Bridging State and History Representations: Understanding Self-Predictive RL

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, Pierre-Luc Bacon

Mila, Université de Montréal, Princeton University, McGill

arXiv

Motivation: A Variety of Representations

State Representations

- Model-free RL (deep Q-learning)
- Model-based RL (OFENet)
- Bisimulation (DeepMDP, DBC)
- Self-predictive representations (SPR, TD-MPC, ALM)

History Representations

- Recurrent model-free RL (recurrent Q-learning)
- Belief states (Dreamer), Predictive state representations
- Information states

This paper unifies them with *self-prediction* and provides a simple and principled learning algorithm.

Background: Representations in POMDPs

Notation

- Observation o_t, Action a_t, History $h_t = (o_{1:t}, a_{1:t-1})$.
- In a partially observable MDP, reward $R(h_t, a_t)$ and transition $P(o_{t+1} \mid h_t, a_t)$ depend on histories.
- Encoder $\phi : \mathcal{H}_t \to \mathcal{Z}$ maps a history into a latent state.
- Policy: $\pi(a_t \mid \phi(h_t))$, Value: $Q^\pi(\phi(h_t), a_t)$.
- Below we omit the subscripts on time-steps.

Sufficient statistics of an history for predicting rewards, values, observations, latent states.

1. Q^*-irrelevance abstraction ϕ_{Q^*} [1]. If $\phi_{Q^*}(h^1) = \phi_{Q^*}(h^2)$, then $Q^*(h^1, a) = Q^*(h^2, a)$. *E.g.*, end-to-end recurrent Q-learning on $\mathcal{Q}(\phi(h), a)$ to convergence.

2. Self-predictive abstraction ϕ_L [1, 2]. (1) **Reward Prediction (RP)**, (2) **Latent state Prediction (ZP)** (self-prediction). *E.g.*, bisimulation in MDPs [3] and information states in POMDPs [2].

$$\exists R_z : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}, \text{ s.t. } \mathbb{E}[r \mid h, a] = R_z(\phi_L(h), a), \quad \text{(RP)}$$

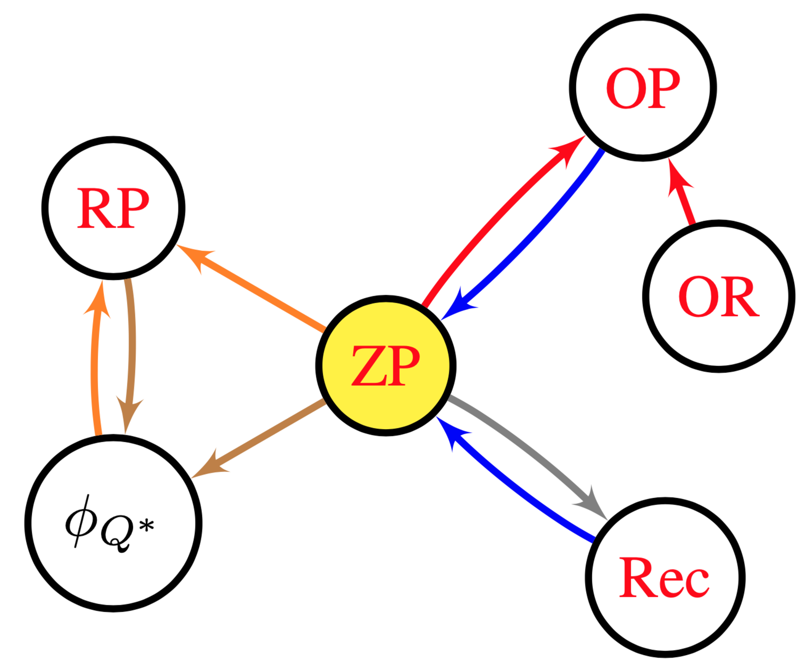
$$\exists P_z : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z}), \text{ s.t. } P(z' \mid h, a) = P_z(z' \mid \phi_L(h), a), \quad \text{(ZP)}$$

3. Observation-predictive abstraction ϕ_O [1, 2]. (1) **Recurrent encoder (Rec)**, (2) **Reward Prediction (RP)**, (3) **Observation Prediction (OP)**. *E.g.*, belief states in POMDPs [4].

$$\exists P_o : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{O}), \text{ s.t. } P(o' \mid h, a) = P_o(o' \mid \phi_O(h), a), \quad \text{(OP)}$$

Their inclusion relationship (we extend from MDPs [1] to POMDPs): ϕ_O is stronger than ϕ_L; ϕ_L is stronger than ϕ_{Q^*}.

A Unified View on History Representations



An implication graph:
Nodes A and B connected to C by the same-color edges *imply* C.

- ZP: next latent state prediction
- OP: next observation prediction
- OR: observation reconstruction
- RP: reward prediction
- Rec: recurrent encoder (MLPs, RNNs)
- ϕ_{Q^*}: optimal value prediction

Learning Self-Predictive Representations

A simple and principled algorithm

We derive a minimalist algorithm to learn ϕ_L (i.e., RP + ZP): ϕ_{Q^*} (end-to-end Q-learning) + ZP (auxiliary task) $\implies$ RP.

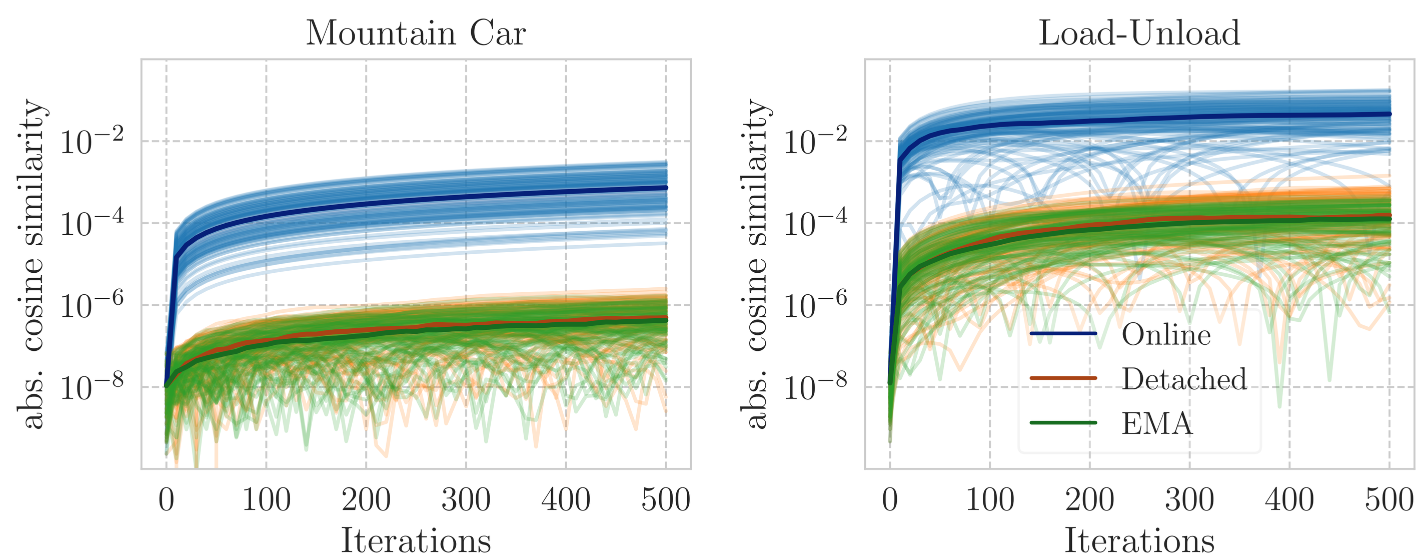
Let $f_\phi : \mathcal{H}_t \to \mathcal{Z}$ be an encoder, $g_\theta : \mathcal{Z} \times \mathcal{A} \to \mathcal{Z}$ be a latent transition model, and $Q_\omega : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ be a latent critic. Sample $(h, a, r, o') \sim \mathcal{D}$ and optimize a single objective:

$$\min_{\phi,\theta,\omega} \underbrace{\mathrm{RL}(Q_\omega; f_\phi(h), a, r, f_{\overline{\phi}}(h'))}_{\phi_{Q^*}: (z,a,r,z'), \text{ diff. thru. } f_\phi(h)} + \lambda \underbrace{\|g_\theta(f_\phi(h), a) - f_{\overline{\phi}}(h')\|_2^2}_{\text{ZP: } \ell_2 \text{ loss, diff. thru. } f_\phi(h)}$$

where $h' = (h, a, o')$ is the next history and $\overline{\phi}$ applies stop-gradient.
This ZP ℓ_2 loss has been widely used in prior work, while our work simplifies them by removing additional components. Furthermore, we show that

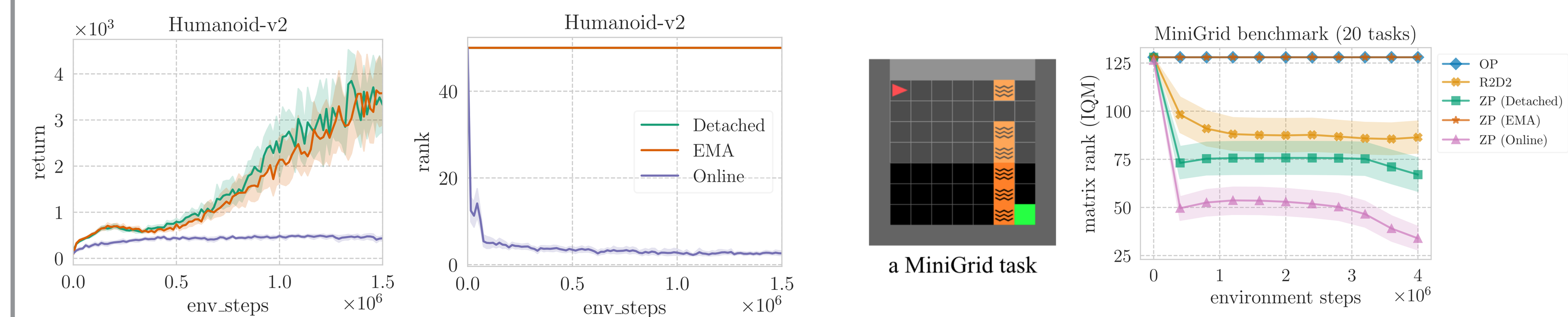
Stop-gradient $\overline{\phi}$ (detached / EMA) provably avoids representational collapse in linear models.

(Extended from [6]) With a linear encoder $f_\phi(h) = \phi^\top h_{-k:} \in \mathbb{R}^d$ and a linear transition model g_θ, if we train g_θ to a stationary point w.r.t. f_ϕ, then $\phi^\top \phi \in \mathbb{R}^{d \times d}$ will retain its initial value during training. Therefore, ϕ will keep full-rank thus avoiding collapse as long as it is orthogonally initialized.

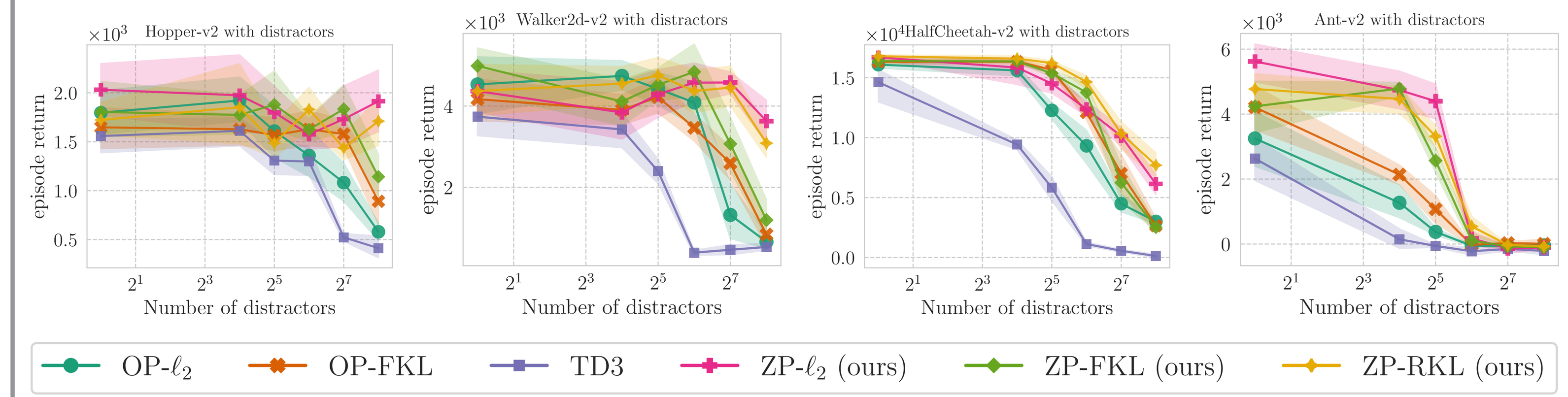


Empirical Findings (ZP)

Stop-gradient in self-prediction mitigates collapse in *deep RL*.

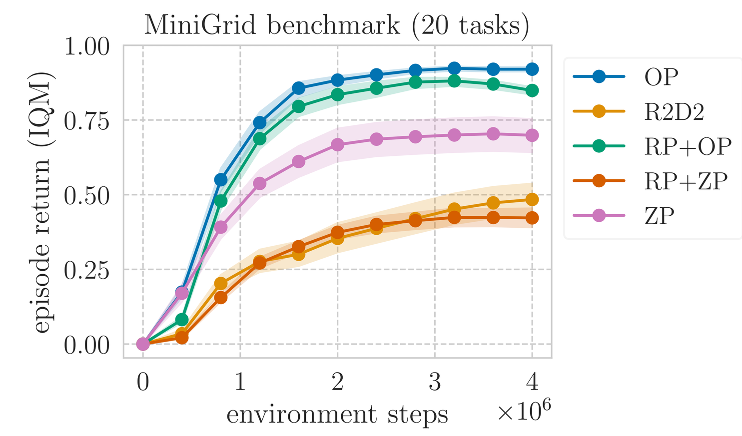


Self-predictive representations are more robust to distractors.



Distracting MuJoCo: concatenating d-dim $\epsilon \sim \mathcal{N}(0, 1)$ vector to the state.
ZP-$*$: self-predictive; OP-$*$: observation-predictive.

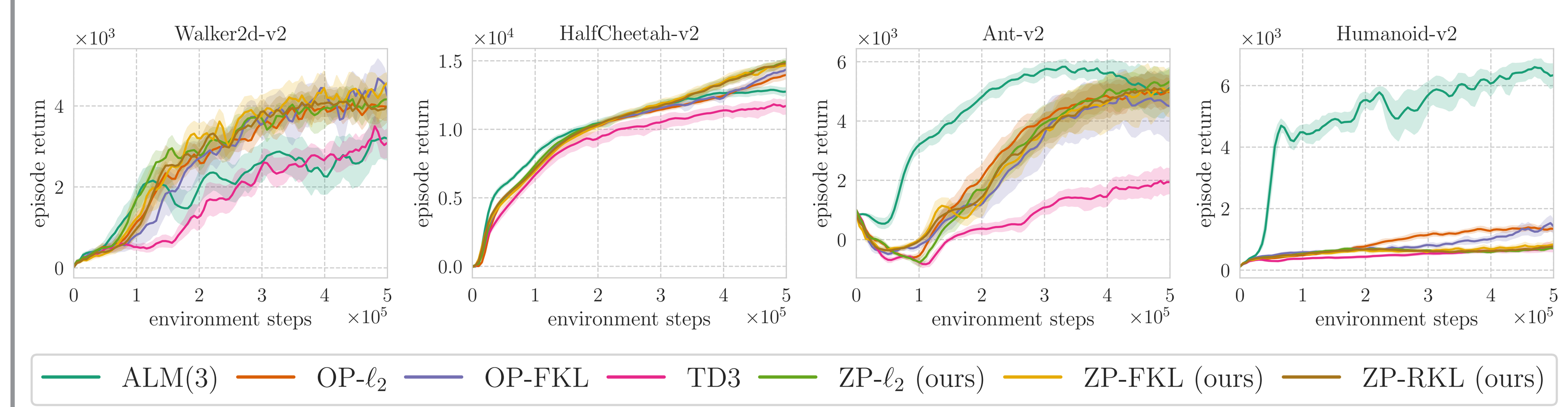
Observation-predictive representations for sparse rewards



MiniGrid: limited view and sparse rewards. Predicting observations provides rich supervision signals.

End-to-end learning (OP, ZP) is better than modular counterpart (RP + OP, RP + ZP) here.

As a baseline: decouple repr. learning from policy optimization



ALM(3)[7] adds intrinsic rewards and SVG-style planning to our algorithm (ZP-RKL). Their main benefit is in the Humanoid task.

References

[1] Li et al. *Towards a unified theory of state abstraction for MDPs*, 2006.
[2] Subramanian et al. *Approximate information state for approximate planning and reinforcement learning in partially observed systems*, 2022.
[3] Givan et al. *Equivalence notions and model minimization in markov decision processes*, 2003.
[4] Kaelbling et al. *Planning and acting in partially observable stochastic domains*, 1998.
[5] Schwarzer et al. *Data-efficient reinforcement learning with self-predictive representations*, 2021.
[6] Tang et al. *Understanding self-predictive learning for reinforcement learning*, 2023.
[7] Ghugare et al. *Simplifying model-based RL: learning representations, latent-space models, and policies with one objective*, 2023.