Partial Observability in DRL

Part 1: POMDPs, (A)DRQN & DVRL
Most of the World is only Partial Observable

Occlusions

Latent Causes

Intentions
From MDP to ...
From MDP to POMDP
Slightly more formal

7-Tuple: \((S, A, T, R, \Omega, O, \gamma)\)

\(s \in S\) is a state from the set of States

\(a \in A\) is an action set of Actions

\(T(s_{t+1} | s_t, a_t)\) is the transition probabilities

\(R: S \times A \to \mathbb{R}\), reward function

\(o \in \Omega\), an observation from the set of observations

\(O(o_{t+1} | s_{t+1}, a_t)\) is the conditional observation probabilities

\(\gamma \in [0,1]\) is the discount factor
From MDP to POMDP: A Problem
How to act on all past information?
Option 1: Remember (RNN)

\[ h_{t-1} \]

\[ o_t \]

\[ a_t \]

\[ h_t \]
How to act on all past information?

Option 1: Remember (RNN)

- Generalization can be hard.
- No notion of stochasticity.
- Continuous cases are hard.
Option 2: Belief

\[ b_t = p_\theta (s_t | o_{\leq t}, a_{\leq t}) \]

Belief state
Option 2: Belief

$$b_t = p_\theta(s_t \mid o_{\leq t}, a_{\leq t})$$

Belief state

States
(Real world markovian states)

Measurements
(Actions & observations)

Belief State
(Estimate)
Option 2: Belief

\[ T = p_\theta (s_t | s_{t-1}, a_{t-1}) \quad \text{Transition Matrix} \]

\[ O = p_\theta (o_t | s_t, a_{t-1}) \quad \text{Observation Matrix} \]

\[ b_t = p_\theta (s_t | o_{\leq t}, a_{\leq t}) \quad \text{Belief state} \]

\[
\begin{align*}
  b_t(s_t) &= \frac{O(o_t | s_t, a_{t-1}) \sum_{s_{t-1} \in S} T(s_t | s_{t-1}, a_{t-1}) b(s_t)}{	ext{Normalization Factor}}
\end{align*}
\]
How to act on all past information?

Option 1: Remember (RNN)
- Generalization can be hard.
- No notion of stochasticity.
- Continuous cases are hard.

Option 2: Belief
- Computationally Expensive.
- Requires model.
- Provides stochasticity.
- Tends to generalize.
Not as clear

Model free
RNN (A)DRQN

Explicit Belief tracking
DVRL

Implicit Belief tracking
Next Session
Not as clear

Model free
RNN (A)DRQN

Explicit Belief tracking
DVRL

Implicit Belief tracking
Next Session
Deep Q-learning approaches for POMDPs

<table>
<thead>
<tr>
<th>Model</th>
<th>Input</th>
<th>Problem Addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN</td>
<td>$s_t$</td>
<td>model-free POMDP</td>
</tr>
<tr>
<td>DBQN</td>
<td>$b_t$</td>
<td>Model-based POMDP</td>
</tr>
<tr>
<td>DRQN</td>
<td>$&lt;o_1, o_2, ..., o_t&gt;$</td>
<td>Model-free POMDP</td>
</tr>
<tr>
<td>DDRQN</td>
<td>$&lt;a_0, a_1, ..., a_{t-1}&gt;$, $&lt;o_1, o_2, ..., o_t&gt;$</td>
<td>Model-free POMDP</td>
</tr>
<tr>
<td>ADRQN</td>
<td>$&lt;(a_0, o_1), (a_1, o_2), ..., (a_{t-1}, o_t)&gt;$</td>
<td>Model-free POMDP</td>
</tr>
</tbody>
</table>

(Zhu et al. 2017)
Deep Q-learning approaches for POMDPs

<table>
<thead>
<tr>
<th>Model</th>
<th>Input</th>
<th>Problem Addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN</td>
<td>$s_t$</td>
<td>model-free POMDP</td>
</tr>
<tr>
<td>DBQN</td>
<td>$b_t$</td>
<td>Model-based POMDP</td>
</tr>
<tr>
<td>DRQN</td>
<td>&lt;$o_1, o_2, ..., o_t&gt;$</td>
<td>Model-free POMDP</td>
</tr>
<tr>
<td>DDRQN</td>
<td>&lt;$a_0, a_1, ..., a_{t-1}$, $o_1, o_2, ..., o_t$&gt;</td>
<td>Model-free POMDP</td>
</tr>
<tr>
<td>ADRQN</td>
<td>&lt;$(a_0, o_1), (a_1, o_2), ..., (a_{t-1}, o_t)$&gt;</td>
<td>Model-free POMDP</td>
</tr>
</tbody>
</table>

(Zhu et al. 2017)
(Action-specific) Deep Recurrent Q-Learning: (A)DRQN

ADRQN

DRQN

\( h_{t-1} \)

\( a_t \)

\( o_t \)

\( a_{t-1} \)

\( o_{t+1} \)

\( h_t \)

Environment

(Hausknecht et al. 2015) (Zhu et al. 2017)
Flickering Frostbite and Pong
(A)DQRN: Results

Figure 2: Training results for Pong
Figure 3: Training results for Frostbite

(Zhu et al. 2017)
(A)DQRN: Results

Train on POMDP, test on MDP

Train on MDP, test on POMDP

(Zhu et al. 2017)
(A)DQRN: Critique

Model-free & Blackbox:
likely to summarize and not generalize
Model free

RNN

(A)DRQN

Explicit Belief tracking

DVRL

Implicit Belief tracking

Next Session
Deep Variational Reinforcement Learning (DVRL)

Particle Filter

Model

$O_{t-1}$

$a_t$

$V$

$\pi$

$L_{ELB}$

$L_{A2C}$

(Igl et al. 2018)
Deep Variational Reinforcement Learning (DVRL)

Model

Particle Filter

$O_{t-1}$

$a_t$

$\mathcal{L}_{\text{ELB}}$

$\mathcal{L}_{\text{A2C}}$

(Igl et al. 2018)
Deep Variational Reinforcement Learning (DVRL)

Model

Particle Filter

$O_{t-1}$

$a_t$

$\mathcal{L}_{ELB}$

$O$

$+$

$\mathcal{L}_{A2C}$

(Igl et al. 2018)
Deep Variational Reinforcement Learning (DVRL)
Deep Variational Reinforcement Learning (DVRL)

Model

Particle Filter

$O_{t-1}$

$a_t$

$V$

$\pi$

$L_{ELB}$

$L_{A2C}$

(Igl et al. 2018)
Brief note on notation

\( a_t = \text{action at time } t \)

\( o_t = \text{observation at time } t \)

\( k \text{ in } [1, K] = \text{number of particles} \)

\( b_t = (h_t, z_t, w_t) = \text{belief at time } t \)

\( z_t = \text{an additional stochastic latent state} \)

\( h_t = \text{latent state of a RNN (in a particle)} \)

\( w_t = \text{importance weight of a particle.} \)
Deep Variational Reinforcement Learning (DVRL)

Model

Particle Filter

$O_{t-1}$

$a_t$

$V$

$\pi$

$L_{\text{ELB}}$

$L_{\text{A2C}}$

(Igl et al. 2018)
DVRL: Particle Filter - Approximating $b_t$

Previous Belief

$$b_{t-1} = (h_{t-1}^k, z_{t-1}^k, w_{t-1}^k)_{k=1}^K$$

Sample new values

$$z_t \sim q_\psi(z_t|h_{t-1}, a_{t-1}, o_t)$$

$$h_t = \psi_\theta^{RNN}(z_t, h_{t-1}, a_{t-1}, o_t)$$

re-weight

$$w_i = \frac{p_\theta(z_i|h_{t-1}, a_{t-1}) p_\theta(o_t|h_{t-1}, z_i, a_{t-1})}{q_\psi(z_t|h_{t-1}, a_{t-1}, o_t)}$$

resample

$$b_t = (h_t^k, z_t^k, w_t^k)_{k=1}^K$$

(Igl et al. 2018)
Deep Variational Reinforcement Learning (DVRL)

\( \mathcal{L}_{\text{ELB}} \)

\( \mathcal{L}_{\text{A2C}} \)

\( V \)

\( \pi \)

\( o_t \)

\( a_{t-1} \)

(Igl et al. 2018)
DVRL: Policy - Summarize the particles

\[ b_t \]

\[ (h_t^1, z_t^1, w_t^1) \]
\[ (h_t^2, z_t^2, w_t^2) \]
\[ (h_t, z_t, w_t) \]
\[ (h_t^k, z_t^k, w_t^k) \]

Summary

\[ \hat{h}_t \]

Value & Policy

\[ V(\hat{h}_t) \]
\[ \pi(a_t|\hat{h}_t) \]

(Igl et al. 2018)
Deep Variational Reinforcement Learning (DVRL)

\( \mathcal{L}_{\text{ELB}} \)

\( \mathcal{O} \)

\( \mathcal{L}_{\text{A2C}} \)

(Igl et al. 2018)
DVRL: Model

\[ w_t = \frac{p_\theta (z_t | h_{t-1}, a_{t-1}) p_\theta (o_t | h_{t-1}, z_t, a_{t-1})}{q_\phi (z_t | h_{t-1}, a_{t-1}, o_t)} \]

\[ p_\theta (o_t | h_{t-1}, z_t, a_{t-1}) \]

(Igl et al. 2018)
DVRL: Model

\[ q_\phi(z_t|h_{t-1}, a_{t-1}, o_t) \quad p_\theta(o_t|h_{t-1}, z_t, a_{t-1}) \]

(Igl et al. 2018)
DVRL: Model

\[ q_\phi (z_t | h_{t-1}, a_{t-1}, o_t) \quad \text{Encoder} \]

\[ p_\theta (o_t | h_{t-1}, z_t, a_{t-1}) \quad \text{Decoder} \]
DVRL: Model

\[ w_t = \frac{p_{\theta} (z_t | h_{t-1}, a_{t-1}) p_{\theta} (o_t | h_{t-1}, z_t, a_{t-1})}{q_{\phi} (z_t | h_{t-1}, a_{t-1}, o_t)} \]

\[ ELBO(\theta, \phi) \approx \sum_{t=1}^{T} \log \left( \frac{1}{K} \sum_{k=1}^{K} w_t^k \right) \]

(Igl et al. 2018)
DVRL: Joint Learning

\[ \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{A2C}} \]

(Igl et al. 2018)
DVRL: Results - noisy MountainHike

(Igl et al. 2018)
DVRL: Results - noisy MountainHike

(Igl et al. 2018)
DVRL: Results - noisy MountainHike

(Igl et al. 2018)
ChopperCommand
Results: Ablation on Atari

(a) Influence of the particle number on performance for DVRL. Only using one particle is not sufficient to encode enough information in the latent state.

(b) Performance of the full DVRL algorithm compared to setting $\lambda^E = 0$ ("No ELBO") or not backpropagating the policy gradients through the encoder ("No joint optim").
DVRL: Critique

The belief state is still a rough approximation.

Is this really the best way to learn it?
Summary

- Extended MDP to POMDP
- (A)DRQN
- DVRL
Discussion

In a POMDP we still assume full access to the reward.

1) This not a realistic case (our perception of the reward depends as much on our observations as the state)

2) If it is realistic, our belief should be updated based on the reward.
Model free

RNN

(A)DRQN

Explicit Belief tracking

DVRL

DPFRL

Implicit Belief tracking

VRM
