Generalizable Episodic Memory for Deep Reinforcement Learning

Hao Hu · MIG · Tsinghua

Advisor： Chongjie Zhang

7/8/2021

Machine Intelligence Group

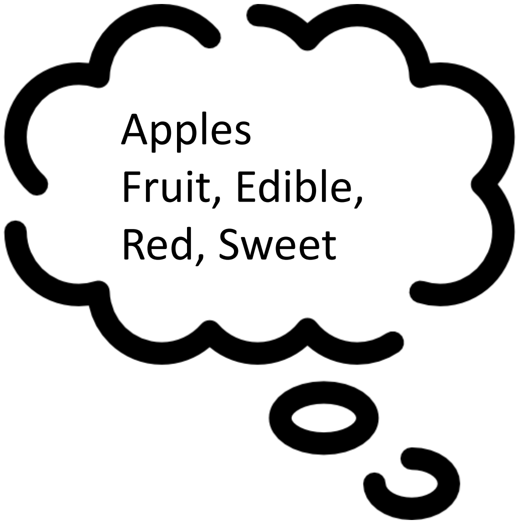
清華大學 交叉信息研究院

Tsinghua University Institute for Interdisciplinary Information Sciences

Semantic Memory v.s. Episodic Memory

Semantic Memory

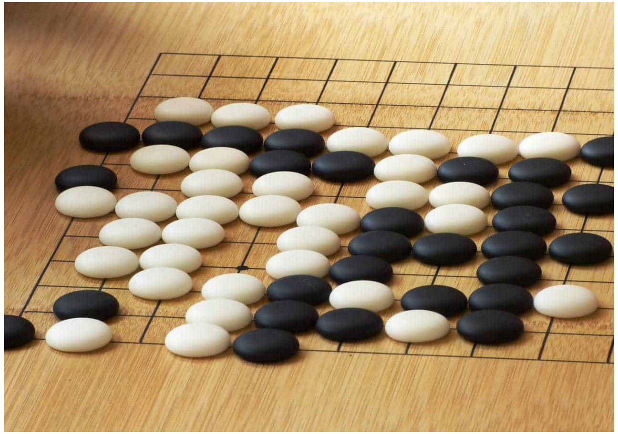
Episodic Memory

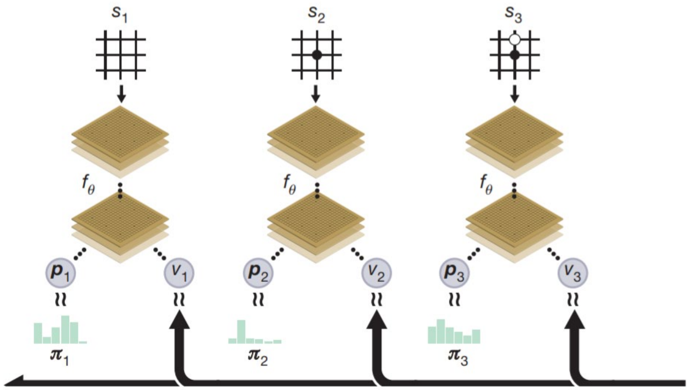


Object Knowledge learned over many interactions

Memories for specific events you have experienced

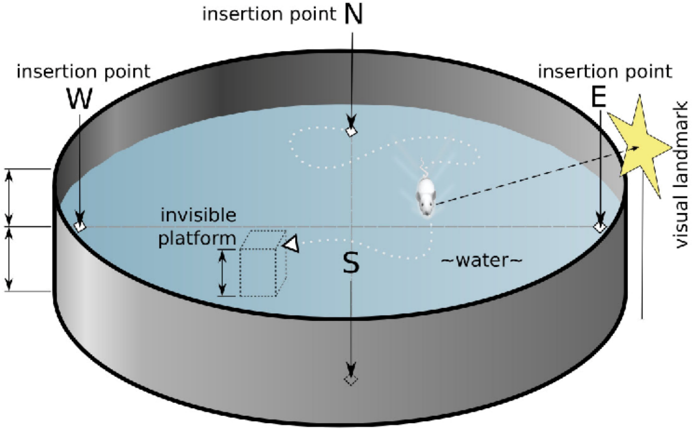
Fast Learning v.s. Slow Learning



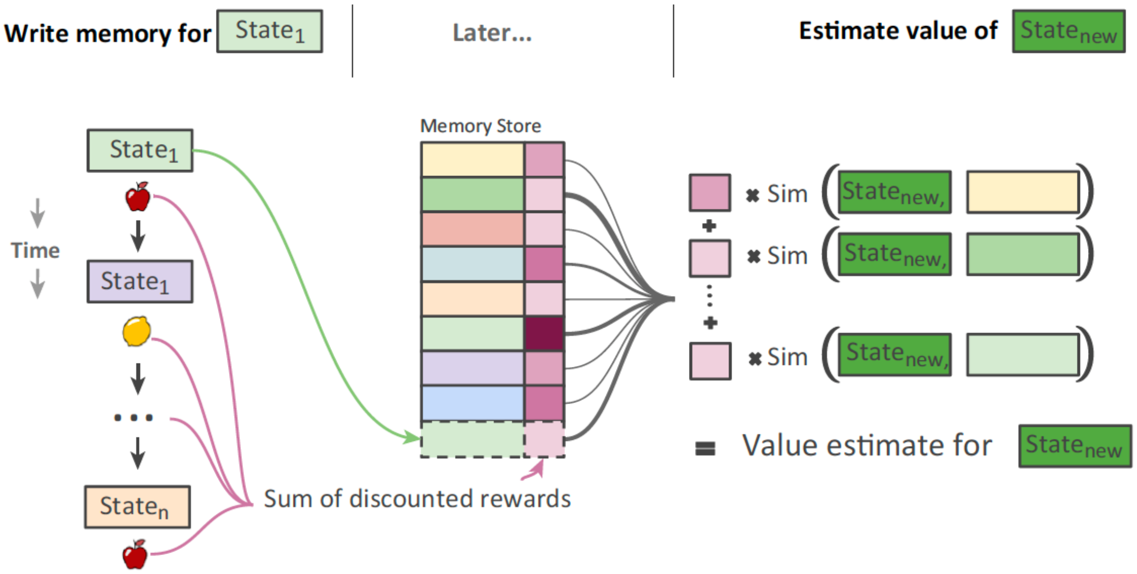


AlphaGo

Fast Learning v.s. Slow Learning



Episodic Control



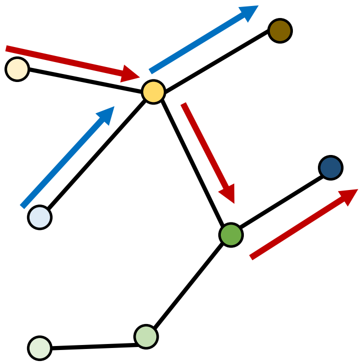
[Blundell, C. et al. Model-free episodic control. 2016]
[Botvinick et al, “Reinforcement Learning, Fast and Slow”, 2019]

Deep RL v.s. Episodic Control

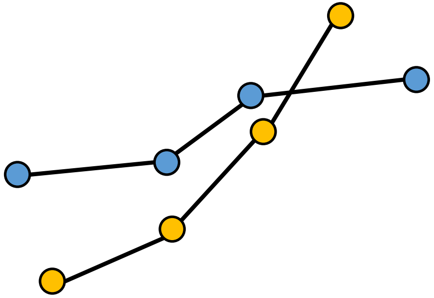
- Conventional Deep RL
  - Parametric
  - Value/Policy Learning
  - Slow gradient-based updates of policy or value functions
- Episodic Control (Learning with memory model)
  - Non-parametric
  - Instance-based learning
  - Rapidly latch onto past successful strategies

Flaws of vanilla episodic control

- No planning



- Not generalizable

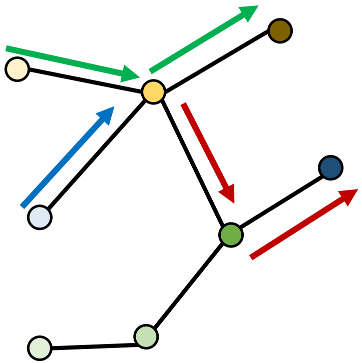


No man ever steps in the same river twice.

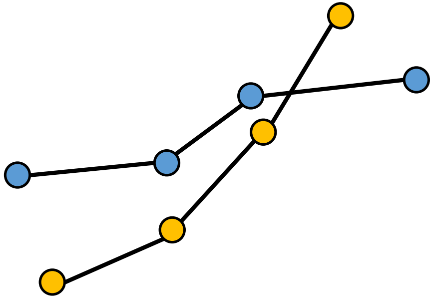
Heraclitus

Flaws of vanilla episodic control

- No planning



- Not generalizable

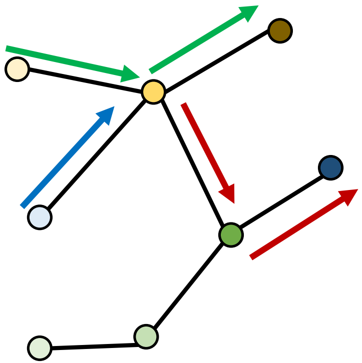


No man ever steps in the same river twice.

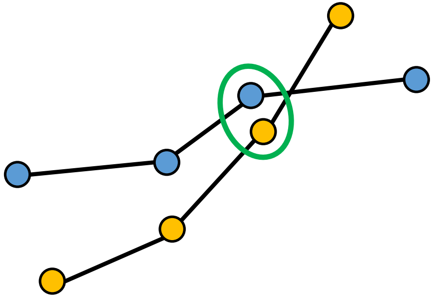
Heraclitus

Flaws of vanilla episodic control

- No planning



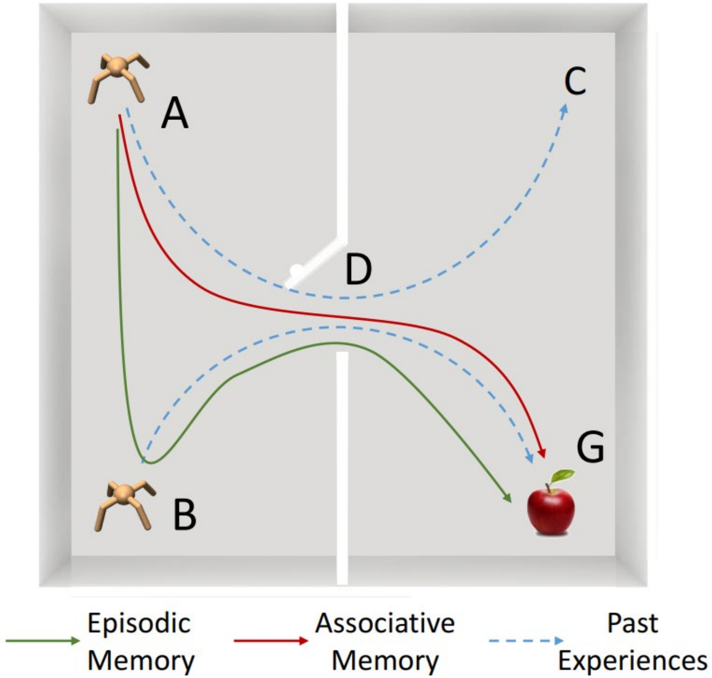
- Not generalizable



No man ever steps in the same river twice.

Heraclitus

Associative Memory



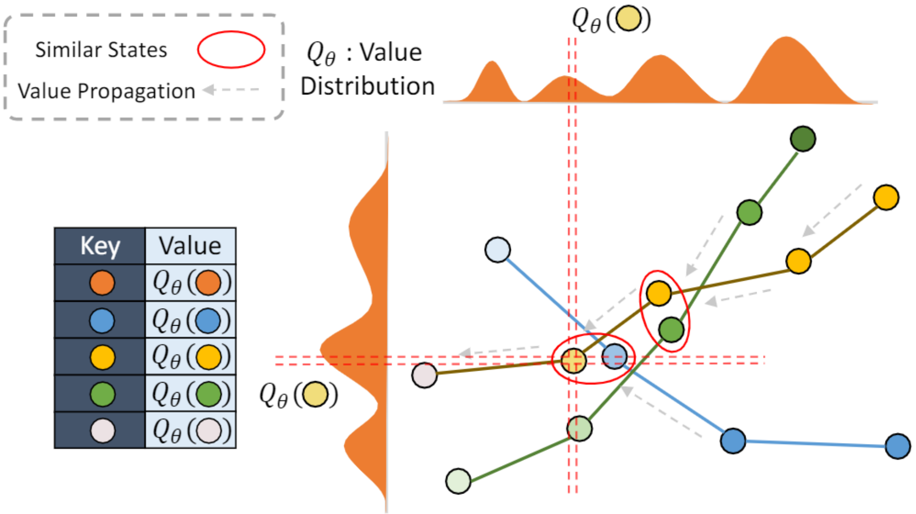
- Graph Induced MDP

$$G = (V, E), V = \{\phi(s) | s \in M\},$$
$$E = \{\phi(s) \rightarrow \phi(s') | (s, a, s') \in M\}$$

- Value propagation in induced MDP

$$Q^{\mathcal{G}}(\phi(s), a) \leftarrow r + \gamma \max_{a'} Q^{\mathcal{G}}(\phi(s'), a')$$

Generalizable Episodic Memory



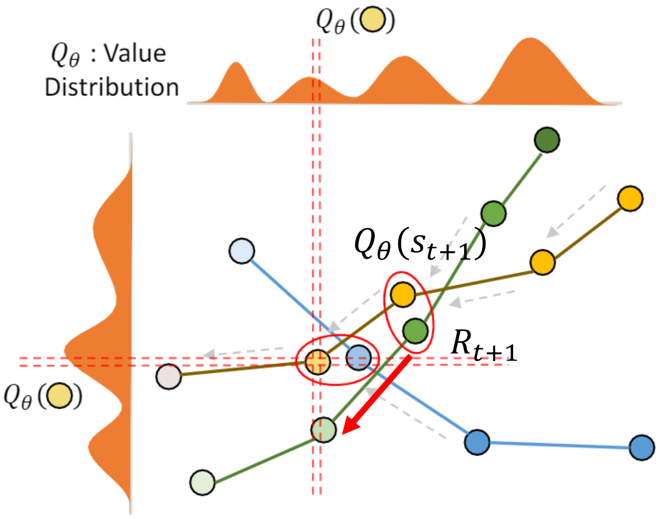
Generalizable Episodic Memory (GEM)

$$\mathcal{L}(Q_\theta) = \mathbb{E}_{(s_t,a_t,R_t)\sim\mathcal{M}}(Q_\theta(s_t,a_t) - R_t)^2.$$

Connecting Experiences

- Implicit planning with memory

$$R_t = \begin{cases} r_t + \gamma \max(R_{t+1}, Q_\theta(s_{t+1})) & \text{if } t < T, \\ r_t & \text{if } t = T. \end{cases}$$



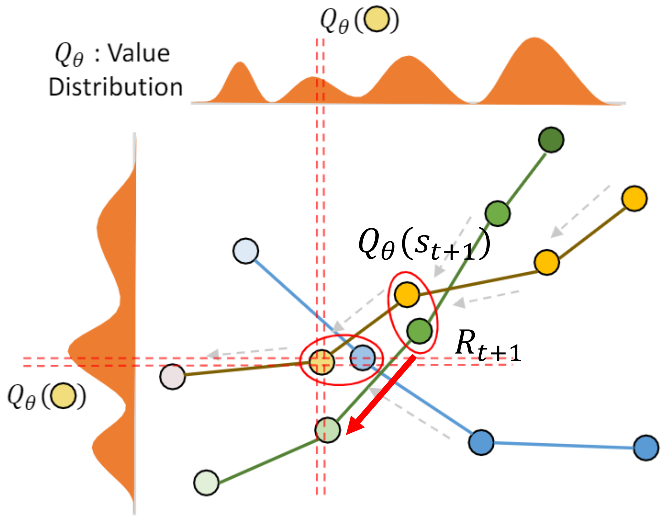
Generalizable Episodic Memory (GEM)

Connecting Experiences

- Equivalently,

$$R_{t,h} = \begin{cases} r_t + \gamma R_{t+1,h-1}, & \text{, if } h > 0, \\ Q_\theta(s_t) & \text{, if } h = 0, \end{cases}$$

$$R_t = R_{t,h^*}, h^* = \arg\max_h R_{t,h},$$



Generalizable Episodic Memory (GEM)

Practical Issues

- Overestimation
  - For a set of unbiased, independent estimators $\tilde{Q}_h = Q_h + \epsilon_h, h \in \{1, \ldots, H\}$,

$$\mathbb{E}\left[\max_h \tilde{Q}_h\right] \geq \max_h \mathbb{E}[\tilde{Q}_h] = \max_h \mathbb{E}[Q_h]$$

e.g.

$$Q_1 = \begin{cases} 1.5, p = \frac{1}{2} \\ 0.5, p = \frac{1}{2} \end{cases}, Q_2 = \begin{cases} 1.6, p = \frac{1}{2} \\ 0.6, p = \frac{1}{2} \end{cases}$$

$$\max(\mathbb{E}[Q_1], \mathbb{E}[Q_2]) = \max(1,1.1) = 1.1$$

$$\mathbb{E}[\max(Q_1, Q_2)] = \frac{1}{2} \times 1.6 + \frac{1}{4} \times 1.5 + \frac{1}{4} \times 0.6 = 1.325 > 1.1$$

Practical Issues

- Double Q-learning

$$\hat{Q} = \max_h Q = Q_{h^*}, h^* = \operatorname{argmax} \mathrm{Q}_h$$

$$Q_{tar} = r_t + \gamma \max_a Q(s_{t+1}, a)$$

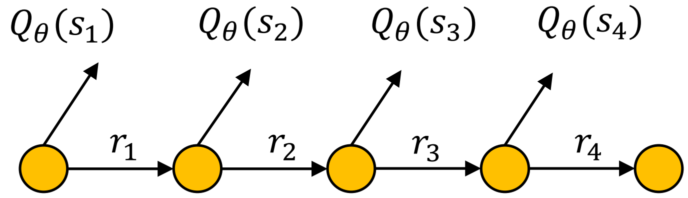
$$\hat{Q}_{Double} = Q^{(2)}_{h^*_{(1)}}, h^*_{(1)} = \operatorname{argmax} Q^{(1)}_h$$

$Q^{(1)}$	0.5	1.2	1.0	0.3
$Q^{(2)}$	0.6	1.1	1.2	0.4

- What double estimator guarantees:

$$\mathbb{E}\left[\hat{Q}_{Double}\right] \leq \max_h \mathbb{E}[Q_h]$$

Twin back-propagation process



$V_h^{(1)}$	4	3.5	3	2
$V_h^{(2)}$	3.8	4	3	2

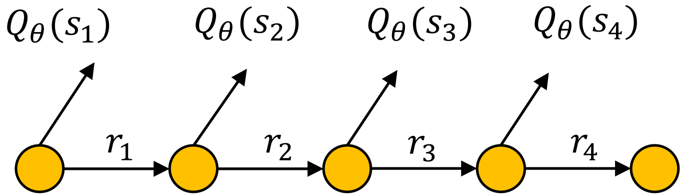
$$h^*_{(2)} = \operatorname{argmax} V_h^{(2)} = 2$$

$$h^*_{(1)} = \operatorname{argmax} V_h^{(1)} = 1$$

$$R^{(1)} = V^{(1)}_{h^*_{(2)}} = 3.5$$

$$R^{(2)} = V^{(2)}_{h^*_{(1)}} = 3.8$$

Twin back-propagation process



$V_h^{(1)}$	4	3.5	3	2
$V_h^{(2)}$	3.8	4	3	2

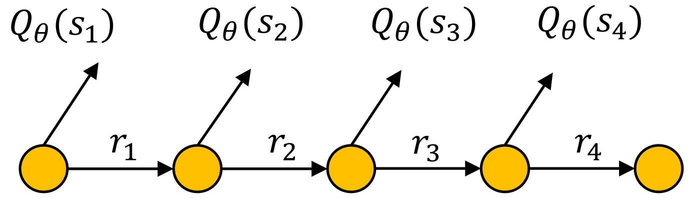
$$h^*_{(2)} = \operatorname{argmax} V_h^{(2)} = 2$$

$$h^*_{(1)} = \operatorname{argmax} V_h^{(1)} = 1$$

$$R^{(1)} = V^{(1)}_{h^*_{(2)}} = 3.5$$

$$R^{(2)} = V^{(2)}_{h^*_{(1)}} = 3.8$$

Twin back-propagation process



$V_h^{(1)}$	4	3.5	3	2
$V_h^{(2)}$	3.8	4	3	2

$$h^*_{(2)} = \operatorname{argmax} V_h^{(2)} = 2$$

$$h^*_{(1)} = \operatorname{argmax} V_h^{(2)} = 1$$

$$R^{(1)} = V^{(1)}_{h^*_{(2)}} = 3.5$$

$$R^{(2)} = V^{(2)}_{h^*_{(1)}} = 3.8$$

Twin back-propagation process

- Twin back-propagation does not overestimate

Theorem 1. Given $\boxed{\text{unbiased and independent}}$ estimators $\tilde{Q}^{\pi}_{(1,2)}(s_{t+h}, a_{t+h}) = Q^{\pi}(s_{t+h}, a_{t+h}) + \epsilon_h^{(1,2)}$, Equation (7) will not overestimate the true objective, i.e.

$$\mathbb{E}_{\tau,\epsilon}\left[R_t^{(1,2)}(s_t)\right] \leq \mathbb{E}_{\tau}\left[\max_{0\leq h\leq T-t-1} Q^{\pi}_{t,h}(s_t)\right], \tag{12}$$

Conservative Estimation

- Clipped Double-Q Learning

$$Q(s, a) = \min\{Q_A(s, a), Q_B(s, a)\}$$

- Asymmetric Loss

$$\mathcal{L}(\theta) = \mathbb{E}[(\delta_t)_+^2 + \alpha(-\delta_t)_+^2]$$

Conservative Estimation

- Conservative estimation as expectile
  - Quantile: minimizer of quantile regression loss

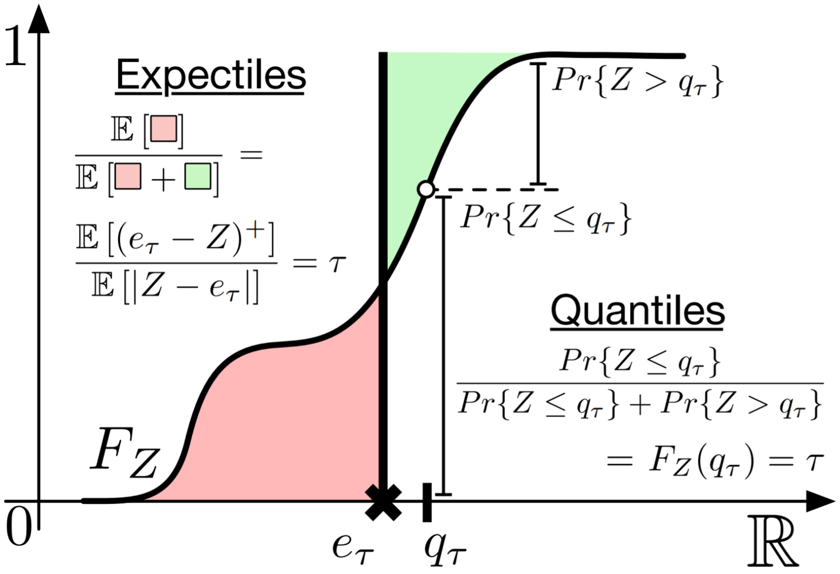
$$QR(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[\left(\tau\mathbb{1}_{\tau>q} + (1-\tau)\mathbb{1}_{\tau\le q}\right)|Z-q|\right]$$

  - Expectile: minimizer of expectile regression loss

$$ER(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[\left(\tau\mathbb{1}_{\tau>q} + (1-\tau)\mathbb{1}_{\tau\le q}\right)(Z-q)^2\right]$$

Conservative Estimation

- Conservative estimation as expectile



[Rowland et al. 2019]

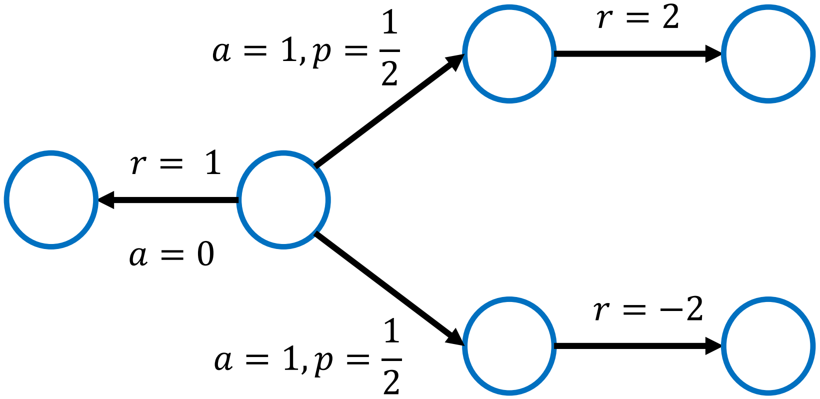
Convergence Analysis

Theorem 2. Algorithm 3 converge to Q^* w.p.1 with the following conditions:

1. The MDP is finite, i.e. $|\mathcal{S} \times \mathcal{A}| \leq \infty$
2. $\gamma \in [0, 1)$
3. The Q-values are stored in a lookup table
4. $\alpha_t(s, a) \in [0, 1], \sum_t \alpha_t(s, a) = \infty, \sum_t \alpha_t^2(s, a) \leq \infty$
5. The environment is fully deterministic, i.e. $P(s'|s, a) = \delta(s' = f(s, a))$ for some deterministic transition function f

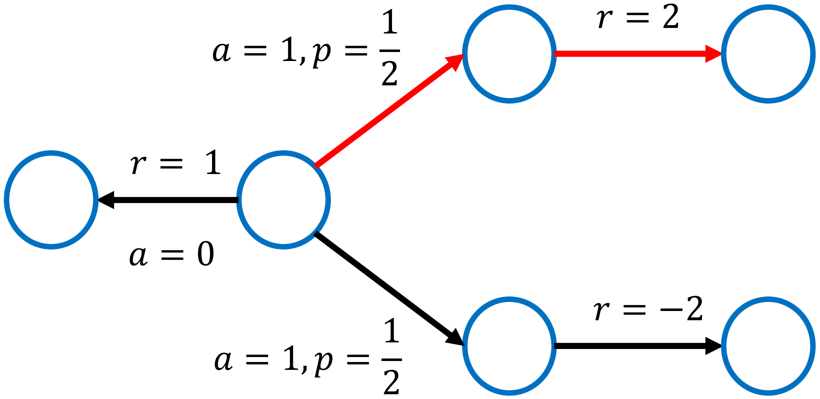
Practical Issues

- Stochastic Environments



Practical Issues

- Stochastic Environments

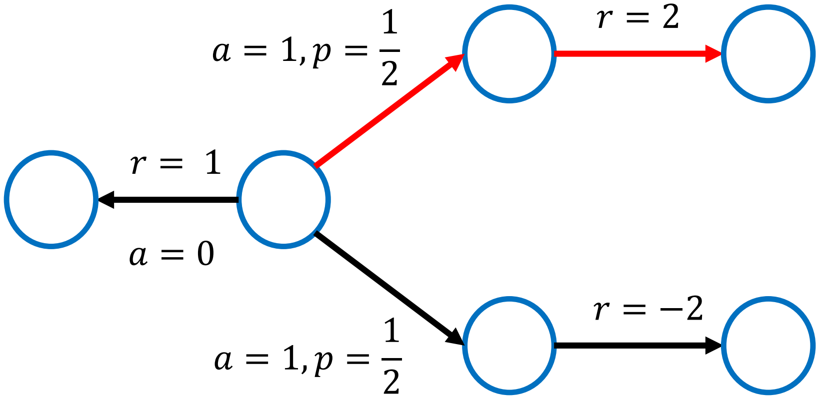


Environment Randomness makes planning within memory fail

But to what extent?

Practical Issues

- Stochastic Environments



Definition 4.1. We define $Q_{max}(s_0, a_0)$ as the maximum value possible to receive starting from (s_0, a_0), i.e.,

$$Q_{max}(s_0, a_0) := \max_{\substack{(s_1, \cdots, s_T), (a_1, \cdots, a_T) \\ s_{i+1} \in supp(P(\cdot|s_i, a_i))}} \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$$

An MDP is said to be nearly-deterministic with parameter μ, if $\forall s \in \mathcal{S}, a \in \mathcal{A}$,

$$Q_{max}(s, a) \leq Q^*(s, a) + \mu$$

where μ is a dependency threshold to bound the stochasticity of environments.

26

Practical Issues

- Stochastic Environments
  - For a nearly-deterministic environment with factor μ, GEM's performance can be bounded by

$$V^{\pi}(s) \geq V^{*}(s) - \frac{2\mu}{1-\gamma}$$

Off-Policy Trade-offs

- Off-Policy evaluation for π with behavior μ
- Consider a general operator $\mathcal{T}$ and assume it has a fix point $\tilde{Q}$
  - Concentration rate of the operator

$$\Gamma(\mathcal{T}) = \sup_{Q_1 \neq Q_2} \frac{\|\mathcal{T}(Q_1 - Q_2)\|_\infty}{\|Q_1 - Q_2\|_\infty}$$

  - the variance and bias of the operator

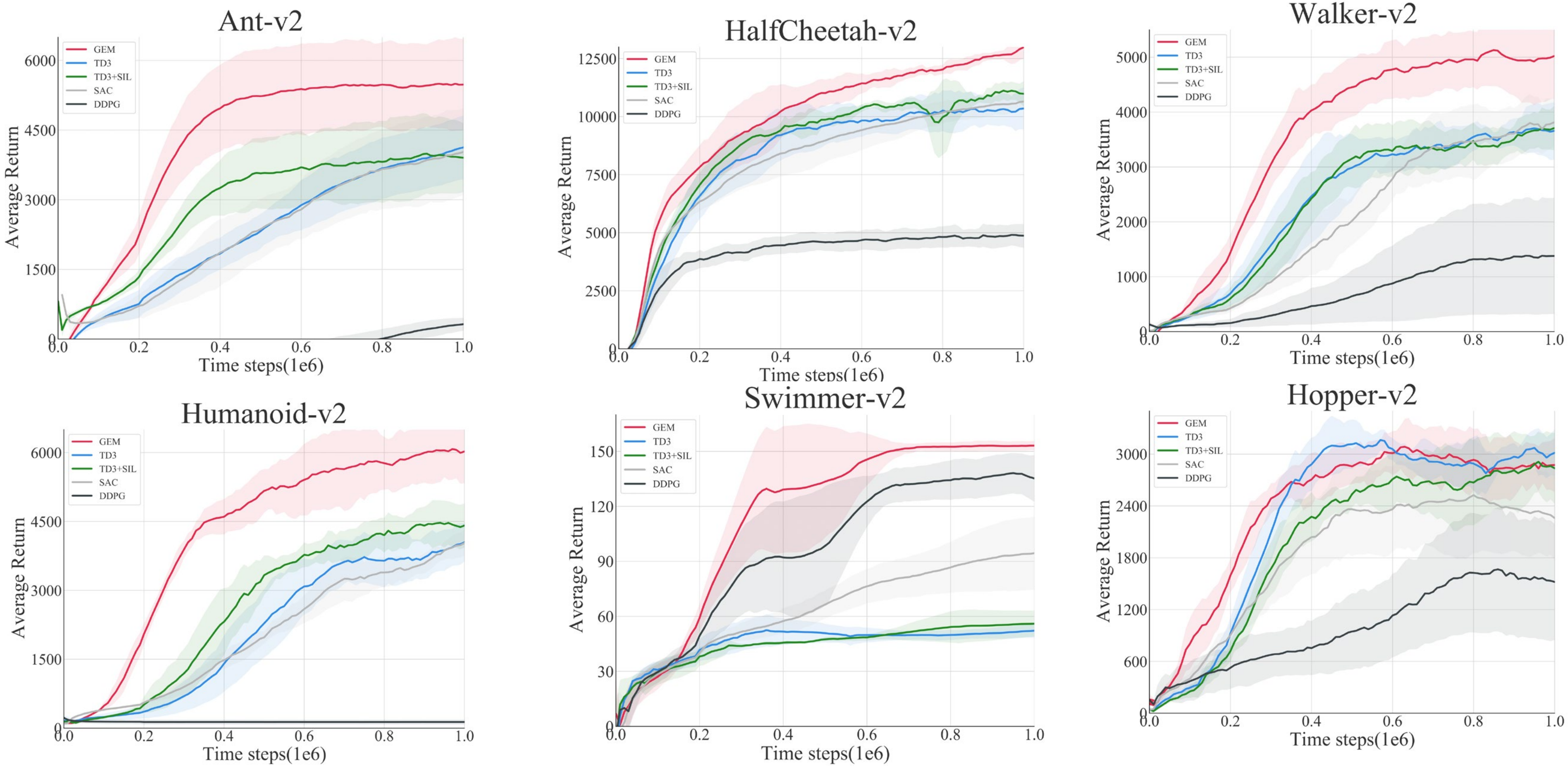
$$\mathbb{B}(\mathcal{T}) = \left\|\tilde{Q} - Q^\pi\right\|_2, \mathbb{V}(\mathcal{T}) = \mathbb{E}_\mu[\|\tilde{\mathcal{T}}Q - \mathcal{T}Q\|_2^2]$$

Off-Policy Trade-offs

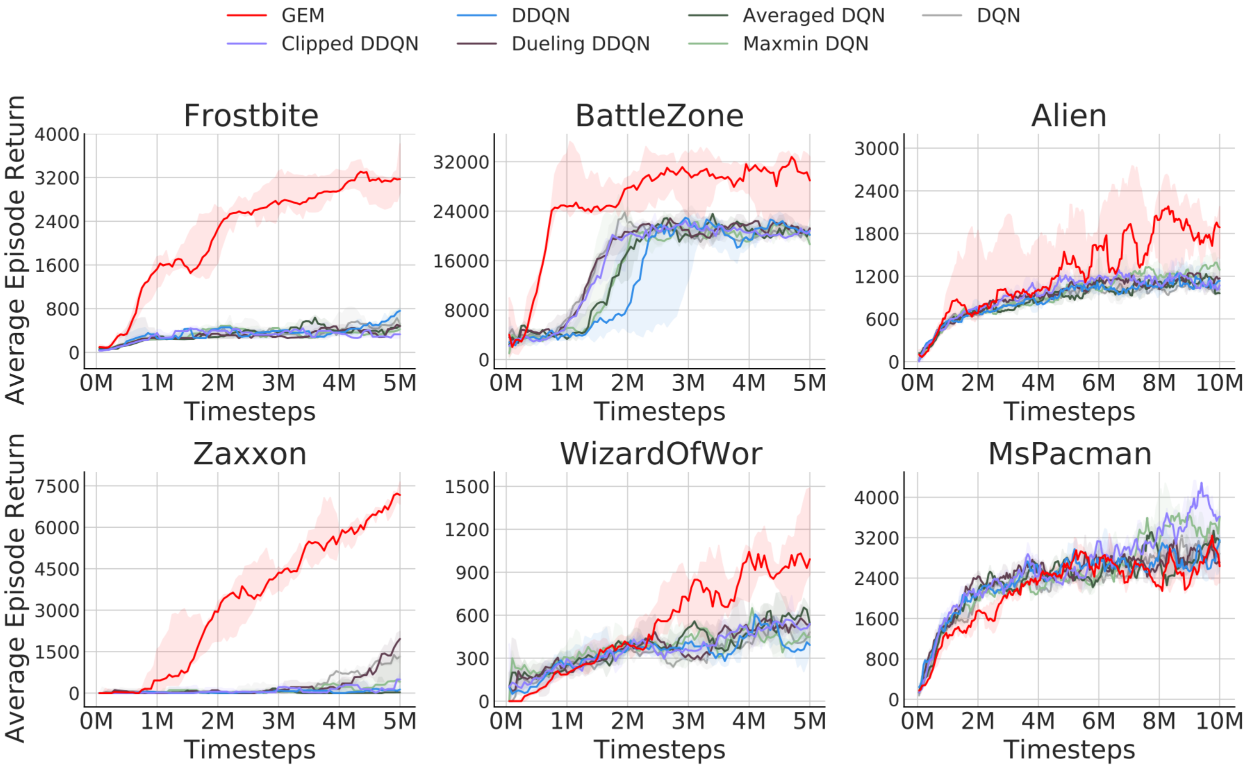
- An information-theoretic lower bound [Rowland et al.]:

$$\sup_{M \in \mathcal{M}} \left\{ \mathbb{B}(\mathcal{T}) + \sqrt{\mathbb{V}(\mathcal{T})} + \frac{2r_{max}}{1-\gamma} \Gamma(\mathcal{T}) \right\} \geq I(\mathcal{M})$$

Experiments

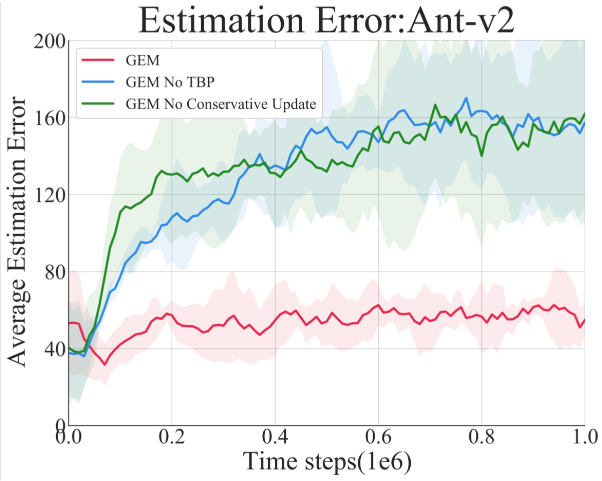


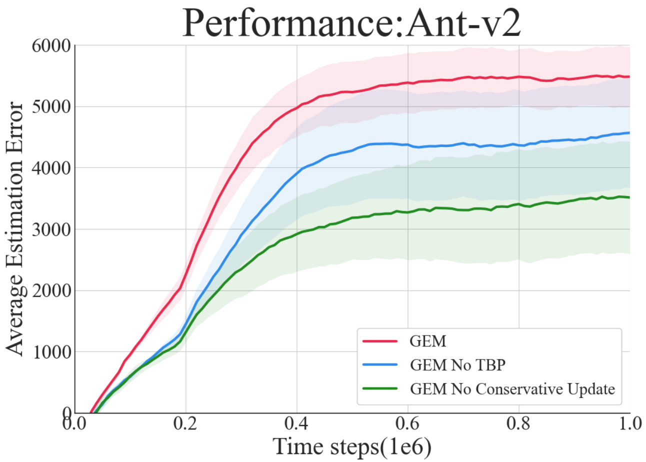
Experiments



Experiments

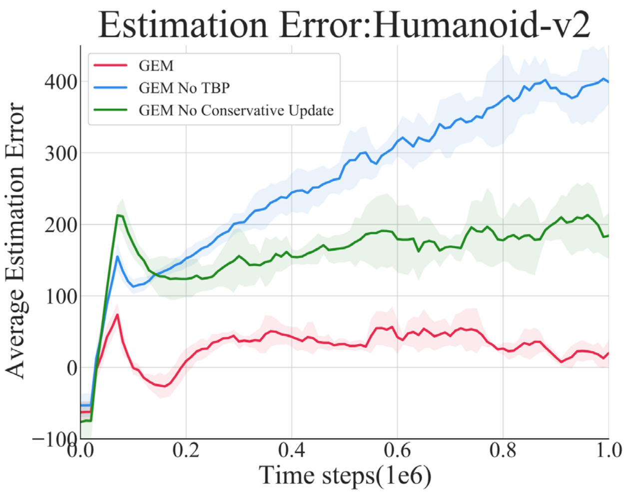
- Ablation study for overestimation

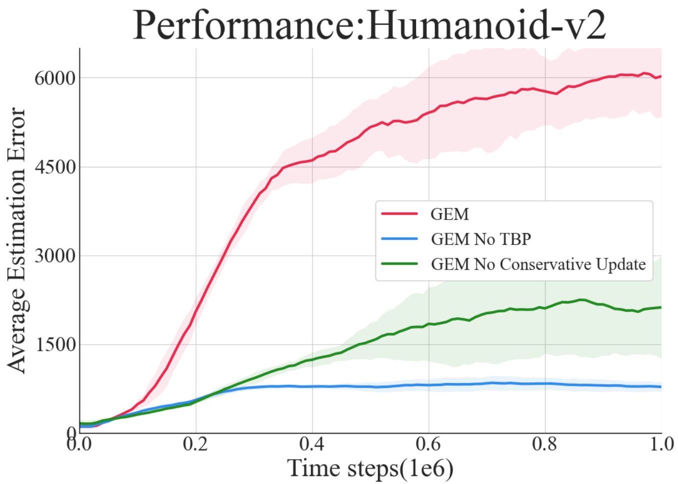




Experiments

- Ablation study for overestimation





Summary

- Episodic memory-based method offers a way for sample-efficient learning
- GEM uses a neural network for natural generalization of discrete memory tables
- TBP reduces overestimation error in planning
- GEM convergences to optimal in deterministic environments and offers trade-offs in stochastic ones

References

[1] Tsividis, Pedro A., et al. "Human learning in Atari." (2017).

[2] Gamrian, Shani, and Yoav Goldberg. "Transfer learning for related reinforcement learning tasks via image-to-image translation." International Conference on Machine Learning. PMLR, 2019.

[3] Blundell, Charles, et al. "Model-free episodic control." arXiv preprint arXiv:1606.04460 (2016).

[4] Hu, Hao, et al. "Generalizable Episodic Memory for Deep Reinforcement Learning." arXiv preprint arXiv:2103.06469 (2021).

[5] Zhang, Jin, et al. "MetaCURE: Meta Reinforcement Learning with Empowerment-Driven Exploration." arXiv preprint arXiv:2006.08170 (2020).

[6] Hafner, Danijar, et al. "Dream to control: Learning behaviors by latent imagination." arXiv preprint arXiv:1912.01603 (2019).

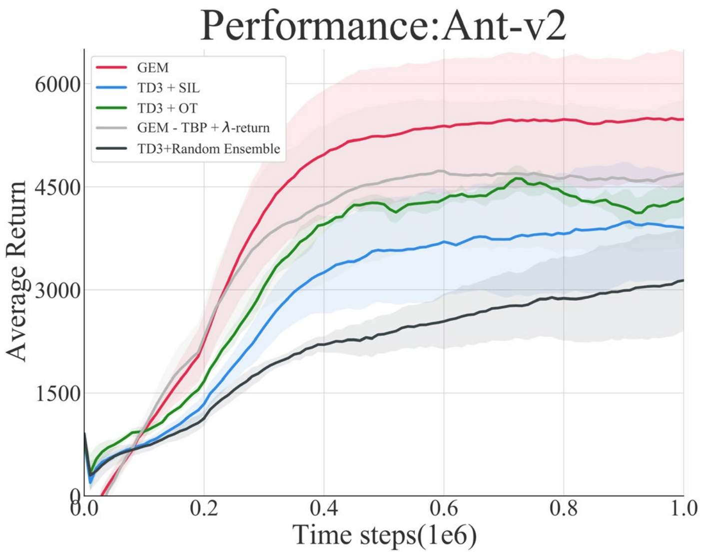
[7] Rowland M, Dadashi R, Kumar S, et al. Statistics and samples in distributional reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2019: 5528-5536.

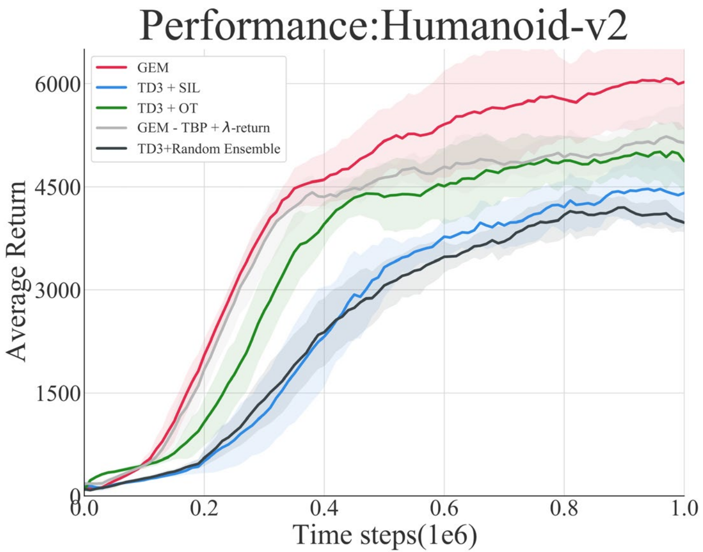
[8] Tang Y. Self-imitation learning via generalized lower bound q-learning[J]. arXiv preprint arXiv:2006.07442, 2020.

[9] David Silver. Tutorial: Deep Reinforcement Learning. https://icml.cc/2016/tutorials/deep_rl_tutorial.pdf

Thanks!

Additional Comparison





Games	GEM	EMDQN	MFEC	NEC
Frostbite	**3030**	596.3	925.1	2747.4
BattleZone	**34600**	28300	19053.6	13345.5
Zaxxon	8180	7740	6288.1	**10082.4**

Table 1: Comparison with existing episodic-memory methods at 10M frames.