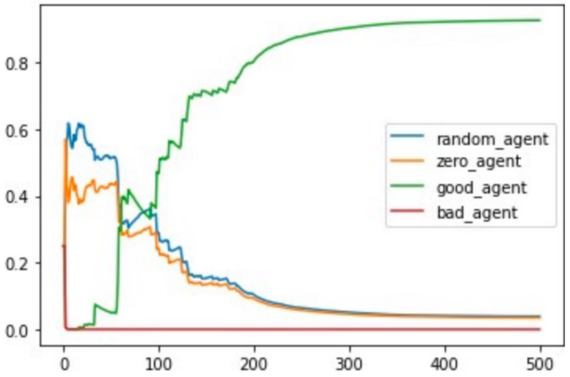
The expert used：

- Optimal:Complete trained agent
- Good：Partially trained function
- Bad : -good
- zero: Return 0 only

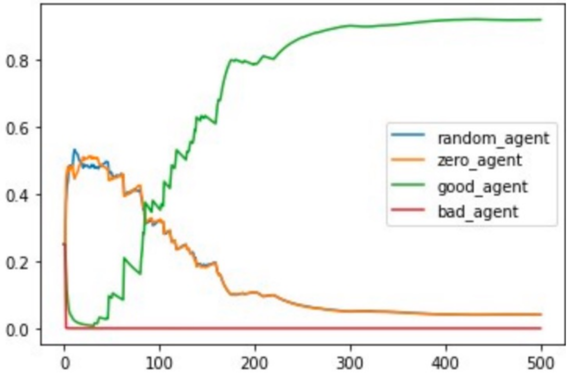
CartPole

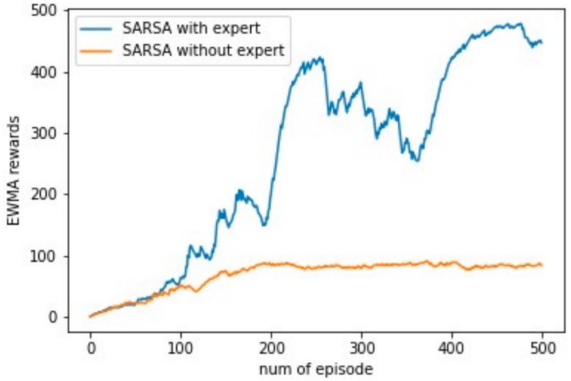


Q learning



SARSA


random_agent
zero_agent
good_agent
bad_agent


SARSA with expert
SARSA without expert
EWMA rewards
num of episode