SOLVING CONTINUOUS CONTROL VIA EPISODIC MEMORY

IGOR KUZNETSOV, ANDREY FILCHENKOV

ITMO UNIVERSITY

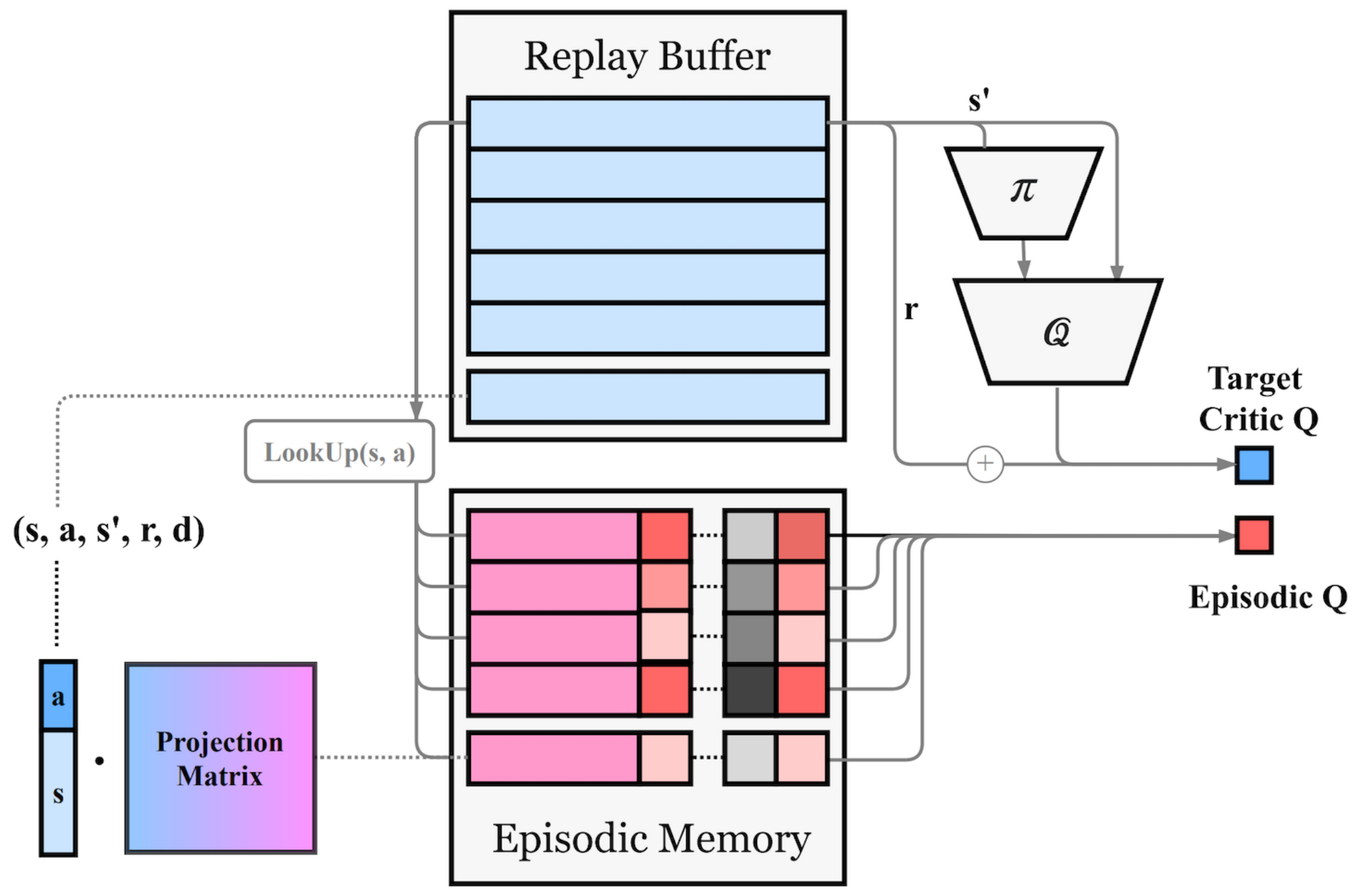
ABSTRACT

We present EMAC, a model-free off-policy algorithm that uses episodic memory to solve continuous control tasks. We use episodic memory module to store representations of (s, a) pairs with corresponding returns. Stored memories allow us to reduce Q-value overestimation and introduce episodic prioritisation. Based on DDPG, EMAC exceeds DDPG, TD3 on all tested environments and SAC on 3 out of 5 environments

Keywords: episodic memory, off-policy algorithms, continuous control

MODEL ARCHITECTURE

- We use table-based structure for storing memories
- Random Projection operator is applied to concatenation of (s, a)
- **Add** operation appends new $(s, a) : R$ to the end of module
- **Lookup** operation searches for the similar (s, a) tuple in projected space and returns corresponding MC returns



Q-VALUE OVERESTIMATION

- Given the (s, a) pair, we can get MC returns from the similar states and actions that were seen in the past.
- We propose the new critic objective

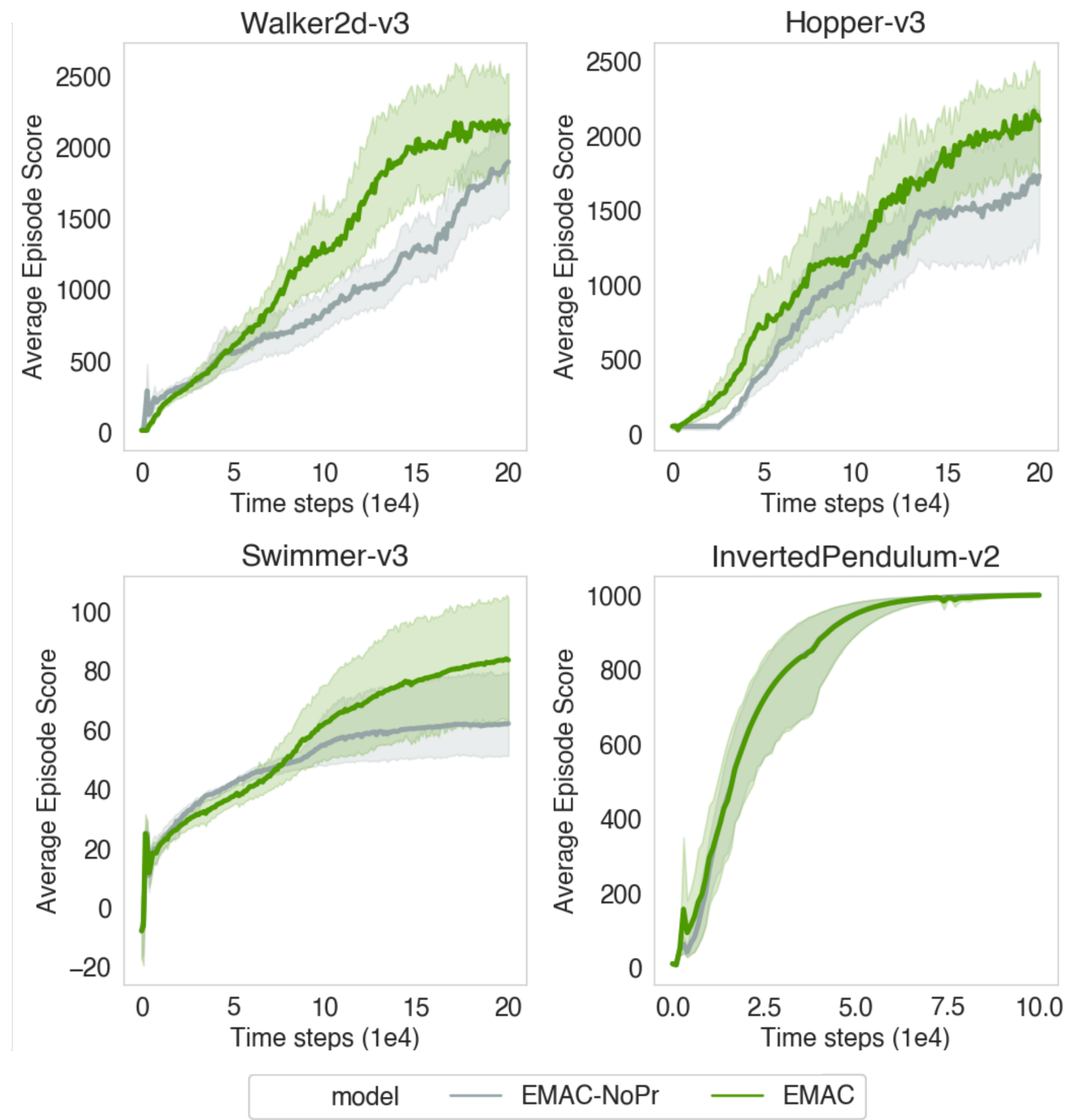
$$J_Q = (Q(s_t, a_t) - Q')^2 + \alpha(Q(s_t a_t) - Q_M)^2 \quad (1)$$

- MC returns from the past suboptimal policy are generally less than the true Q-estimate
- Therefore Q_M can be seen as a penalty to the Q-overestimation of a critic

EPISODIC PRIORITIZATION

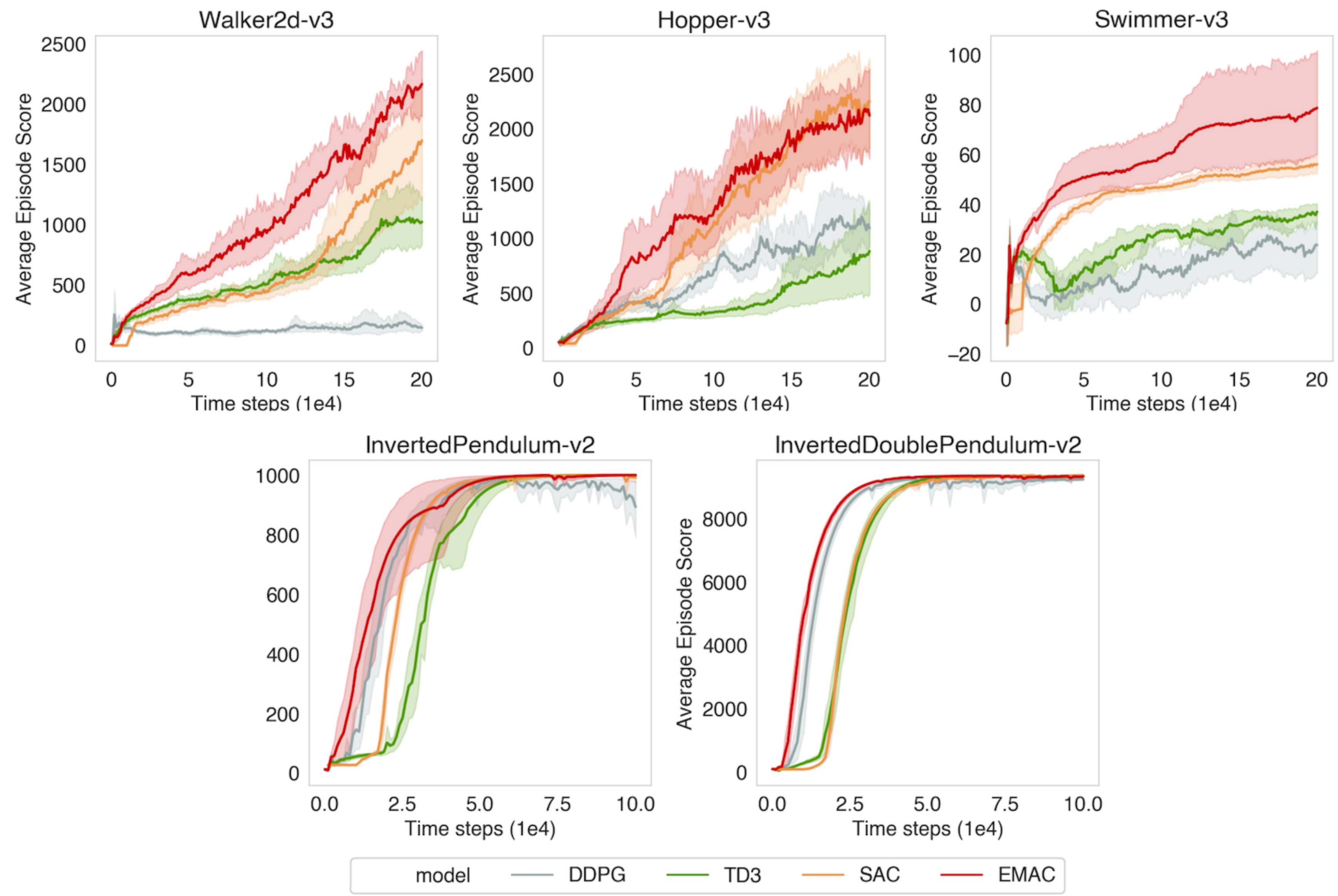
We exploit stored MC returns (p_i) as a measure of prioritization, which improves performance

$$P(i) = \frac{p_i^\beta}{\sum_k p_k^\beta}, \quad (2)$$

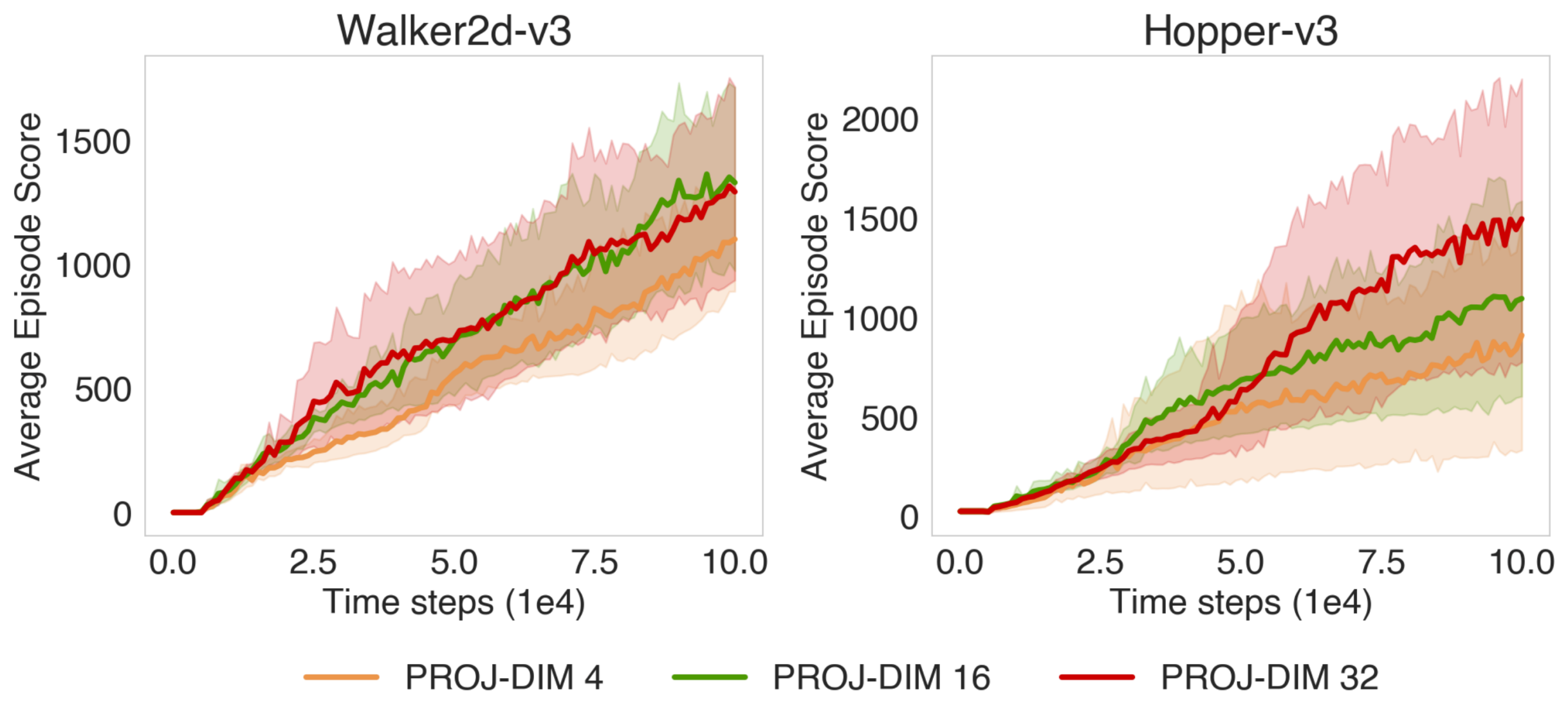


RESULTS

We evaluate the algorithm on 5 OpenAI environments: Walker2d, Hopper, Swimmer, InvertedPendulum, InvertedDoublePendulum on 100k environment steps. EMAC outperforms DDPG and TD3 on all environments and SAC on 3 environments.



DESIGN CHOICE



- During the **lookup** operation we search for $K = 1$ or $K = 2$ nearest (s, a) pairs
- Memory module capacity is equal to the capacity of replay buffer, which is possible due to the low-data regime
- The projected dimension size is set to the minimal size 4 for speed-up **lookup**
- The search for the similar (s, a) pairs is performed with $L2$ distance and vectorized with CUDA

REFERENCES

[1] Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *ICML*, 2017.

[2] Zichuan Lin, Tianqi Zhao, Guangwen Yang, and Lintao Zhang. Episodic memory deep q-networks. In *IJCAI*, 2018.

FUTURE RESEARCH

Current memory representation with random projections has its limits as it does not reflect topological structure of states and actions. Therefore, we plan to use more complex differential memory representations in future. Another direction of research is explicitly introducing short-term working memory mechanisms alongside episodic memory. The motivation is to mimic human learning system with long-term and short-term memory mechanisms, exploiting the benefits from both.

CONTACT INFORMATION

Email igorkuznetsov14@gmail.com
Twitter @schatty_
Web https://schatty.github.io