IVDR: Imitation learning with Variational inference and Distributional Reinforcement learning to find Optimal Driving Strategy

Kihyung Joo and Simon S. Woo, 2021, 20th IEEE International Conference on Machine Learning and Applications (ICMLA)

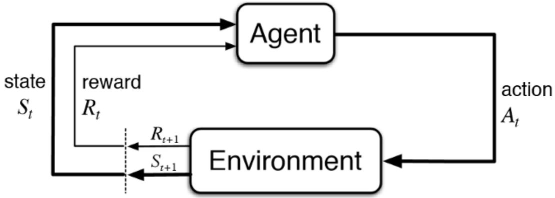
Lin, Syue-Cian Gordon
林學謙

Department of electrical engineering
National Tsing Hua University

January 6, 2023

Paper Aim

Due to various traffic conditions, it is not easy to apply a rule-based driving method.
Reinforcement learning can deal with complex conditions.
Through imitation learning, RL can converge faster.
Finally, variational inference can overcome a local minimum, and choose an optimal policy.



Environment

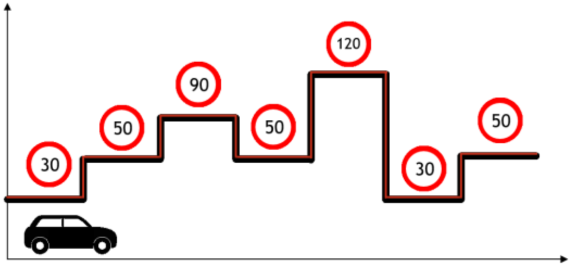


Figure 1. Illustration of the car driving with the different speed limits, where the number in the red circle indicates the speed limit, and the X-axis is the time and the Y-axis is the speed of a car. The agent observes the current and future speed limits to determine whether to accelerate or decelerate.

The state consists of the current vehicle speed, past acceleration, current speed limit, future speed limit, the farther future speed limit, the distance remaining to the future speed limit, and the remaining distance to the farther future speed limit.

On-policy VS Off-policy

Policy definition

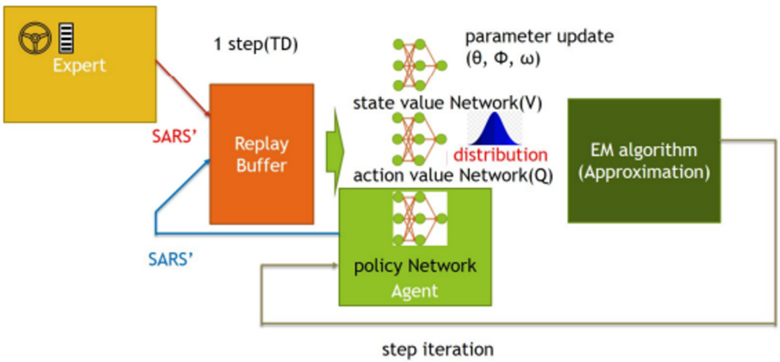
Given state

$$\pi(a|s)$$

The policy provides the probability of the action

- On-policy
  - 會受到reward影響，可能無法選出最短路徑
  - 保守路徑
- Off-policy
  - 會有最短路徑，但收斂的速度可能較慢

IVDR archetecture pipeline



Imitation learning

Imitation learning

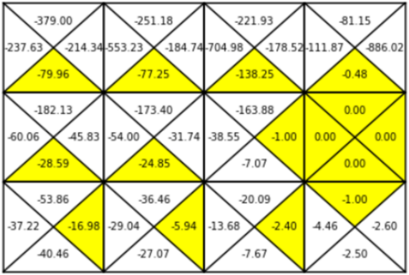
- 屬於off-policy的一種
- 根據expert的數據進行訓練
- Behavior Cloning 行為複製
- 使用Dataset Aggregation來增加數據

Dataset aggregation

1. 第一批資料進行訓練，訓練出π_n
2. 並且讓π_n放進環境中看其observations
3. experts來糾正actorπ_n ，產生新的dataset
4. 並且與先前的資料一起訓練actor π_{n+1}
5. Repeat

Swamp			
S_0			S_T

	1	2	3	4
1	-107.24	-170.08	-115.82	-175.2
2	-35.92	-34.36	-21.39	0.0
3	-25.57	-14.84	-9.48	-1.39



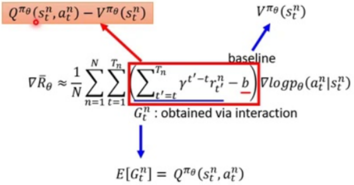
Reinforcement Learning: Actor Critic

Critic

- Neural Network
- State value Network
- State-action value Network
- input state to get the state value and state-action value

Actor

- Neural Network
- input state to get the action



Variational Inference

- 用來解決難以計算的後驗分佈$p(z|x, \theta)$
- z is a latent variable
- simplifying $p(x|\theta)$

latent variable meaning

- 隱變量使得數學模型變得較為簡單
- 隱變量也可以想成原本模型隱含的特性

Ex. 數學模型相當複雜，結果其行為可能跟Gaussian Mixture Model差不多

$$p(x) = \frac{p(x, z)}{p(z|x)}$$

$$p(z|x) = \frac{p(x, z)}{p(x)}$$

Variational Inference

假設存在$q(z|\theta) = p(z|\theta, x)$

$$p(x|\theta) = E_{q(z|\theta)}[p(x|\theta)] = E_{q(z|\theta)}[log\frac{p(x,z|\theta)}{p(z|x,\theta)}]$$

$$= E_{q(z|\theta)}[log\frac{p(x,z|\theta)}{q(z|\theta)}\frac{q(z|\theta)}{p(z|x,\theta)}]$$

$$= E_{q(z|\theta)}[log\frac{p(x,z|\theta)}{q(z|\theta)}] + KL(q(z|\theta)||p(z|x,\theta))$$

目的是為了求$p(x|\theta)$

Evidence lower bound

$$E_{q(z|\theta)}[log\frac{p(x,z|\theta)}{q(z|\theta)}]$$

Variational Inference

Maximize evidence lower bound, we can minimize KL divergence.
When the KL divergence is smaller than tolerance, $q(z|\theta)$ is equal to $P(z|x,\theta)$.
Evidence lower bound $\approx E_{q(z|\theta)}[log\, p(x,z|\theta)] \approx Q(\theta,\theta^t)$.
$E_{q(z|\theta)}[log\, q(z|\theta)]$ is conditional entropy, and it is constant.
Because of this, we can get the approximation of evidence lower bound.
Our target is the maximum of evidence lower bound.

EM Algorithms

EM algorithms

E-step:

$$Q(\theta, \theta^{old}) = E_{q(z|\theta^{old})}[\log p(x, z|\theta)]$$

M-step:

$$\theta^{new} = argmax_{\theta} Q(\theta, \theta^{old})$$

這裡的θ可以當作是新的點，因為實際操作EM algorithms，需要將初始值帶入(θ^t)，然後透過gradient來求出各個最大值的位置。

Realistic

The loss function of neural network

Distributional Reinforcement Learning

Common state-action function

Given State and action，使用期望值對累積回報進行建模，回傳一個scalar，而非變數。

Distributional state-action function

Given state and action，直接對於累積回報進行建模，回傳一個分佈，可以獲得更多的資訊。

因為環境的隨機性，使用Distributional state-value function可以更好的應對，此題的環境為車速的多樣變化。(如果已經收斂了，將難以逃脫出local minimun)

Algorithms

Algorithm 1 Imitation with Variational Inference and Distributional Reinforcement learning (IVDR)

Initialize parameter vectors : $\phi,\ \overline{\phi},\ \omega,\ \theta,\ D\ \leftarrow\ \{\}$
$D^{expert} \leftarrow \{(s_t{}^{expert}, a_t{}^{expert} r_t{}^{expert} s_{t+1}{}^{expert}), \cdots\}$
for each iteration **do**
  for each episode step **do**
    $a_t \sim \pi^q (a|s_t; \theta))$
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    $D \leftarrow D \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
    $D \leftarrow D \cup \left\{\left(s_t^{expert}, a_t^{expert}, r_t^{expert}, s_{t+1}^{expert}\right)\right\}$
    for each gradient step **do**
      $\phi \leftarrow \phi - \lambda \nabla_\phi J^V(\phi)$
      $J^Z(\omega) \leftarrow \sim N\left(J^Q(\omega), \sigma^2\right)$
      $\omega \leftarrow \omega - \lambda \nabla_\omega J^Z(\omega)$
      $\theta \leftarrow \theta - \lambda \nabla_\theta J^{\pi^q}(\theta)$
      $\overline{\phi} \leftarrow \tau \overline{\phi} + (1-\tau) \overline{\phi}$
    end for
  end for
end for

Algorithms

$$J^V(\phi) = E_{s_t \to D}[\frac{1}{2}(V_\phi(s_t) - E_{a_t \to \pi_\theta}[Q_w(s_t, a_t)])^2]$$

$$J^Q(w) = E_{(h_t, r_t, s_{t+1})} \to D[\frac{1}{2}(r_t + \gamma V_{\bar{\phi}} - Q_w(h_t))^2]$$

$$J^{\pi^q}_{IVDR}(\theta) = E_{h_t \to D}[\log \pi_\theta(a_t|s_t) * (\alpha - (Z_w(h_t) - V_{\bar{\phi}}))]$$

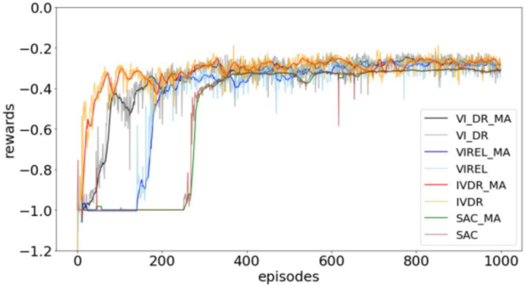
$$J^{\pi^q}_{virel}(\theta) = E_{h_t \to D}[\log \pi_\theta(a_t|s_t) * (\alpha - (Q_w(h_t) - V_{\bar{\phi}}))]$$

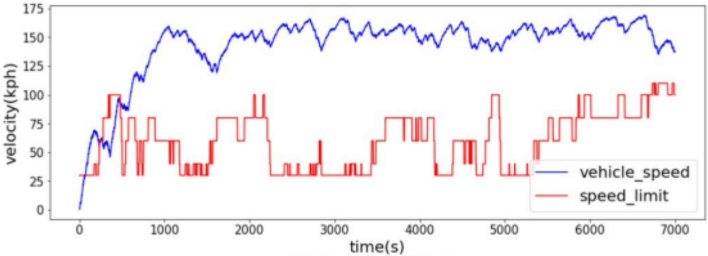
Z_w	Distributional function from Q_w
h_t	s_t and a_t

Result

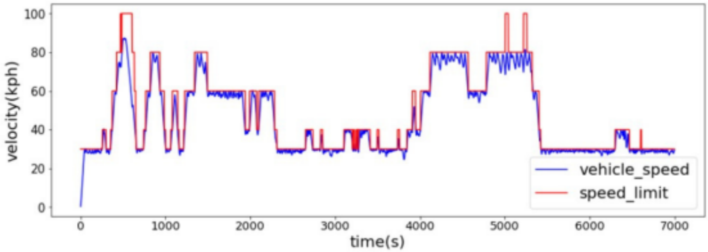
Table I
PERFORMANCE COMPARISON BETWEEN MODELS

	Unit	**IVDR**	**SAC**	**VIREL**	**VI_DR**
Learning speed (Threshold : -0.4)	Iteration	**40**	289	173	80
Average of rewards	Score	-0.27	-0.31	-0.28	-0.27
Standard deviation of rewards	Score	0.027	**0.007**	0.025	0.032





< Before learning >



< After IVDR learning >

[1] 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)

[2] odie's whisper, 漫談Variational Inference (一) https://odie2630463.github.io/2018/08/21/vi-1/

[3] Book 李宏毅老師Deep Reinforcement Learning 2018課程筆記 https://hackmd.io/@shaoeChen/Bywb8YLKS/https%3A%2F%2Fhackmd.io%2F%40shaoeChen%2FH1aW8iEhS

[4] Shweta Bhatt, Reinforcement Learning 101 https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292