Partial Observability in DRL

Part 1: POMDPs, (A)DRQN & DVRL

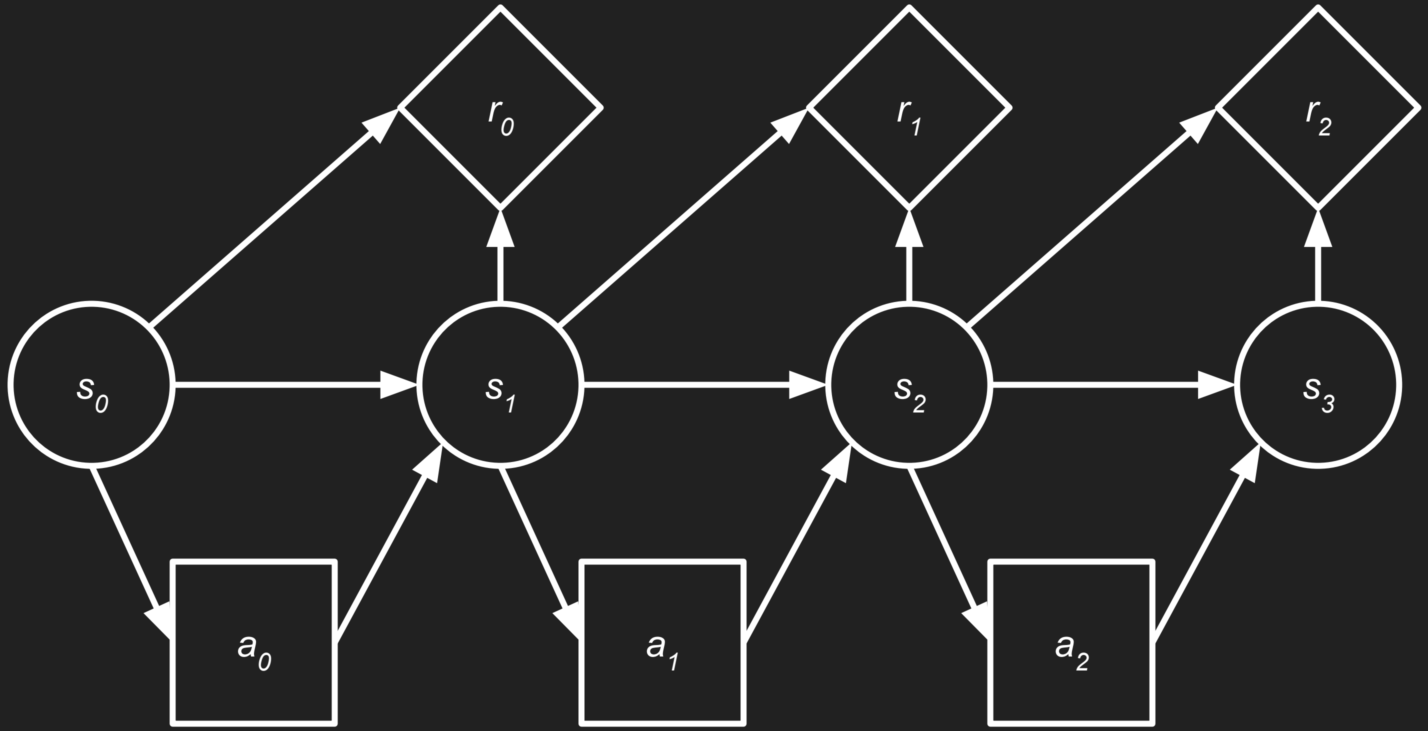
Most of the World is only Partial Observable

Occlusions

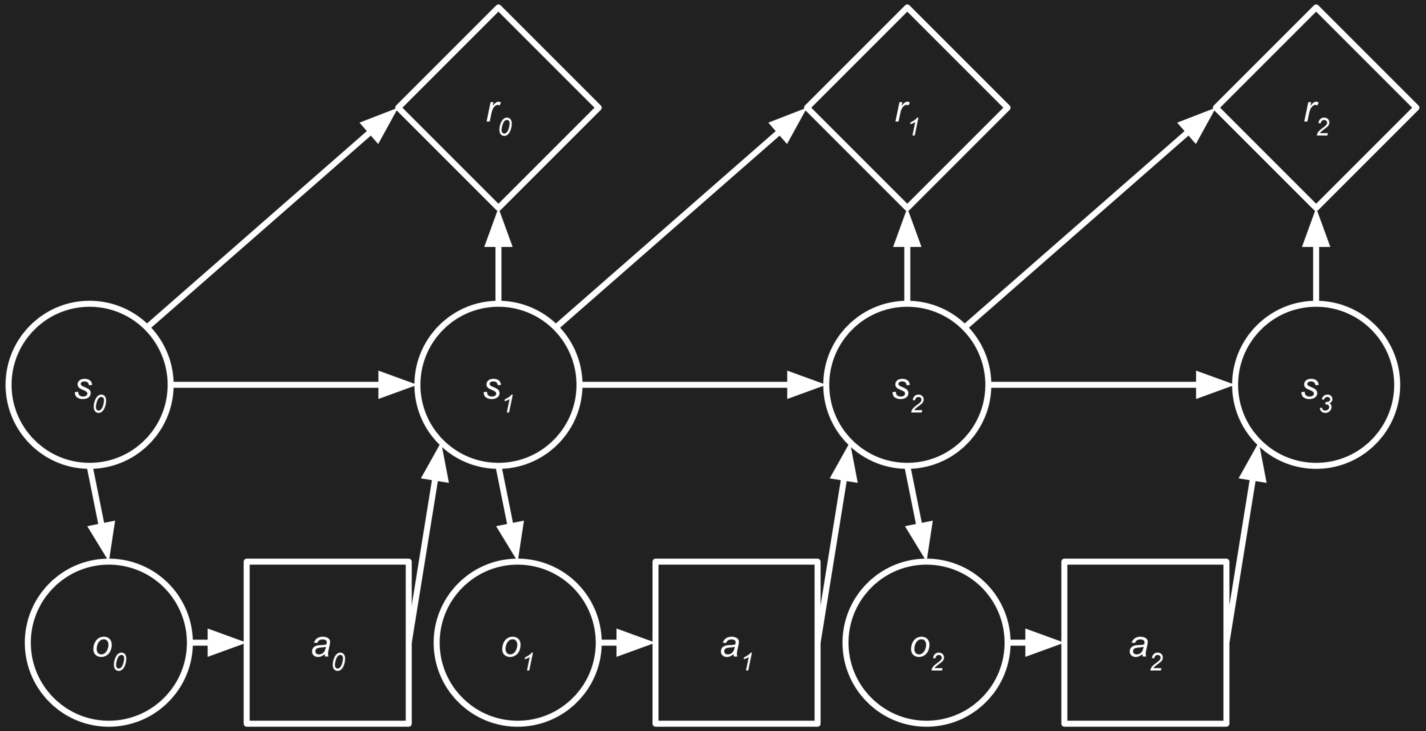
Latent Causes

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

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

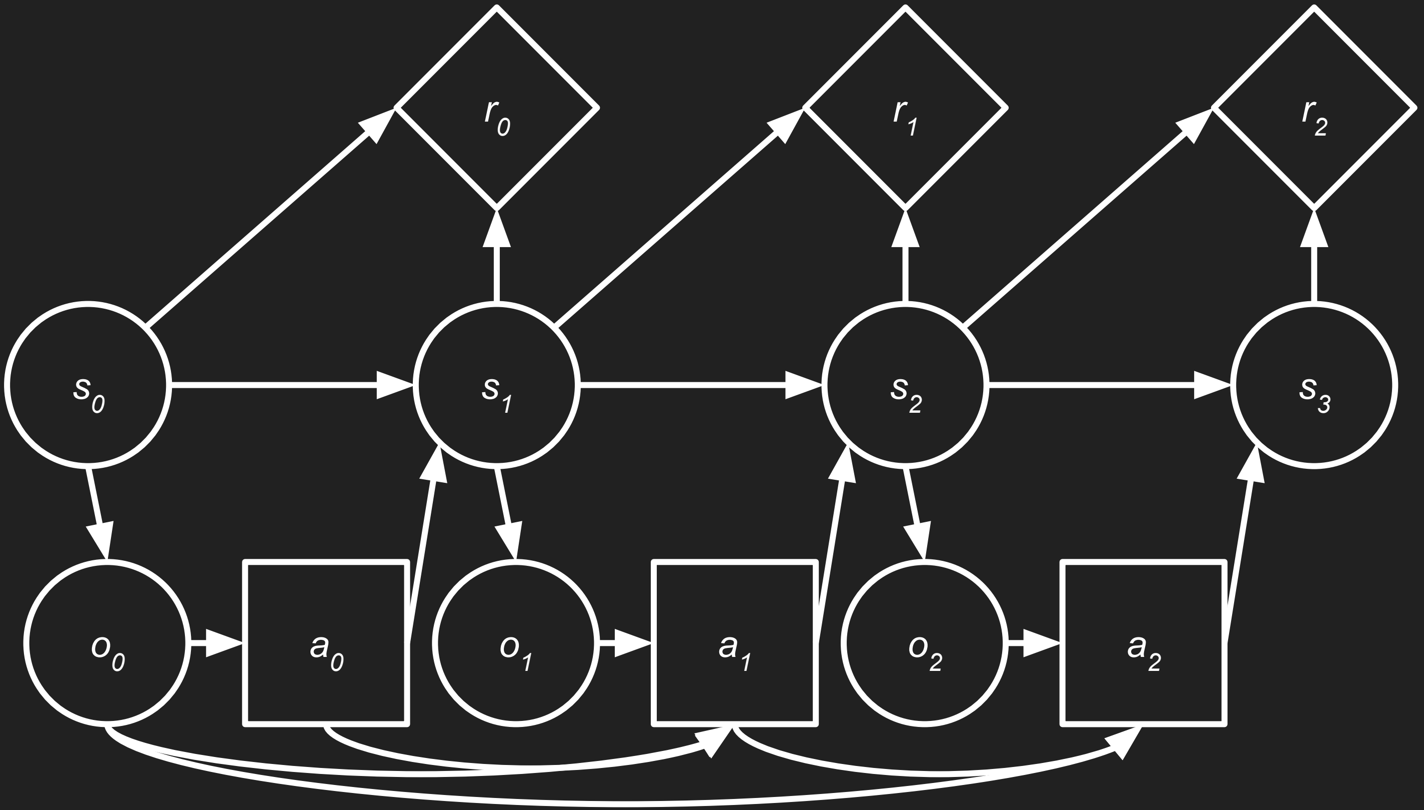
$R: S \times A \rightarrow \mathbb{R}$, reward function

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

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

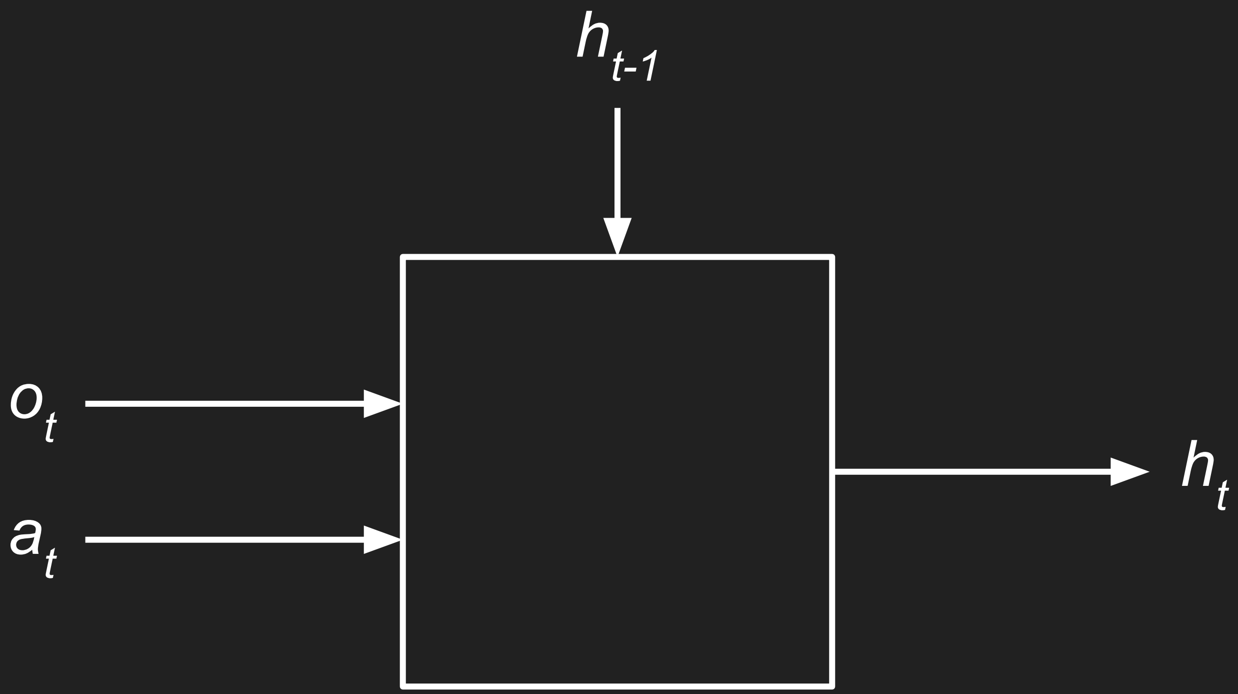
$\gamma \in [0,1]$ is the discount factor

From MDP to POMDP: A Problem



How to act on *all* past information?

Option 1: Remember (RNN)



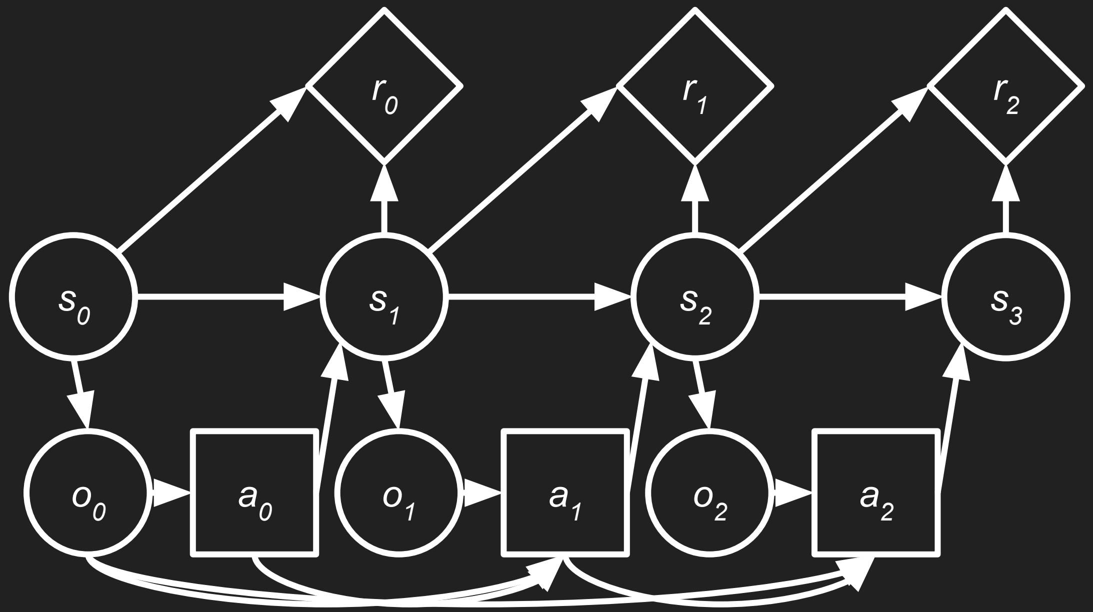
How to act on all past information?

Option 1: Remember (RNN)

- Generalization can be hard.
- No notion of stochasticity.
- Continuous cases are hard.

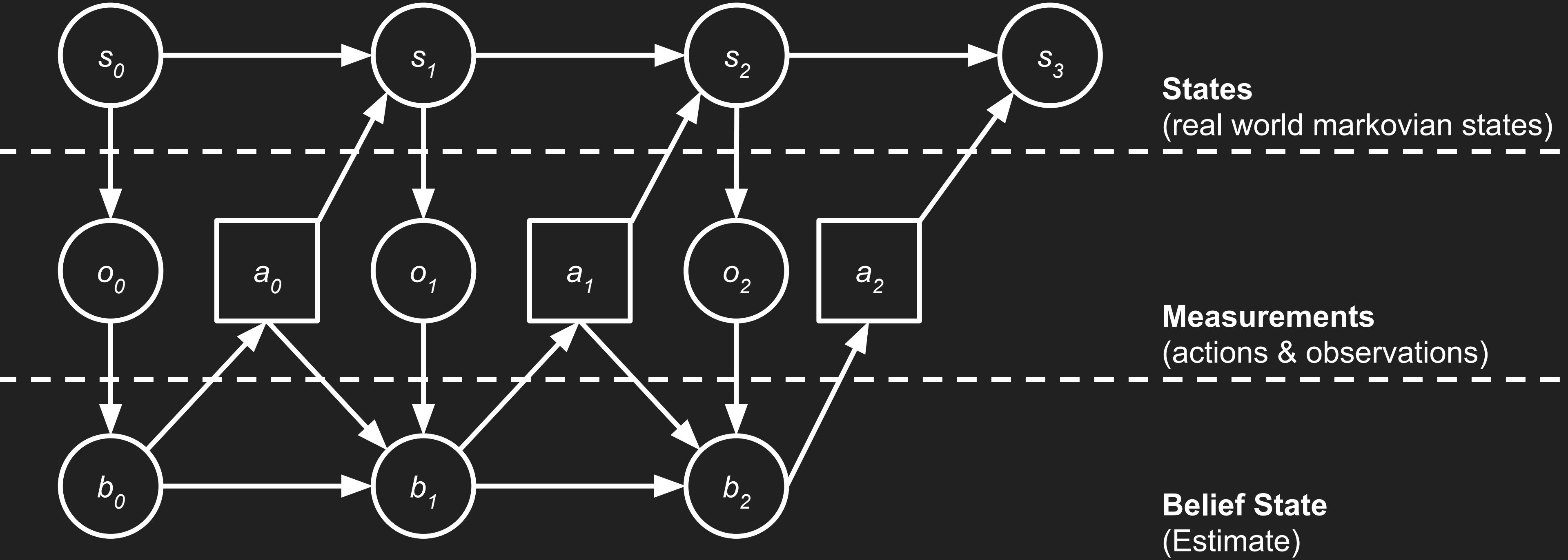
Option 2: Belief

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



Option 2: Belief

$b_t \coloneqq p_\theta\,(s_t \mid o_{\le t}, a_{\le t})$ Belief state



Option 2: Belief

$T = p_\theta(s_t | s_{t-1}, a_{t-1})$ Transition Matrix

$O = p_\theta(o_t, | s_t, a_{t-1})$ Observation Matrix

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

$$b_t(s_t) = \frac{O(o_t|s_t, a_{t-1}) \sum_{s_{t-1} \in S} T(s_t|s_{t-1}, a_{t-1}) b(s_t)}{\text{Normalization Factor}}$$

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

Model	Input	Problem Addressed
DQN	s_t	model-free POMDP
DBQN	b_t	Model-based POMDP
DRQN	$<o_1,o_2,\ldots,o_t>$	Model-free POMDP
DDRQN	$<a_0,a_1,\ldots,a_{t-1}>$ $<o_1,o_2,\ldots,o_t>$	Model-free POMDP
ADRQN	$<(a_0,o_1),(a_1,o_2),\ldots,(a_{t-1},o_t)>$	Model-free POMDP

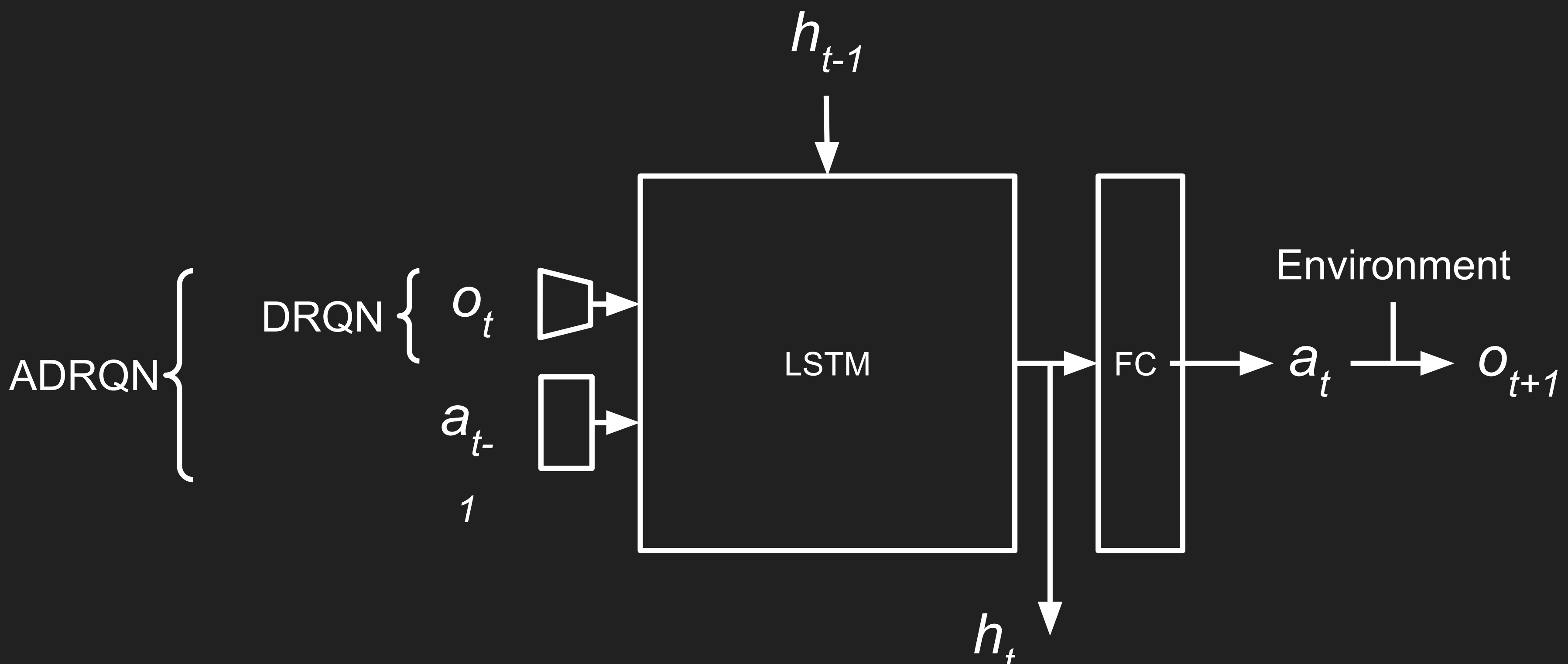
(Zhu et al. 2017)

Deep Q-learning approaches for POMDPs

Model	**Input**	**Problem Addressed**
DQN	s_t	model-free POMDP
DBQN	b_t	Model-based POMDP
DRQN	$<o_1,o_2,\ldots,o_t>$	Model-free POMDP
DDRQN	$<a_0,a_1,\ldots,a_{t-1}>$ $<o_1,o_2,\ldots,o_t>$	Model-free POMDP
ADRQN	$<(a_0,o_1),(a_1,o_2),\ldots,(a_{t-1},o_t)>$	Model-free POMDP

(Zhu et al. 2017)

(Action-specific) Deep Recurrent Q-Learning: (A)DRQN



(Hausknecht et al. 2015) (Zhu et al. 2017)

Flickering Frostbite and Pong

(A)DQRN: Results

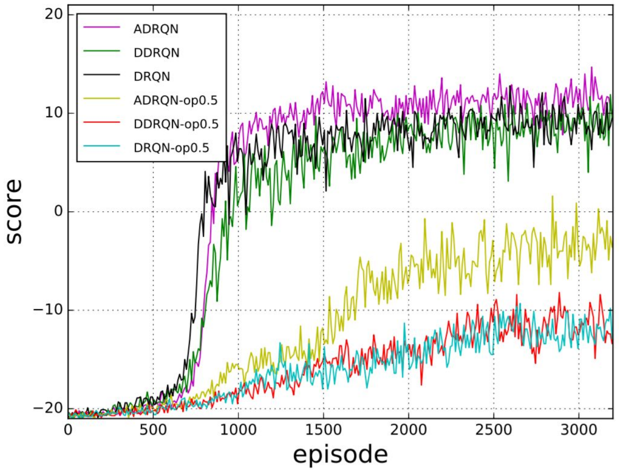


Figure 2: Training results for Pong

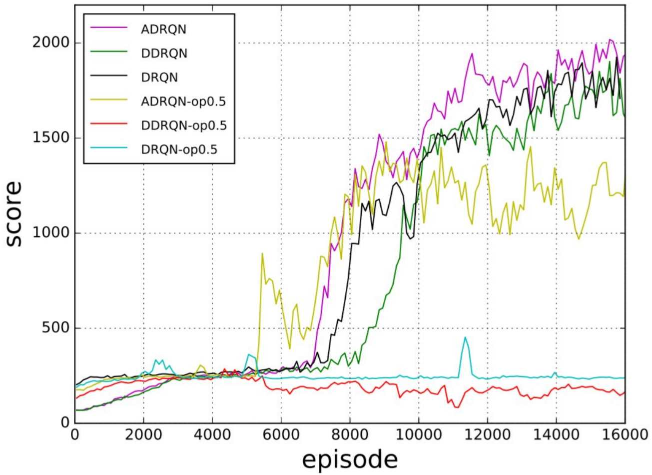
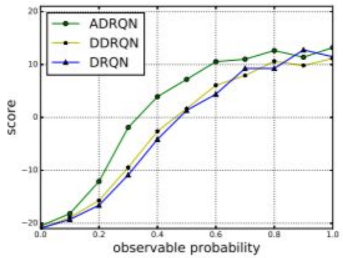


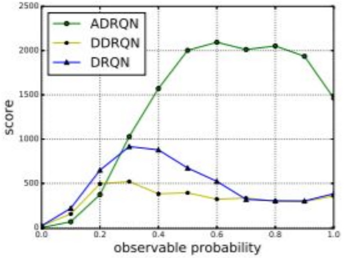
Figure 3: Training results for Frostbite

(Zhu et al. 2017)

(A)DQRN: Results

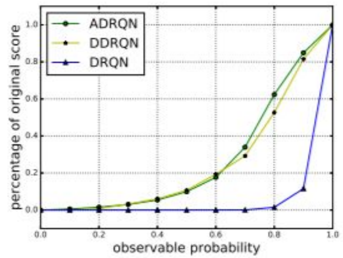


(a) Game Pong

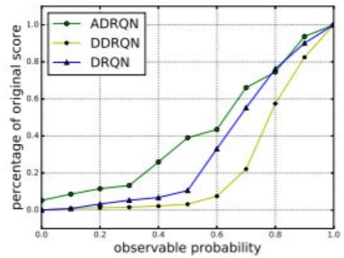


(b) Game Frostbite

Train on POMDP, test on MDP



(a) Game Pong



(b) Game Frostbite

Train on MDP, test on POMDP

(Zhu et al. 2017)

(A)DQRN: Critique

Model-free & Blackbox:

likely to summarize and not generalize

Next

Model free

RNN
(A)DRQN

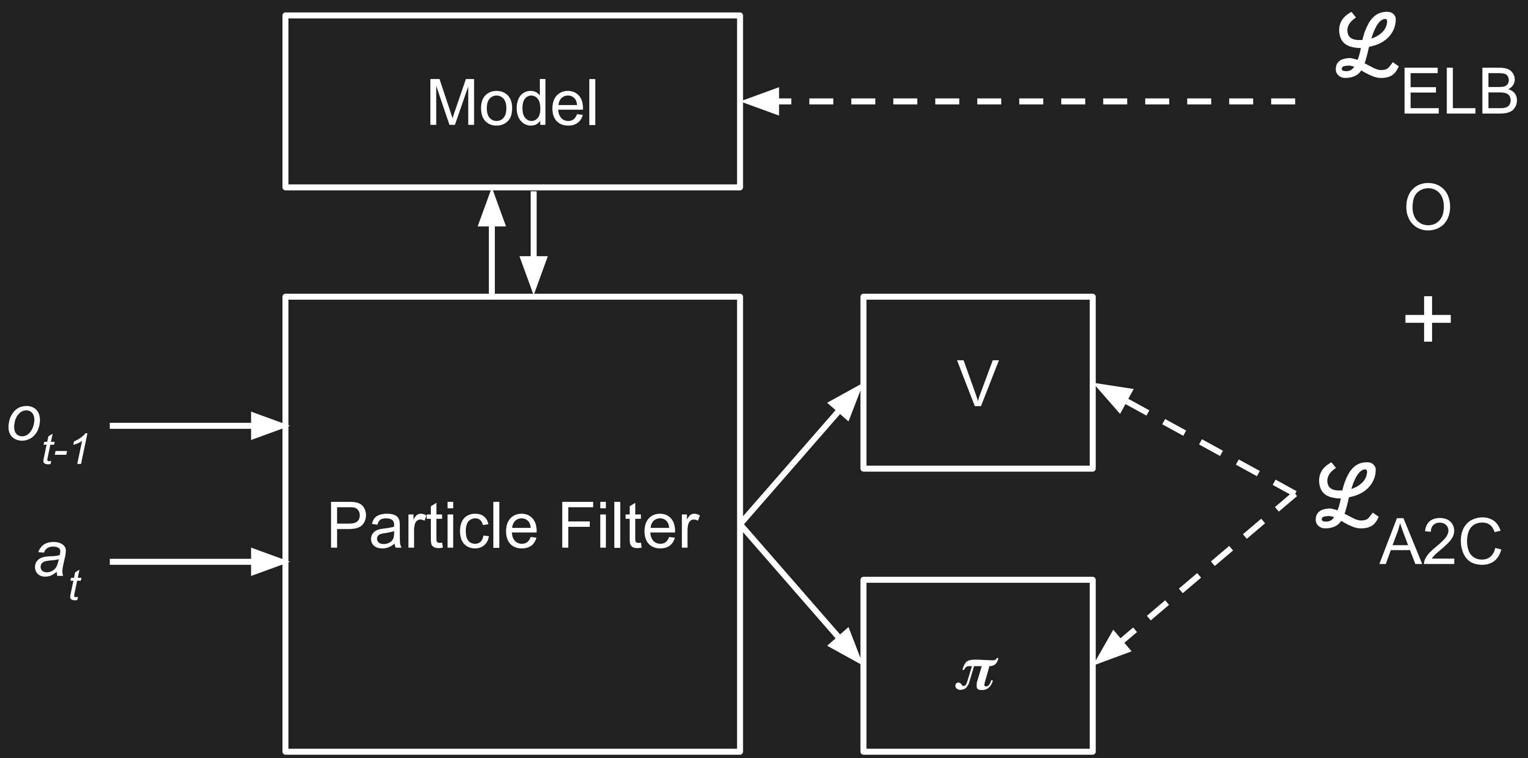
Explicit
Belief tracking

DVRL

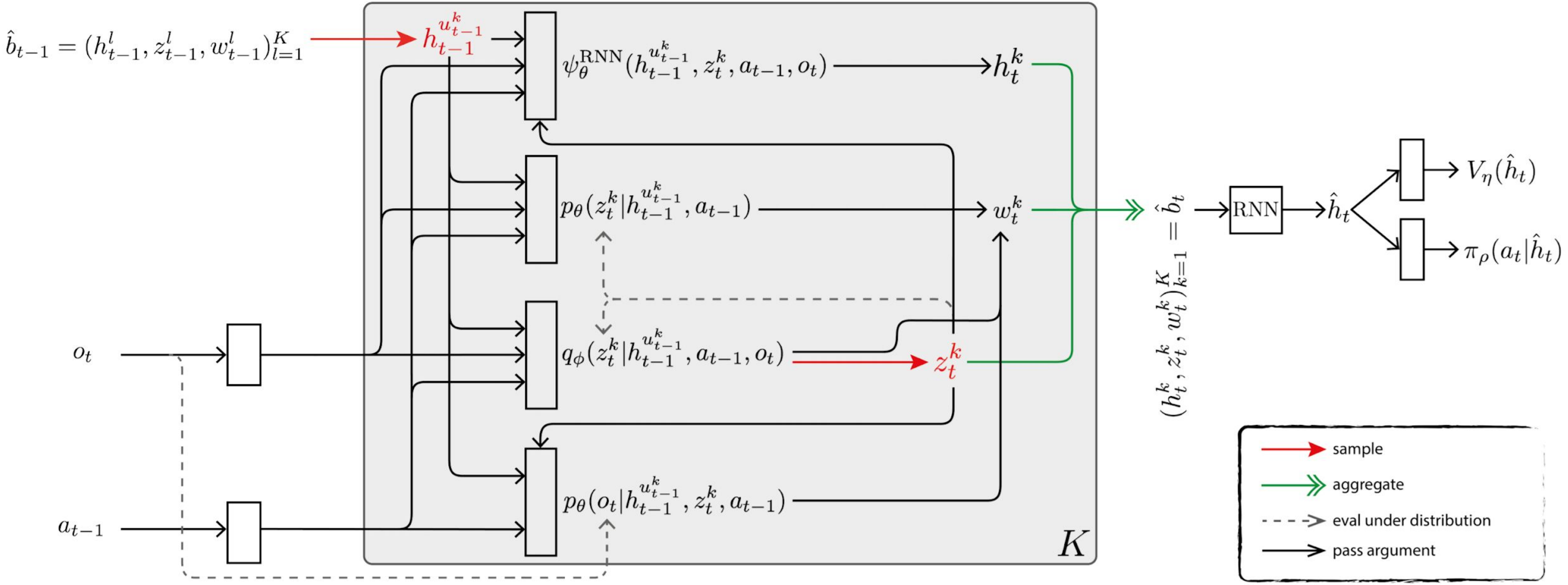
Implicit
Belief tracking

Next Session

Deep Variational Reinforcement Learning (DVRL)

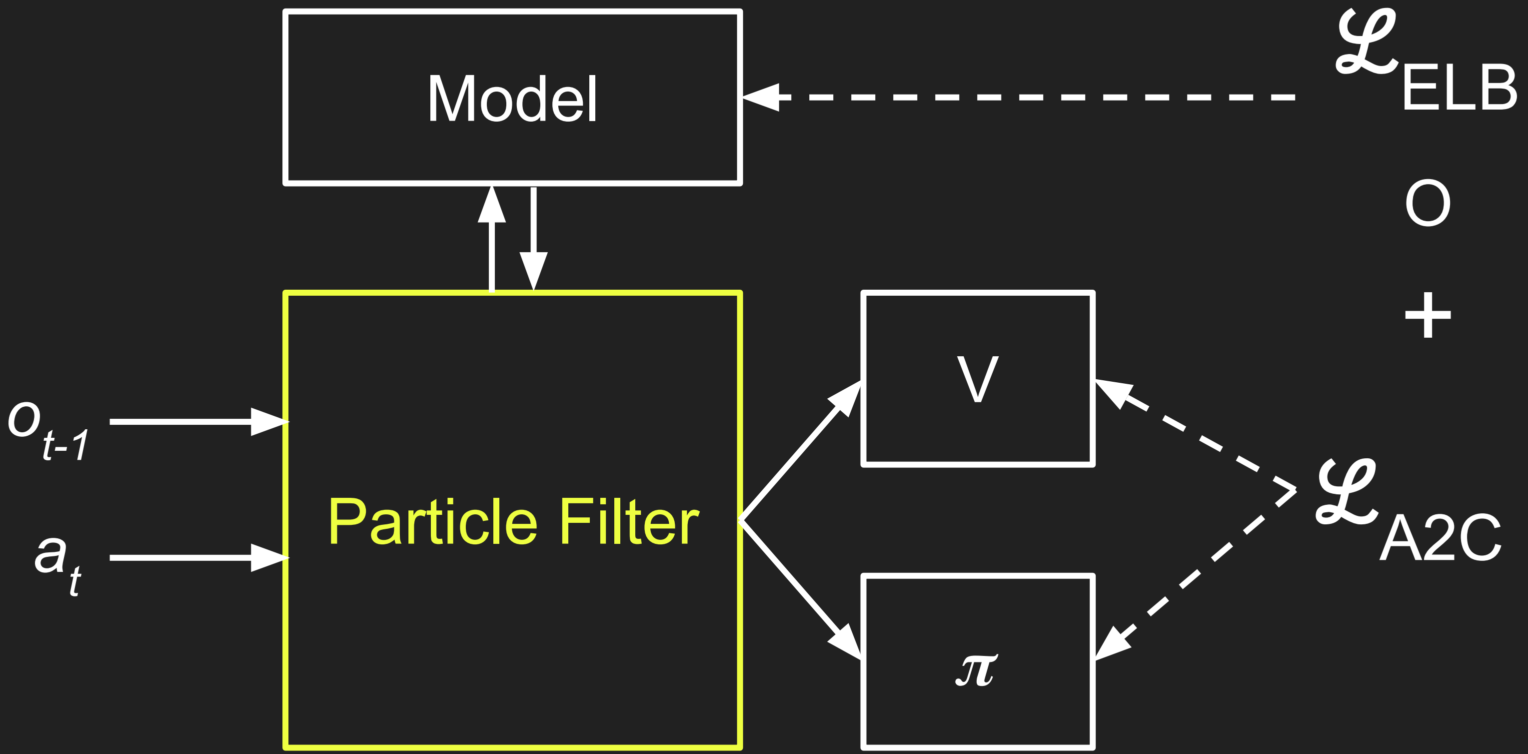


(Igl et al. 2018)



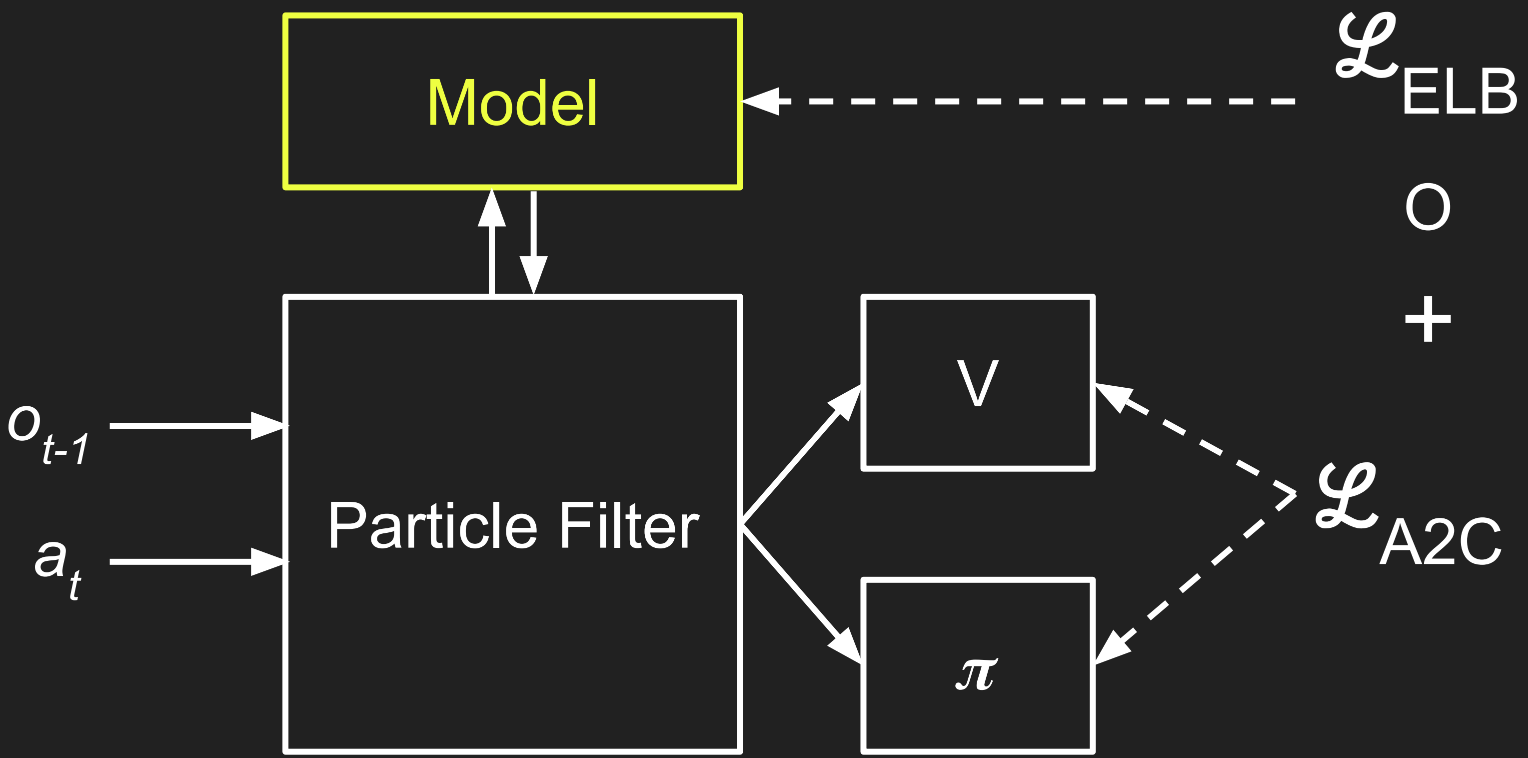
(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



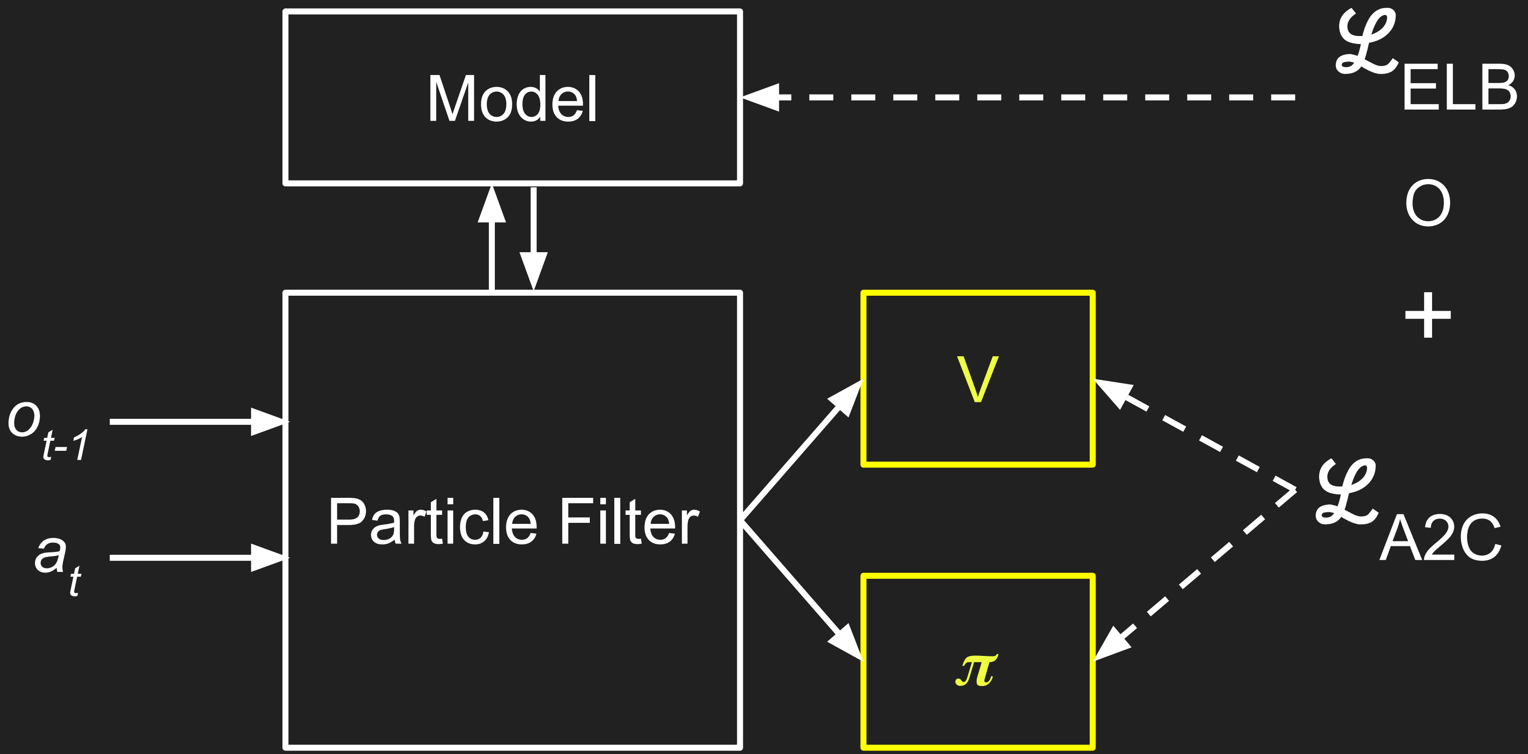
(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



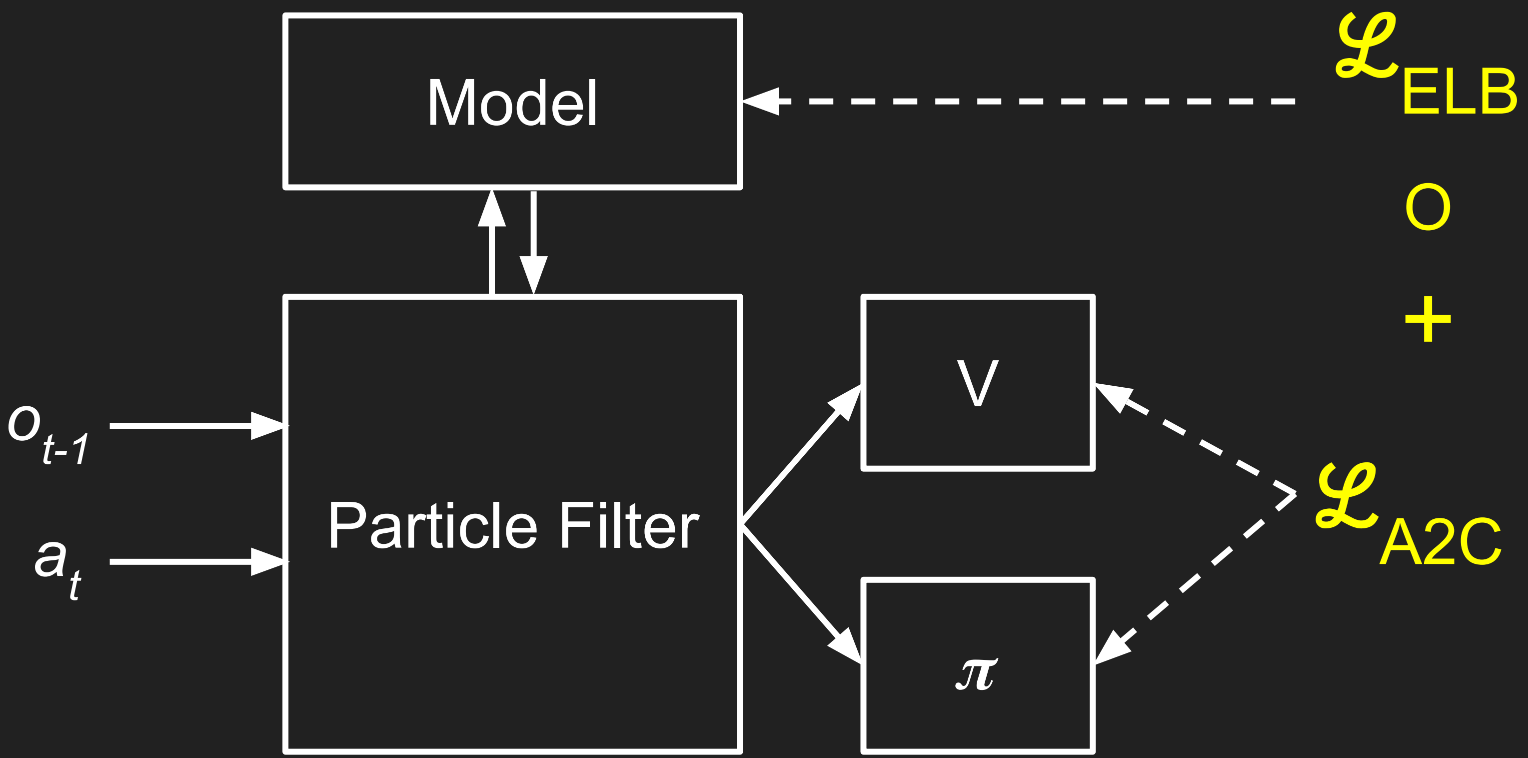
(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



(Igl et al. 2018)

Brief note on notation

a_t = *action at time t*

o_t = *observation at time t*

k in [1,K] = number of particles

$b_t = (h_t, z_t, w_t)$ *belief at time t*

z_t = *an additional stochastic latent state*

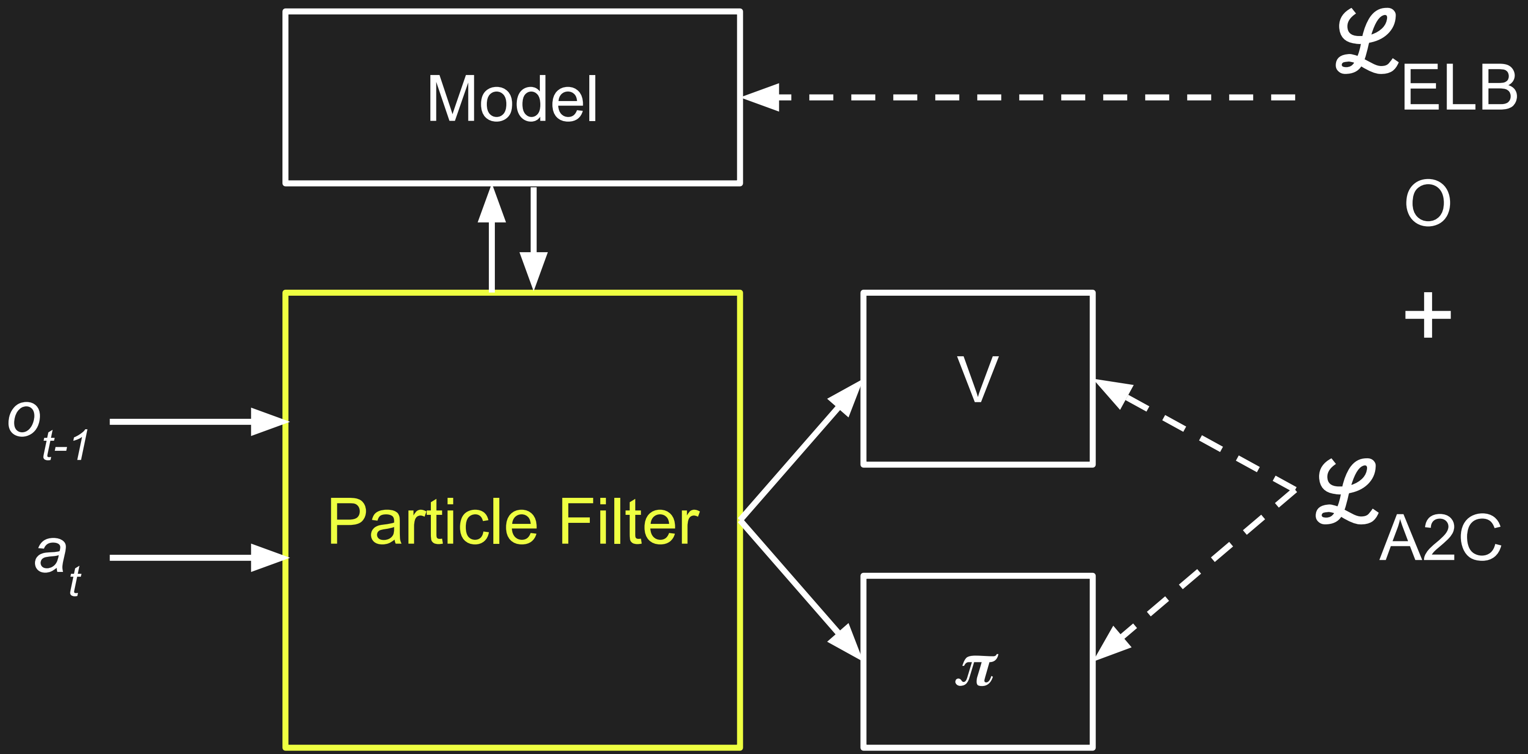
h_t = *latent state of a RNN (in a particle)*

Latent Summary of state

w_t = *importance weight of a particle.*

Likelihood of that latent state

Deep Variational Reinforcement Learning (DVRL)



(Igl et al. 2018)

DVRL: Particle Filter - Approximating b_t

Previous Belief

$$b_{t-1} = \left(h^k_{t-1}, z^k_{t-1}, w^k_{t-1}\right)^K_{k=1}$$

Sample new values

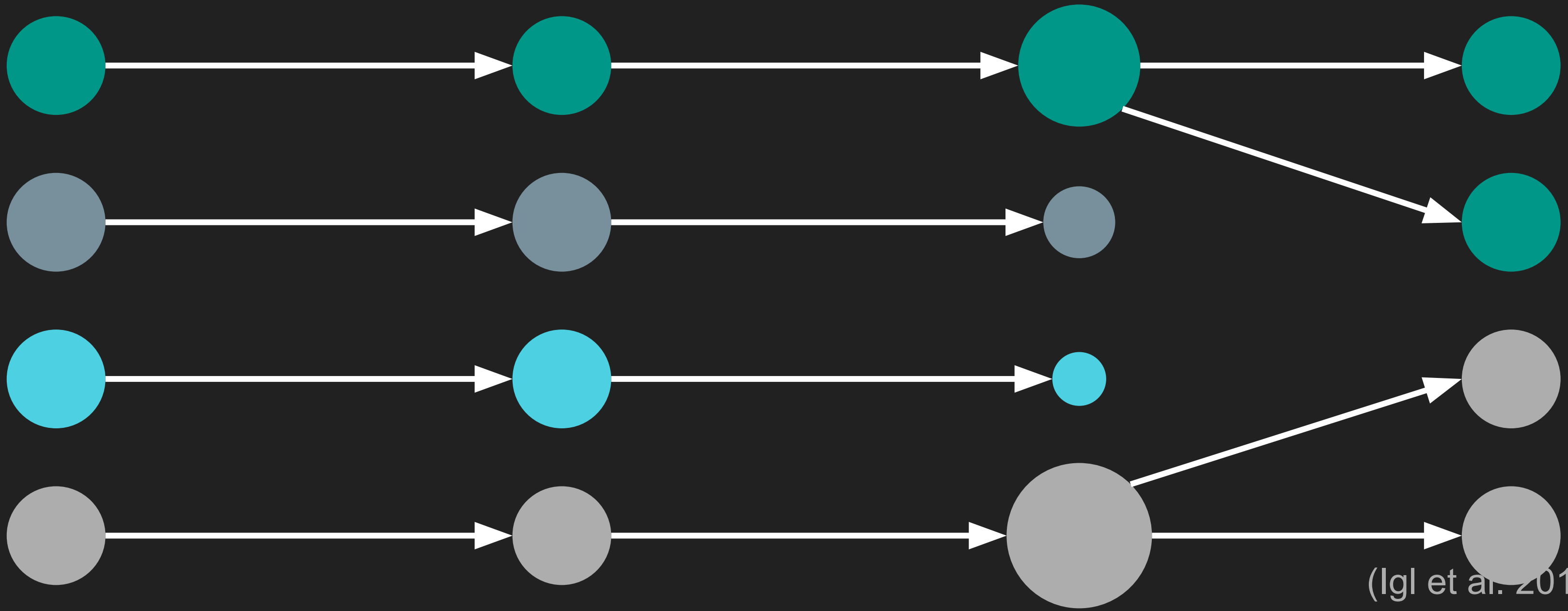
$$z_t \sim q_\phi(z_t | h_{t-1}, a_{t-1}, o_t)$$
$$h_t = \psi^{RNN}_\theta(z_t, h_{t-1}, a_{t-1}, o_t)$$

re-weight

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

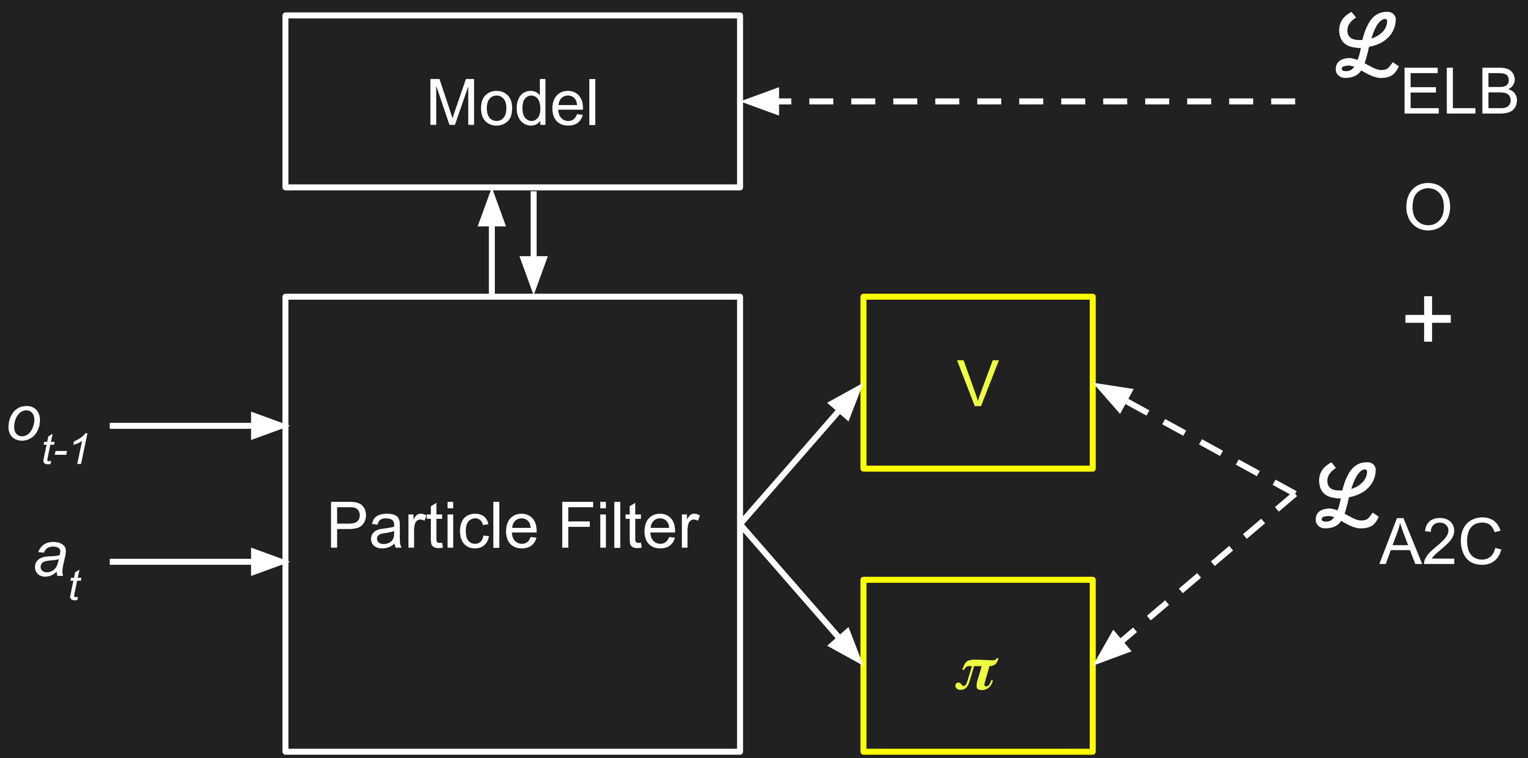
resample

$$b_t = \left(h^k_t, z^k_t, w^k_t\right)^K_{k=1}$$



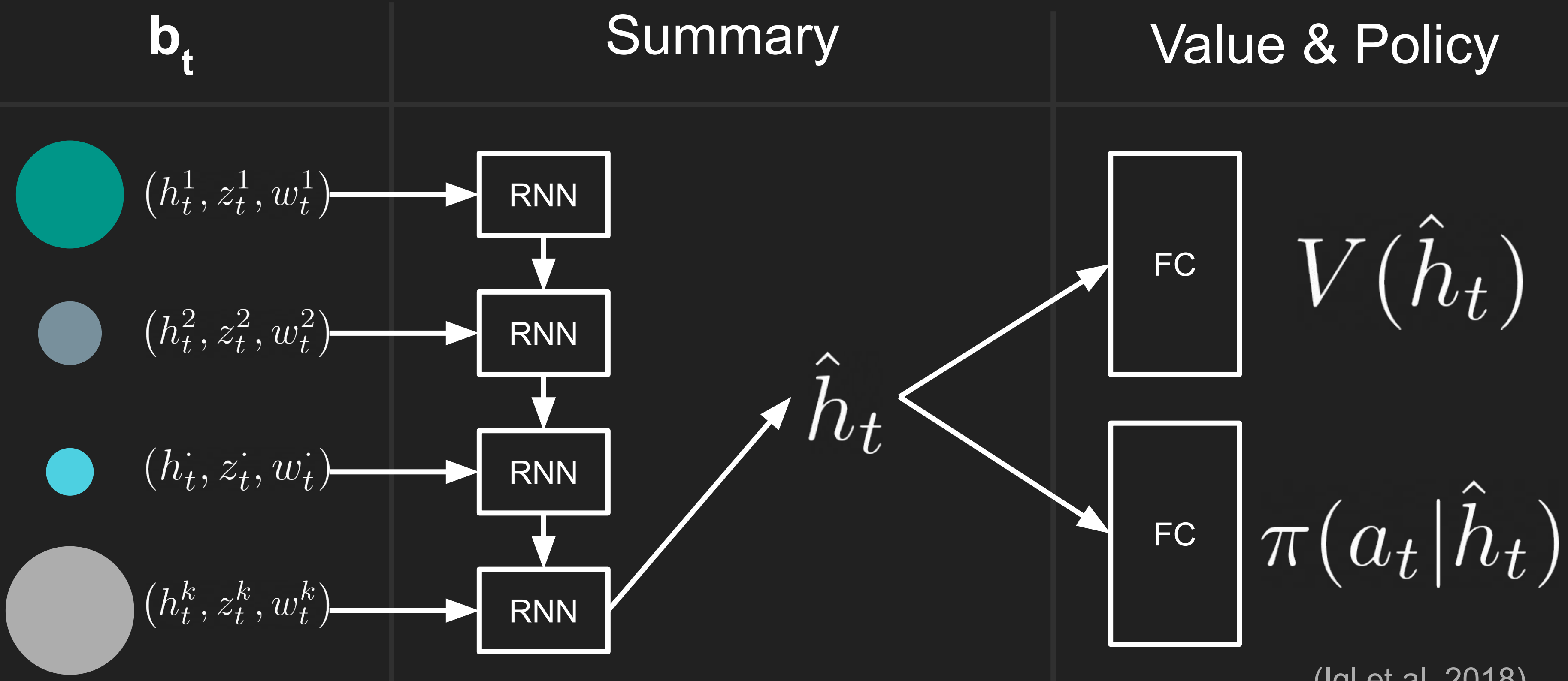
(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



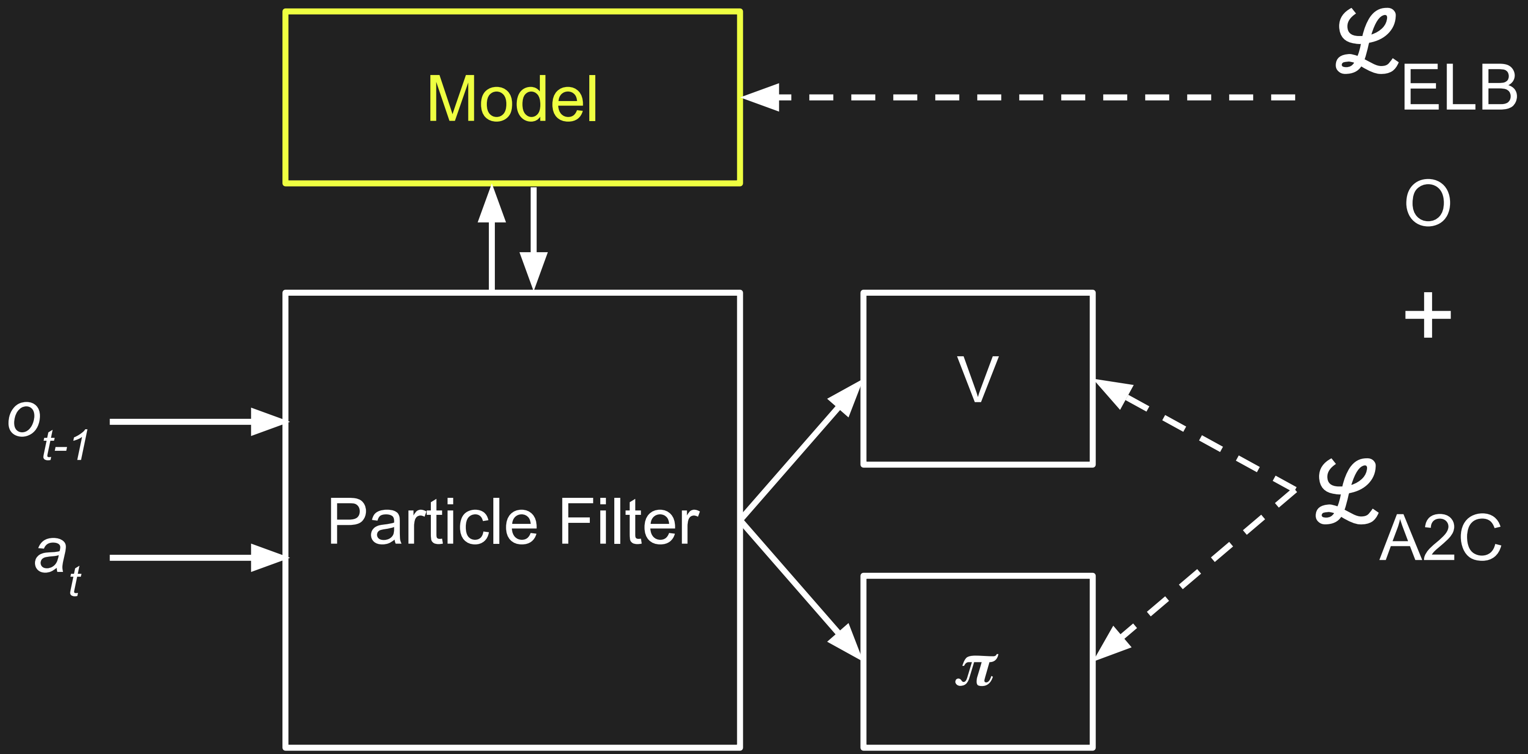
(Igl et al. 2018)

DVRL: Policy - Summarize the particles



(Igl et al. 2018)

Deep Variational Reinforcement Learning (DVRL)



(Igl et al. 2018)

DVRL: Model

$$w_t = \frac{p_\theta\left(z_t \middle| h_{t-1}, a_{t-1}\right) \boxed{p_\theta\left(o_t \middle| h_{t-1}, z_t, a_{t-1}\right)}}{q_\phi\left(z_t \middle| h_{t-1}, a_{t-1}, o_t\right)}$$

$$p_\theta\left(o_t \middle| h_{t-1}, z_t, a_{t-1}\right)$$

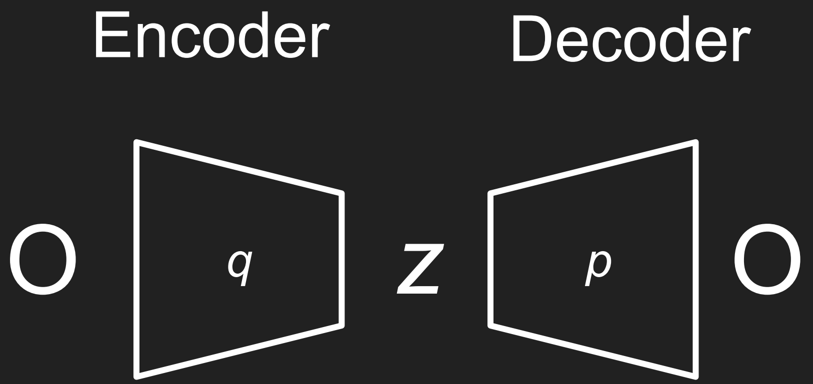
(Igl et al. 2018)

DVRL: Model

$$q_\phi\left(z_t \middle| h_{t-1}, a_{t-1}, o_t\right) \qquad p_\theta(o_t | h_{t-1}, z_t, a_{t-1})$$

(Igl et al. 2018)

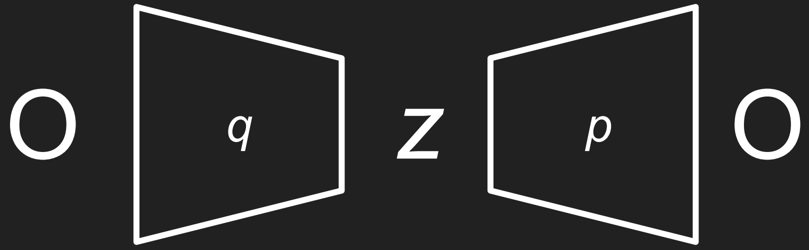
DVRL: Model



$$q_\phi\left(z_t | h_{t-1}, a_{t-1}, o_t\right) \qquad p_\theta(o_t | h_{t-1}, z_t, a_{t-1})$$

(Igl et al. 2018)

DVRL: Model



Prior

Decoder

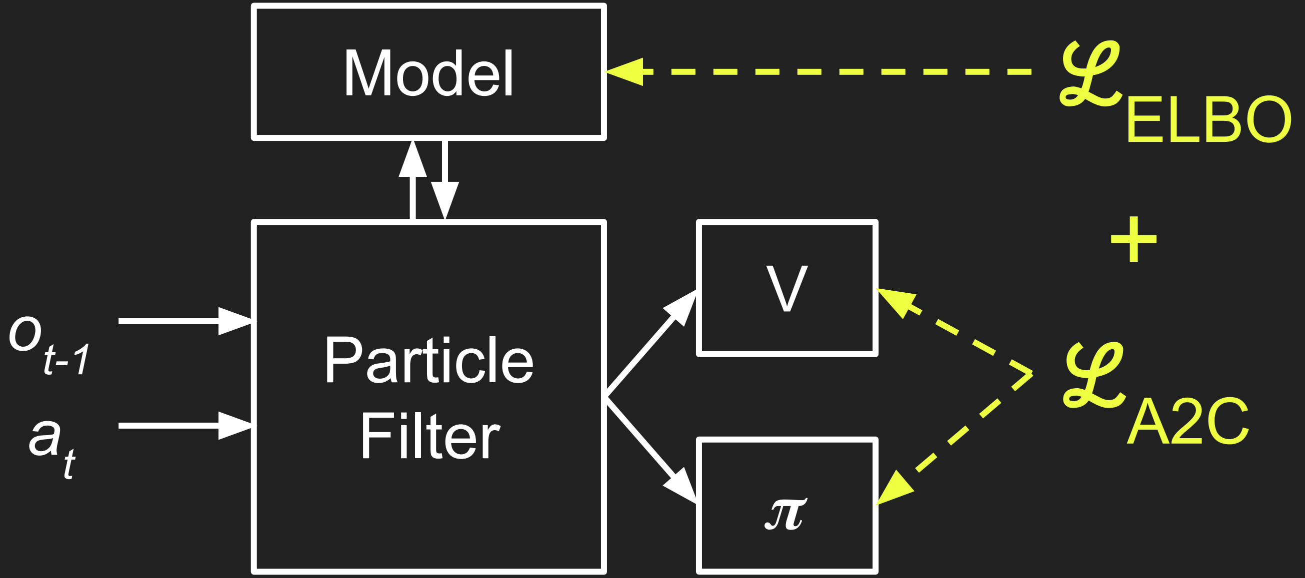
$$w_t = \frac{p_\theta\left(z_t | h_{t-1}, a_{t-1}\right) p_\theta\left(o_t | h_{t-1}, z_t, a_{t-1}\right)}{q_\phi\left(z_t | h_{t-1}, a_{t-1}, o_t\right)}$$

Encoder

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

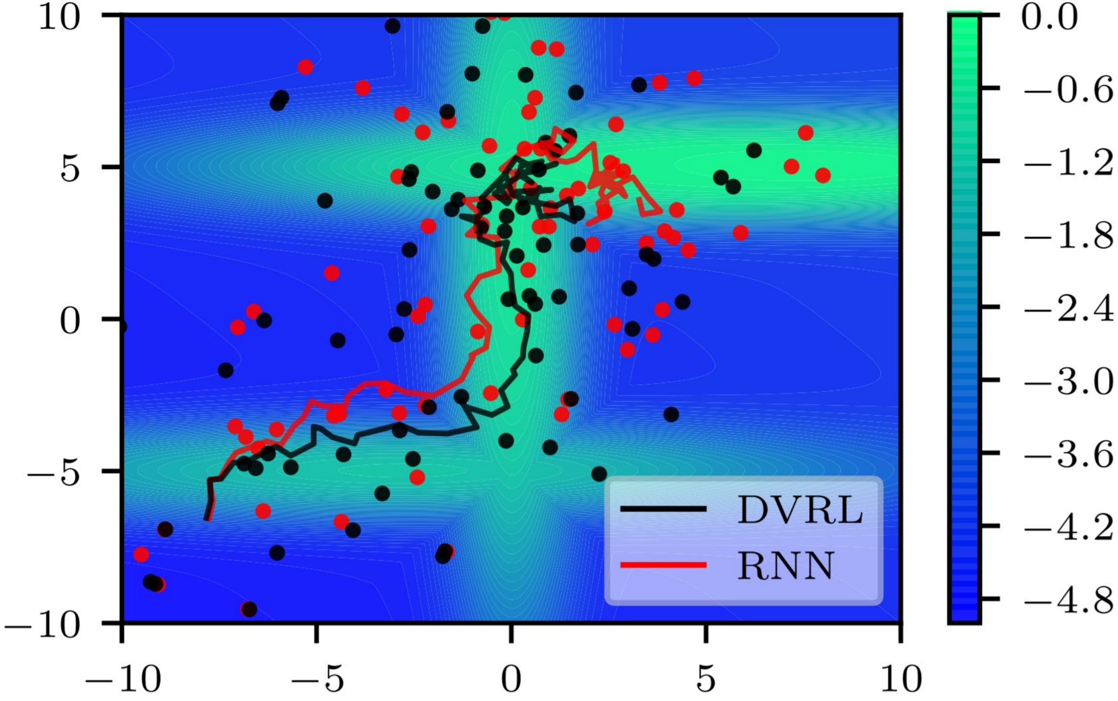
(Igl et al. 2018)

DVRL: Joint Learning



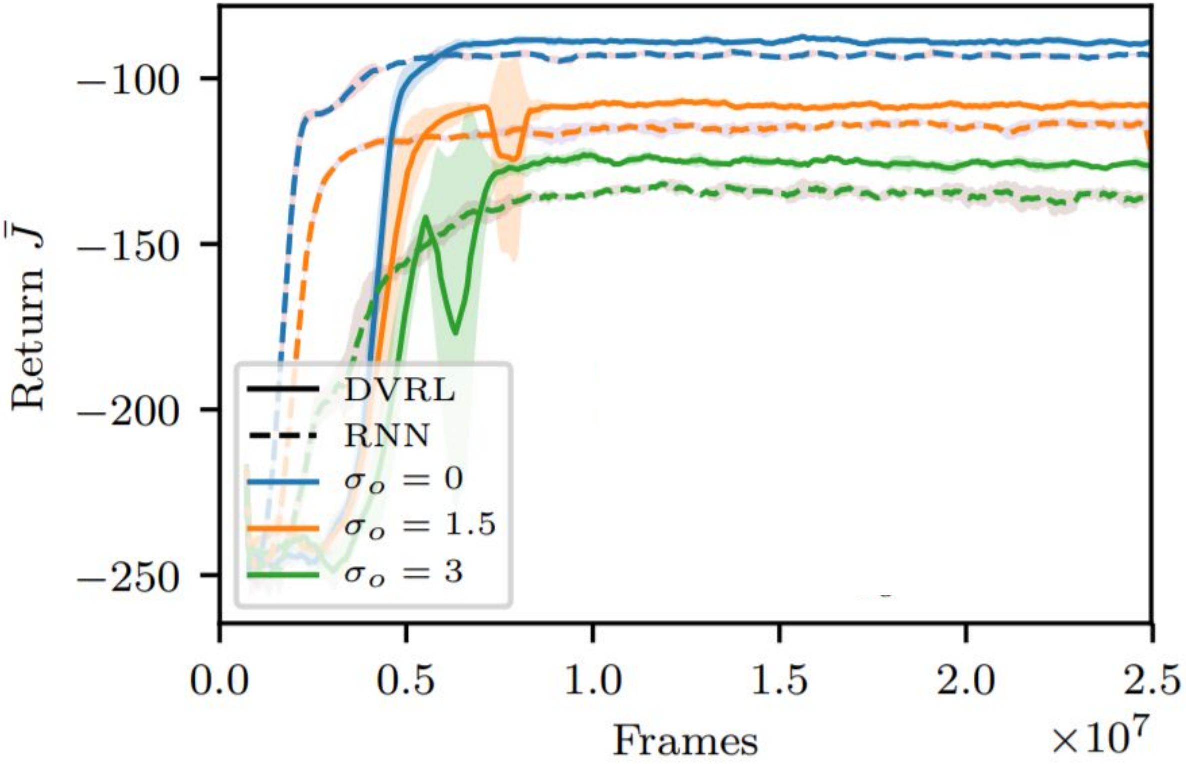
(Igl et al. 2018)

DVRL: Results - noisy MountainHike



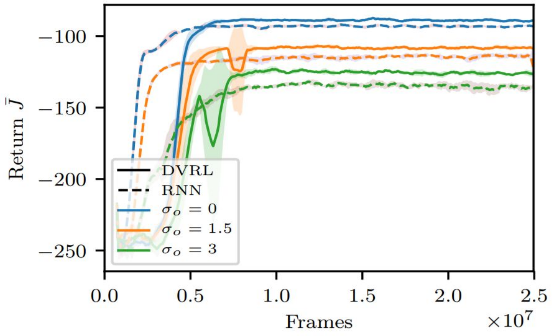
(Igl et al. 2018)

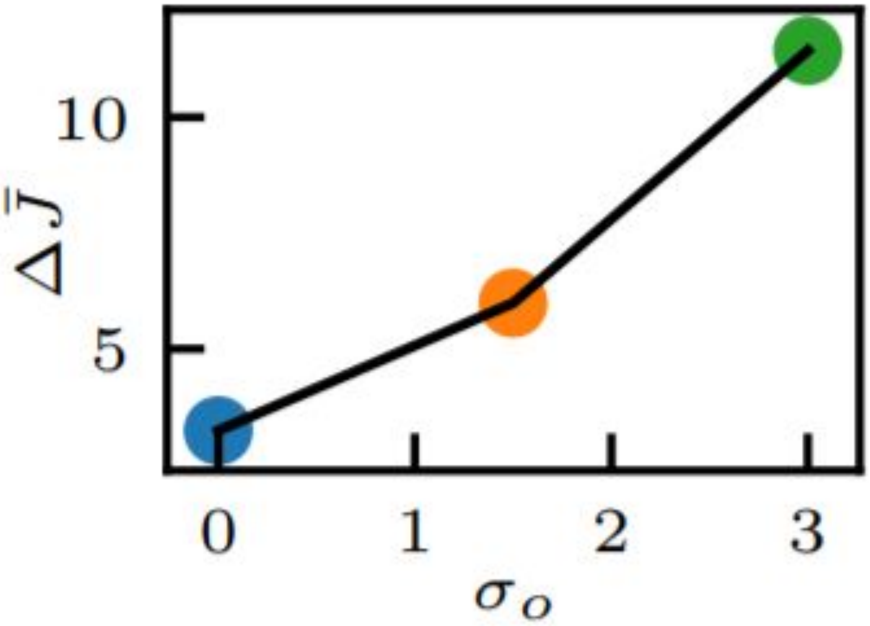
DVRL: Results - noisy MountainHike



(Igl et al. 2018)

DVRL: Results - noisy MountainHike

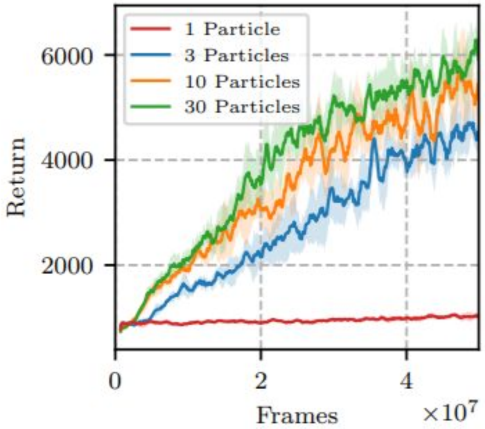




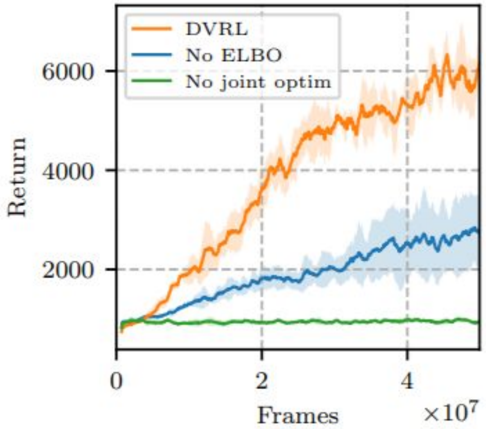
(Igl et al. 2018)

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").

(Igl et al. 2018)

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.

Next

Model free

RNN
(A)DRQN

Explicit Belief tracking

DVRL
DPFRL

Implicit Belief tracking

VRM

References

Hausknecht, M., & Stone, P. (2015, September). *Deep recurrent q-learning for partially observable mdps*. In 2015 AAAI Fall Symposium Series.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., & Whiteson, S. (2018). *Deep variational reinforcement learning for pomdps*. arXiv preprint arXiv:1806.02426.

Zhu, P., Li, X., Poupart, P., & Miao, G. (2017). *On improving deep reinforcement learning for pomdps.* arXiv preprint arXiv:1704.07978.