Towards Robust Bisimulation Metric Learning

Mete Kemertas, Tristan T. Aumentado-Armstrong
(equal contribution)

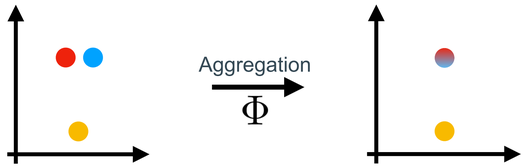
Computer Science
UNIVERSITY OF TORONTO

NEURAL INFORMATION PROCESSING SYSTEMS

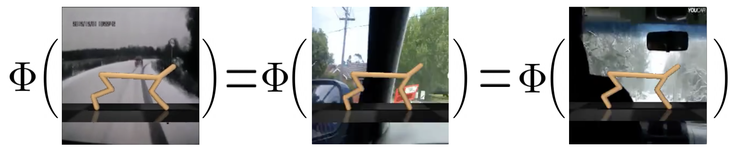
Bisimulation Metrics

Why learn state similarity metrics in RL?

1) Compact representations



2) Distractor invariance



3) Theoretical guarantees in value function approximation (VFA)

- Solving the ϵ-aggregated MDP is approx. equivalent to solving the ground MDP

$$|V^*(\mathbf{s}) - V^*(\Phi(\mathbf{s}))| \leq O(\epsilon)$$

Policy-Independent Bisimulation Metrics [1]

$$d(\mathbf{s}_i, \mathbf{s}_j) = \max_{\mathbf{a} \in \mathcal{A}} (1-c)|R(\mathbf{s}_i, \mathbf{a}) - R(\mathbf{s}_j, \mathbf{a})| \; + \; cW_1(d)(\mathcal{P}(\cdot|\mathbf{s}_i, \mathbf{a}), \mathcal{P}(\cdot|\mathbf{s}_j, \mathbf{a}))$$

Immediate reward difference — Recursive "future" distance

On-Policy Bisimulation Metrics [3]

$$d_\pi(\mathbf{s}_i, \mathbf{s}_j) = c_R|r_i^\pi - r_j^\pi| \; + \; c_T W_p(d_\pi)(\mathcal{P}^\pi(\cdot|\mathbf{s}_i), \mathcal{P}^\pi(\cdot|\mathbf{s}_j))$$

- VFA guarantees above hold for bisimulation metrics when $c_T \geq \gamma$

Main Theoretical Results

1) VFA guarantees hold for non-optimal policies

$$|V^\pi(\mathbf{s}) - V^\pi(\Phi(\mathbf{s}))| \leq O(\epsilon)$$

2) VFA guarantees as a function of c_T and model error $\mathcal{E}_\mathcal{P}$

$$|V^\pi(\mathbf{s}) - V^\pi(\Phi(\mathbf{s}))| \leq O(\epsilon + \frac{c_T}{1-c_T}\mathcal{E}_\mathcal{P})$$

- Large c_T amplifies VFA guarantee loss due to modelling errors.

3) Expected bisimulation distance (over stationary distribution)

$$\mathbb{E}[d_\pi(\mathbf{s}_i, \mathbf{s}_j)] \approx \frac{c_R}{1-c_T}\mathbb{E}[|r_i^\pi - r_j^\pi|]$$

On-Policy Bisimulation: Issues & Remedies

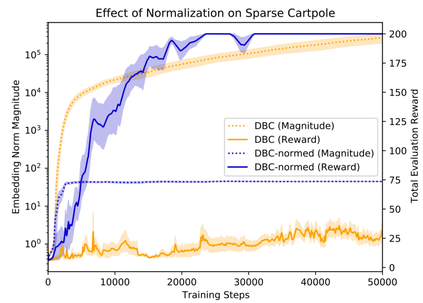
- Learn state representations that respect bisimulation as in DBC [2]

$$||\phi(\mathbf{s}_i) - \phi(\mathbf{s}_j)|| \approx d_\pi(\mathbf{s}_i, \mathbf{s}_j)$$

Embedding Normalization

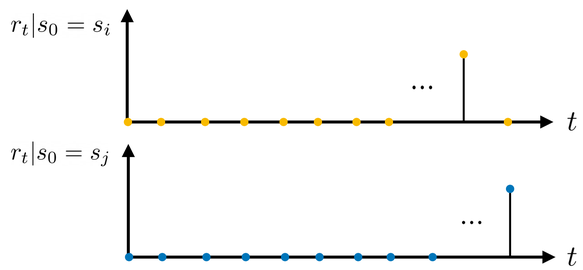
- Bisimulation metrics are theoretically **bounded**.
- Using predicted dynamics may **violate** this and cause divergence.
- Solution: threshold the embedding magnitude

$$||\phi(\mathbf{s})|| \leq \frac{c_R(R_{\max} - R_{\min})}{2(1-c_T)}$$



Intrinsic Rewards

- Sparse reward sequences lead to premature collapse in bisimulation space.



$$d(\mathbf{s}_i, \mathbf{s}_j) = O(c_T^t) \approx 0, \text{for large } t \text{ since } c_T \in [0, 1).$$

- We use curiosity-driven forward model error to enrich the reward signal

$$r_{I,t} = ||\widehat{\phi}_\mu(\mathbf{s}_{t+1}) - \phi(\mathbf{s}_t)||_2^2$$

Predicted next state — Current latent state

Inverse Dynamics (ID) Regularization

- Prevent pathologies by regularizing embedding space to stay informative

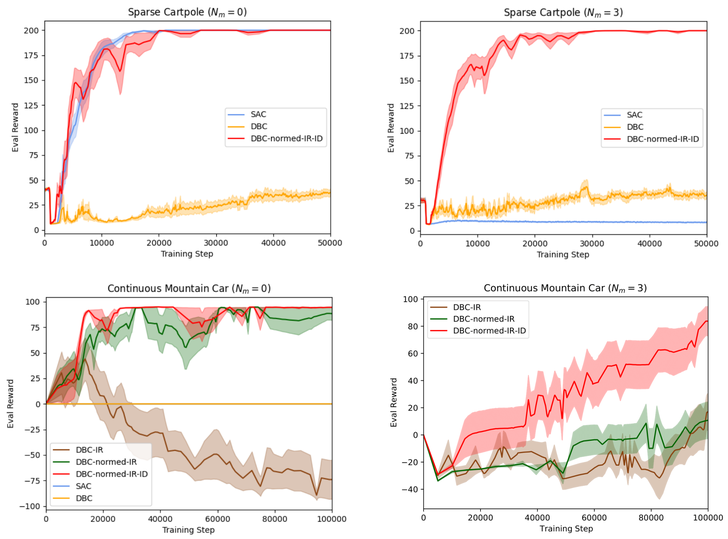
Given $\phi(\mathbf{s}_t), \phi(\mathbf{s}_{t+1})$ → **Predict** $\mathbf{a}_t$

- Also promotes distraction-invariance [4]

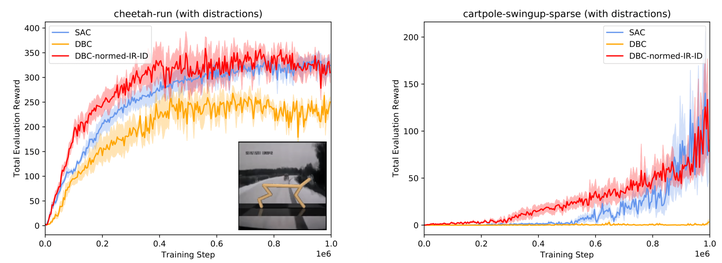
Empirical Results

Modified OpenAI Gym Classical Control Tasks with sparsity & distraction

Low distraction — *High distraction*



Deepmind Control Suite (DMC) with background video distractions



Summary

- Bisimulation metrics have been used to promote distraction-invariance.
- We extended theoretical VFA bounds for bisimulation metrics to **(i)** non-optimal policies and **(ii)** approximate dynamics.
- Our analysis revealed pathologies under sparse rewards and learned dynamics.
- We propose remedies and show performance gains on a suite of tasks.

References

[1] Ferns et al (2011). "Bisimulation metrics for continuous Markov decision processes." In: SIAM Journal on Computing.

[2] Zhang et al (2020). "Learning invariant representations for reinforcement learning without reconstruction." In: International Conference on Learning Representations.

[3] Castro (2020). "Scalable methods for computing state similarity in deterministic Markov decision processes." In: Proceedings of the AAAI Conference on Artificial Intelligence.

[4] Pathak et al (2017). "Curiosity-driven exploration by self-supervised prediction." In: International Conference on Machine Learning.