Generalizable Episodic Memory for Deep Reinforcement Learning

Hao Hu, Jianing Ye, Guangxiang Zhu, Zhizhou Ren, Chongjie Zhang

Machine Intelligence Group

清華大學 交叉信息研究院
Tsinghua University Institute for Interdisciplinary Information Sciences

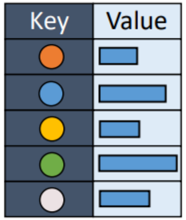
Episodic Control

- Learning

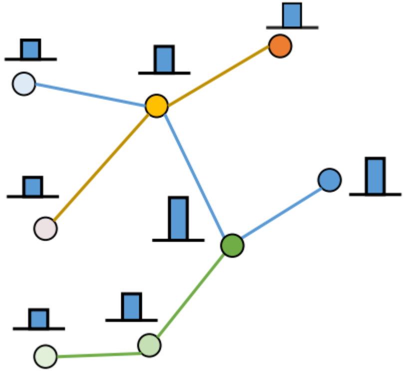
$$Q^{EM}(s,a) = \begin{cases} R, & \text{if } (s,a) \notin EM, \\ \max\{R, Q^{EM}(s,a)\}, & \text{otherwise.} \end{cases}$$

- Execution

$$\widehat{Q}^{EM}(s,a) = \begin{cases} \frac{1}{k}\sum_{i=1}^{k} Q(s_i,a) & \text{if } (s,a) \notin Q^{EM}, \\ Q^{EM}(s,a) & \text{otherwise,} \end{cases}$$

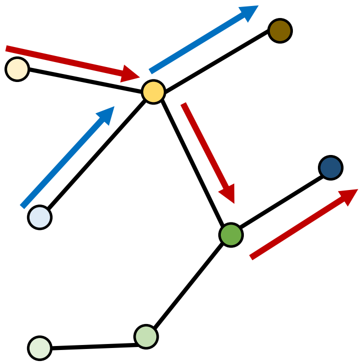


Memory Table

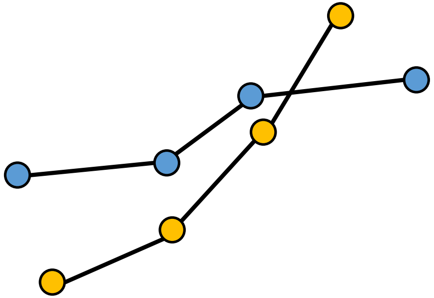


Flaws of vanilla episodic control

- No planning



- Not generalizable

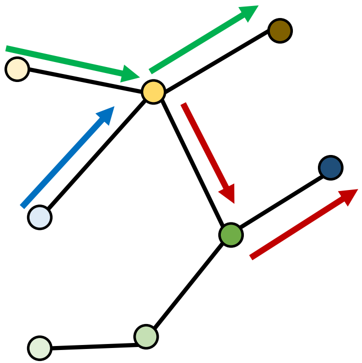


No man ever steps in the same river twice.

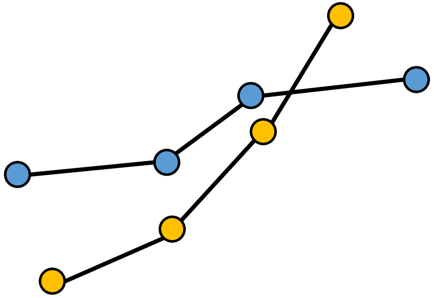
Heraclitus

Flaws of vanilla episodic control

- No planning



- Not generalizable

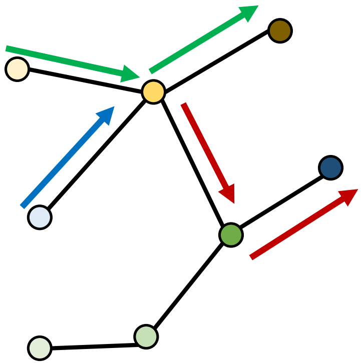


No man ever steps in the same river twice.

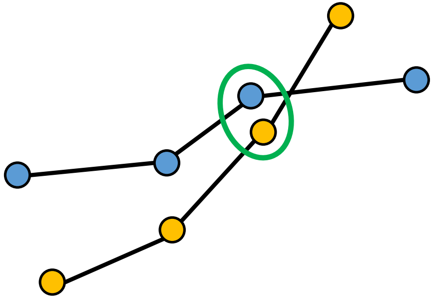
Heraclitus

Flaws of vanilla episodic control

- No planning



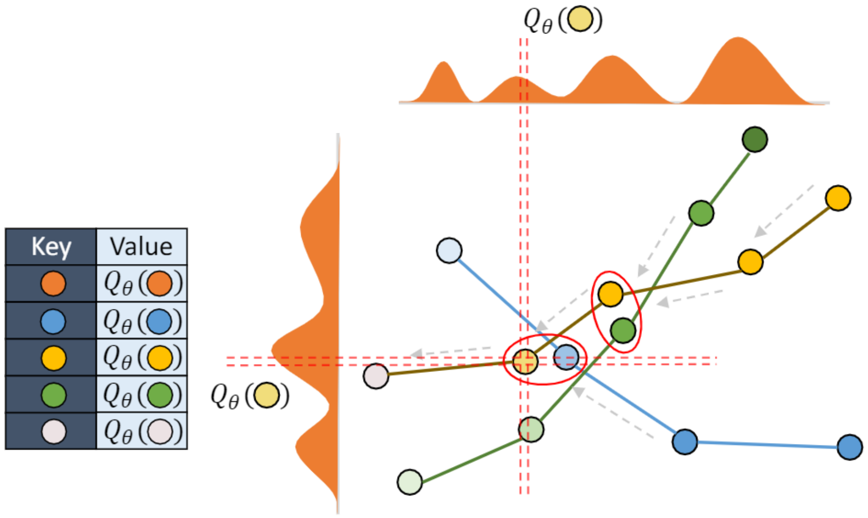
- Not generalizable



No man ever steps in the same river twice.

Heraclitus

Generalizable Episodic Memory



Learn by memorizing discrete tables

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

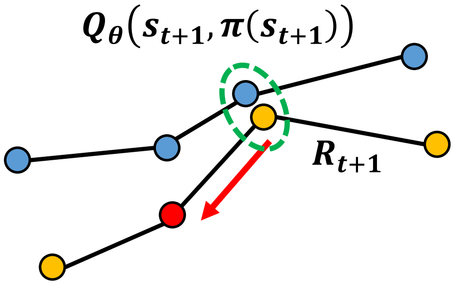
Implicit Planning with Memory

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

Equivalently,

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

$$R_t = V_{t,h^*}, h^* = \arg\max_{h>0} V_{t,h},$$



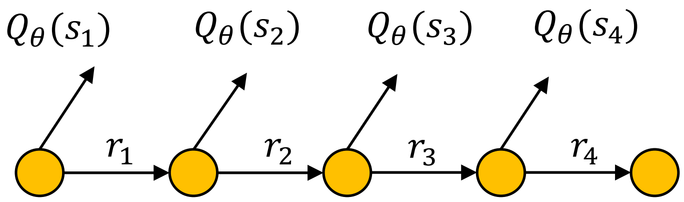
Practical Issues: Overestimation

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

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

- This can be derived directly from Jensen's Inequality.

Twin back-propagation process



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

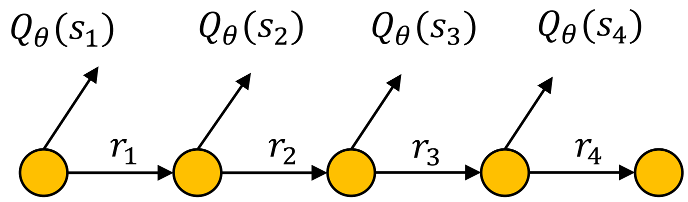
$$h^*_{(2)} = \operatorname{argmax} V_h^{(2)} = 2$$

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

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

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

Twin back-propagation process



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

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

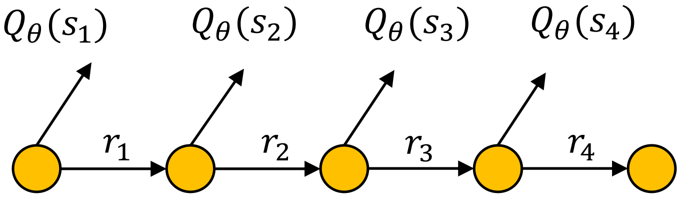
$$h^*_{(1)} = \operatorname{argmax} V_h^{(1)} = 1$$

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

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

10

Twin back-propagation process



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

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

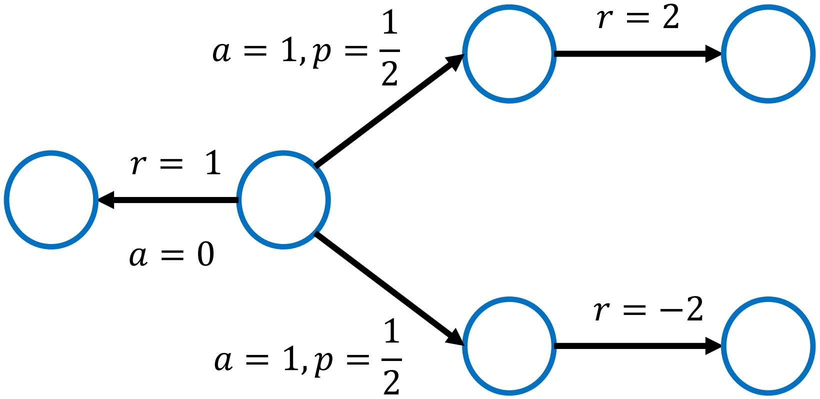
$$h^*_{(1)} = \operatorname{argmax} V_h^{(2)} = 1$$

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

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

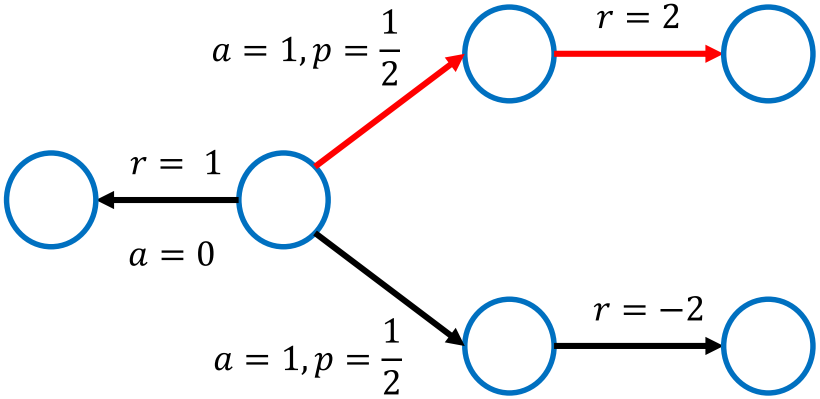
Generalizable Episodic Memory

- Practical Issues: Stochastic Environments



Generalizable Episodic Memory

- Practical Issues: Stochastic Environments

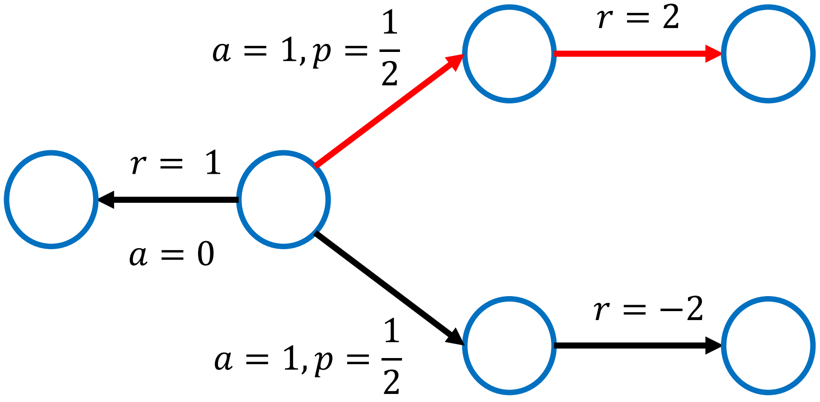


Environment Randomness makes planning fail!

But to what extent?

Generalizable Episodic Memory

- 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.,

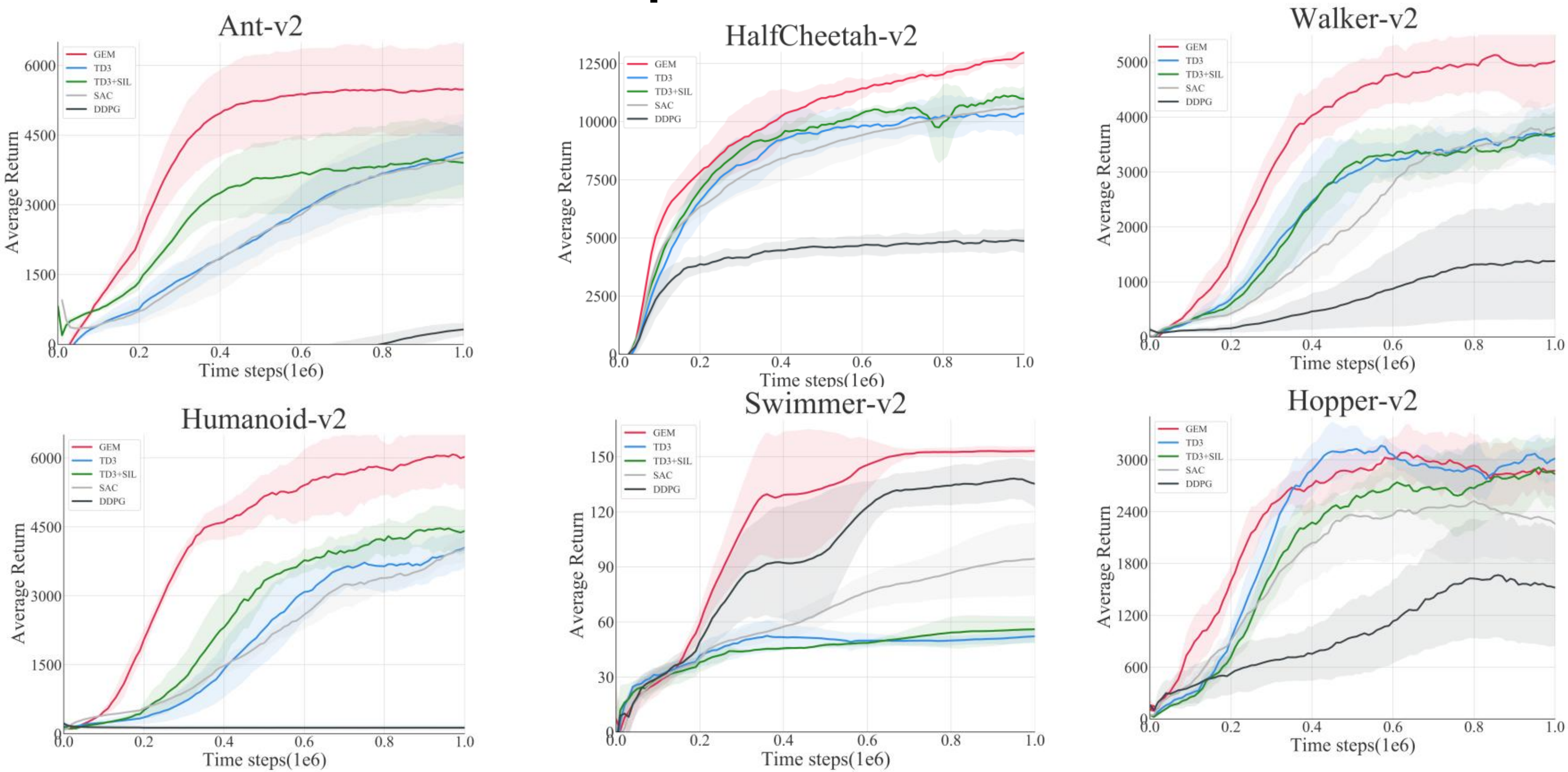
$$Q_{max}(s_0, a_0) := \max_{\substack{(s_1, \cdots, s_T), (a_1, \cdots, a_T) \\ s_{i+1} \in supp(P(\cdot|s_i, a_i))}} \sum_{t=0}^{T} \gamma^t r(s_t, a_t)$$

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

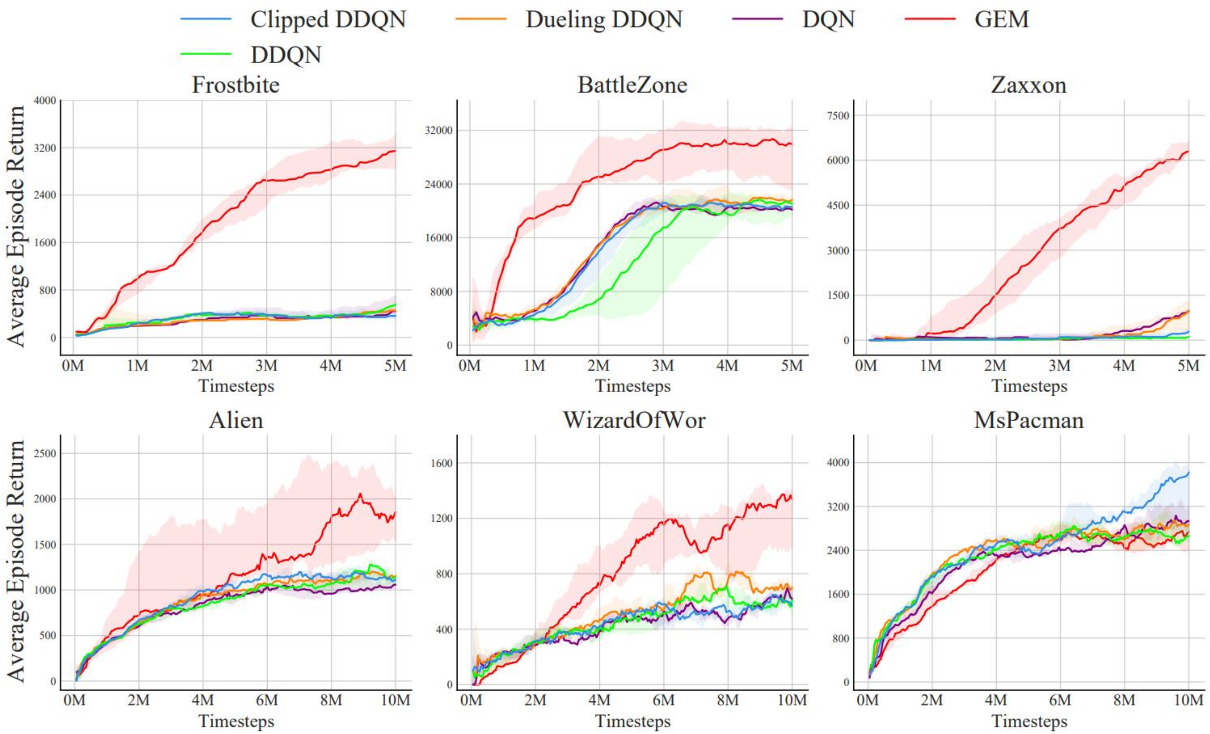
$$Q_{max}(s, a) \leq Q^*(s, a) + \mu$$

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

Experiments

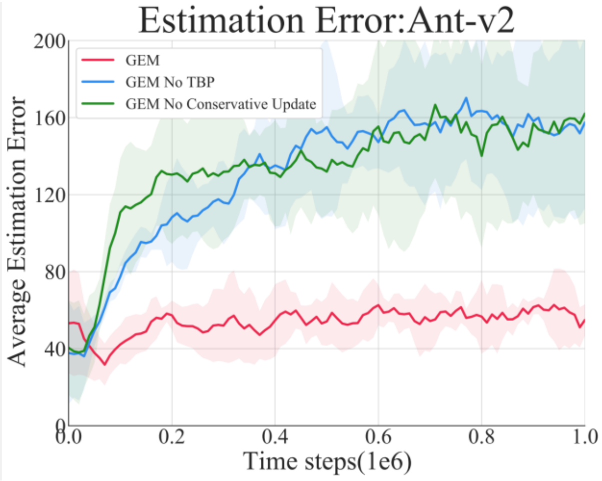


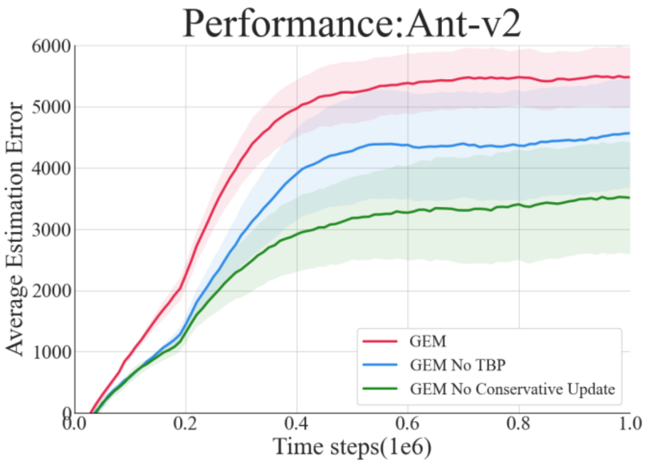
Experiments



Experiments

- Reducing overestimation





Summary

Thanks!

- Check out our paper for more details
- Code available at https://github.com/MouseHu/GEM
- Happy to answer questions by email:

  hu-h19@mails.tsinghua.edu.cn chongjie@tsinghua.edu.cn