Understanding Self-Predictive Learning for Reinforcement Learning

Yunhao Tang [1] **Zhaohan Daniel Guo** [1] **Pierre Harvey Richemond** [1] **Bernardo Ávila Pires** [1] **Yash Chandak** [2] **Rémi Munos** [1] **Mark Rowland** [1] **Mohammad Gheshlaghi Azar** [1] **Charline Le Lan** [3] **Clare Lyle** [1] **András György** [1] **Shantanu Thakoor** [1] **Will Dabney** [1] **Bilal Piot** [1] **Daniele Calandriello** [1] **Michal Valko** [1]

Abstract

We study the learning dynamics of self-predictive learning for reinforcement learning, a family of algorithms that learn representations by minimizing the prediction error of their own future latent representations. Despite its recent empirical success, such algorithms have an apparent defect: trivial representations (such as constants) minimize the prediction error, yet it is obviously undesirable to converge to such solutions. Our central insight is that careful designs of the optimization dynamics are critical to learning meaningful representations. We identify that a faster paced optimization of the predictor and semi-gradient updates on the representation, are crucial to preventing the representation collapse. Then in an idealized setup, we show self-predictive learning dynamics carries out spectral decomposition on the state transition matrix, effectively capturing information of the transition dynamics. Building on the theoretical insights, we propose bidirectional self-predictive learning, a novel self-predictive algorithm that learns two representations simultaneously. We examine the robustness of our theoretical insights with a number of small-scale experiments and showcase the promise of the novel representation learning algorithm with large-scale experiments.

1. Introduction

Self-prediction is one of the fundamental concepts in reinforcement learning (RL). In value-based RL, temporal difference (TD) learning (Sutton, 1988) uses the value function prediction at the next time step as the prediction target for the current time step $V(x_t) \leftarrow R(x_t) + \gamma V(x_{t+1})$, a procedure also known as *bootstrapping*. We can understand TD-learning as self-prediction specialized to value learning, where the value function makes predictions about targets constructed from itself.

Recently, the idea of self-prediction has been extended to representation learning with much empirical success (Schwarzer et al., 2021; Guo et al., 2020; 2022). In self-predictive learning, the aim is to learn a representation Φ jointly with a transition function P which models the transition of representations in the latent space $\Phi(x_t) \to \Phi(x_{t+1})$, by minimizing the prediction error

$$\|P\left(\Phi(x_t)\right) - \Phi(x_{t+1})\|_2^2 .$$

Intuitively, minimizing the prediction error should encourage the algorithm to learn a compressed latent representation $\Phi(x)$ of the state x. However, despite the intuitive construct of the prediction error, there is no obvious theoretical justification why minimizing the error leads to meaningful representations at all. Indeed, the *trivial* solution $\Phi(x_t) \equiv c$ for any constant vector c minimizes the error but retains no information at all. Comparing self-predictive learning with value learning, the key difference lies in that the value function is *grounded* in the immediate reward $R(x_t)$. In contrast, the prediction error that motivates self-predictive learning is apparently not grounded in concrete quantities in the environment.

A number of natural questions ensue: how do we reconcile the apparent defect of the prediction error objective, with the empirical success of practical algorithms built on such an objective? What are the representations obtained by self-predictive learning, and are they useful for downstream RL? With obvious theory-practice conflicts in place, it is difficult to establish self-predictive learning as a principled approach to representation learning in general.

We present the first attempt at understanding self-predictive learning for RL, through a theoretical lens. In an idealized setting, we identify key elements to ensure that the self-predictive algorithm avoids collapse and learns meaningful representations. We make the following theoretical and algorithmic contributions.

Key algorithmic elements to prevent collapse. We identify two key algorithmic components: (1) the two time-scale optimization of the transition function P and representation

[1]DeepMind [2]Stanford University [3]University of Oxford. Correspondence to: Yunhao Tang <robintyh@deepmind.com>.

Preliminary work.

Φ; and (2) the semi-gradient update on Φ, to ensure that the representation maintains its capacity throughout learning (Section 3). As a result, self-predictive learning dynamics does not converge to trivial solutions starting from random initializations.

Self-prediction as spectral decomposition. With a few idealized assumptions in place, we show that the learning dynamics locally improves upon a trace objective that characterizes the information that the representations capture about the transition dynamics (Section 3). Maximizing this objective corresponds to spectral decomposition on the state transition matrix. This provides a partial theoretical understanding as to why self-predictive learning proves highly useful in practice.

Bidirectional self-predictive learning. Based on the theoretical insights, we derive a novel self-predictive learning algorithm: bidirectional self-predictive learning (Section 5). The new algorithm learns two representations simultaneously, based on both a forward prediction and a backward prediction. Bidirectional self-predictive learning enjoys more general theoretical guarantees compared to self-predictive learning, and obtains more consistent and stable performance as we validate both on tabular and deep RL experiments (Section 7).

2. Background

Consider a reward-free Markov decision process (MDP) represented as the tuple $(\mathcal{X}, \mathcal{A}, p, \gamma)$ where $\mathcal{X}$ is a finite state space, $\mathcal{A}$ the finite action space, $p : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$ the transition kernel and $\gamma \in [0, 1)$ the discount factor. Let $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ be a fixed policy. For convenience, let $P^\pi : \mathcal{X} \to \mathscr{P}(\mathcal{X})$ be the state transition kernel induced by the policy π. We focus on reward-free MDPs instead of regular MDPs because we do not need reward functions for the rest of the discussion.

Throughout, we assume tabular state representation where each state $x \in \mathcal{X}$ is equivalently encoded as a one-hot vector $x \in \mathbb{R}^{\mathcal{X}}$. This representation will be critical in establishing results that follow. In general, a representation matrix $\Phi \in \mathbb{R}^{|\mathcal{X}| \times k}$ embeds each state $x \in \mathcal{X}$ as a k-dimensional real vector $\Phi^T x \in \mathbb{R}^k$. In practice, we tend to have $k \ll |\mathcal{X}|$ where $|\mathcal{X}|$ is the cardinal of $\mathcal{X}$. The representation is generally shaped by learning signals such as TD-learning or auxiliary objectives. Good representations should entail sharing information between states, and facilitate downstream tasks such as policy evaluation or control.

2.1. Self-predictive learning

We introduce a mathematical framework for analyzing self-predictive learning, which seeks to capture the high level properties of its various algorithmic instantiations (Schwarzer et al., 2021; Guo et al., 2020; 2022). Throughout, assume we have access to state tuples $x, y \in \mathcal{X}$ sampled sequentially as follows,

$$x \sim d, y \sim P^\pi(\cdot|x),$$

where we use $x \sim d$ to denote sampling the state from a distribution defined by the probability vector $d \in \mathbb{R}^{|\mathcal{X}|}$. To model the transition in the representation space $\Phi^T x \to \Phi^T y$, we define $P \in \mathbb{R}^{k \times k}$ as the latent prediction matrix. The predicted latent at the next time step from $\Phi^T x$ is $P^T \Phi^T x$. The goal is to minimize the reconstruction loss in the latent space,

$$\min_{\Phi, P} L(\Phi, P) \coloneqq \mathbb{E}_{x \sim d, y \sim P^\pi(\cdot|x)} \left[\left\| P^T \Phi^T x - \Phi^T y \right\|_2^2 \right]. \tag{1}$$

As alluded to earlier, naively optimizing Equation (1) may lead to trivial solutions such as $\Phi^* = 0$, which also produces the optimal objective $L(\Phi^*, P) = 0$. We next discuss how specific optimization procedures entail learning meaningful representation Φ and prediction function P.

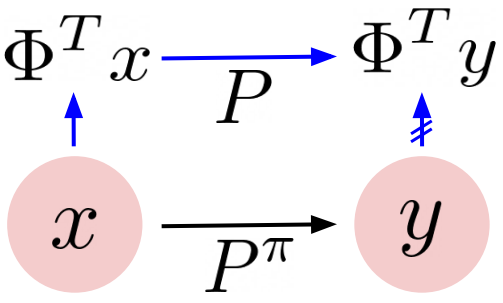


Figure 1. A diagram that outlines the conceptual components of self-predictive learning. The black arrow indicates sampling the transition $y \sim P^\pi(\cdot|x)$. the blue arrow indicates algorithmic components of self-predictive learning: predicting the next state representation $\Phi^T y$ from the first state representation $\Phi^T x$ using prediction matrix P. In practice as shown in Equation (2), self-predictive learning stops the gradient on the prediction target.

3. Understanding learning dynamics of self-predictive learning

Assume the algorithm proceeds in discrete iterations $t \geq 0$. In practice, the update of Φ follows a semi-gradient update through $L(\Phi, P)$ by stopping the gradient via the prediction target $\Phi^T y$

$$\Phi_{t+1} \leftarrow \Phi_t - \eta \nabla_{\Phi_t} \mathbb{E} \left[\left\| P_t^T \Phi_t^T x - \text{sg}\left(\Phi_t^T y\right) \right\|_2^2 \right], \tag{2}$$

where sg stands for stop-gradient and $\eta > 0$ is a fixed learning rate. In the expectation, $x \sim d, y \sim P^\pi(\cdot|x)$ unless otherwise stated.

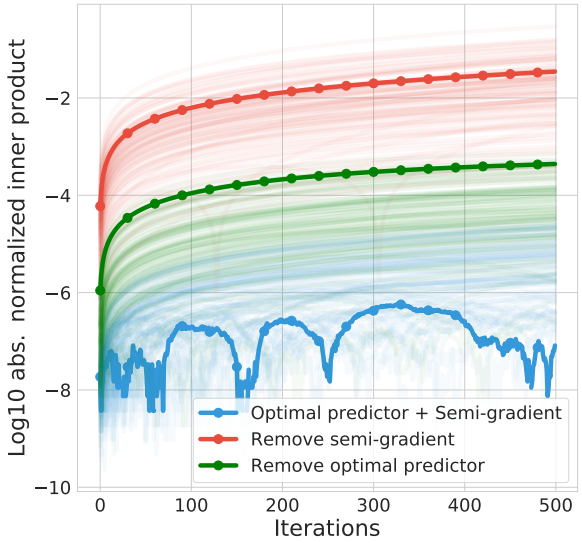


Figure 2. Absolute value of the inner product between the two (normalized) columns of Φ, versus the number of iterations, for different variants of Algorithm 1. Each light curve corresponds to one of 100 independent runs over randomly generated MDPs, and the solid curve shows the median over runs. The experiments are based on the discretized dynamics (Equation (2)) with a small but finite learning rate.

3.1. Non-collapse property of self-predictive learning

We identified a key condition to ensure that the solution does not collapse to trivial solutions, that P_t be optimized at a faster pace than Φ_t. Intriguingly, this condition is compatible with a number of empirical observations made in prior work (e.g., Appendix I of Grill et al., 2020 and Chen and He, 2021 have both identified the importance of near optimal predictors). More formally, the prediction function P_t is computed as one optimal solution to the loss function, fixing the representation Φ_t,

$$P_t \in \arg\min_P \mathbb{E}\left[\left\|P^T\Phi_t^T x - \Phi_t^T y\right\|_2^2\right]. \tag{3}$$

In practice, the above assumption might be approximately satisfied by the fact that the prediction function P_t is often parameterized by a smaller neural network (e.g., an MLP or LSTM) compared to the representation Φ_t (e.g., a deep ResNet, Schwarzer et al., 2021; Guo et al., 2020; 2022), and hence can be optimized at a faster pace even with gradient descent. The pseudocode for such a self-predictive learning algorithm is in Algorithm 1.

To understand the behavior of the joint updates in Equations (2) and (3), we propose to consider the behavior of the corresponding continuous time system. Let $t \geq 0$ be the continuous time index, the ordinary differential equation

Algorithm 1 Self-predictive learning.

Representation matrix $\Phi_0 \in \mathbb{R}^{|\mathcal{X}|\times k}$ for $k \leq |\mathcal{X}|$
for $t = 1, 2...T$ **do**
  Compute prediction matrix P_t based on Equation (3).
  Update representation Φ_t based on Equation (2).
end for
Output final representation Φ_T.

(ODE) systems jointly for (Φ_t, P_t) is

$$\begin{aligned} P_t &\in \arg\min_P L(\Phi_t, P), \\ \dot{\Phi}_t &= -\nabla_{\Phi_t}\mathbb{E}\left[\left\|P_t^T\Phi_t^T x - \mathrm{sg}\left(\Phi_t^T y\right)\right\|_2^2\right]. \end{aligned} \tag{4}$$

Our theoretical analysis consists in understanding the behavior of the above ODE system. The key result shows that the learning dynamics in Equation (4) does not lead to collapsed solutions.

Theorem 1. Under the dynamics in Equation (4), the covariance matrix $\Phi_t^T\Phi_t \in \mathbb{R}^{k\times k}$ is constant over time.

The representation matrix decomposes into k representation vectors $\Phi_t = [\phi_{1,t} \cdots \phi_{k,t}]$, where $\phi_{i,t} \in \mathbb{R}^{\mathcal{X}}$ for each $i = 1, \ldots, k$. Geometrically, we can visualize $(\phi_{i,t})_{i=1}^k$ as forming a basis of a k-dimensional subspace of $\mathbb{R}^{\mathcal{X}}$. Throughout the learning process, all basis vectors rotate in the same direction, keeping the relative angles between basis vectors and their lengths unchanged. As a direct implication of the *rotation* dynamics, it is not possible for $(\phi_{i,t})_{i=1}^k$ to start as different vectors but then converge to the same vector. That is, the representation cannot collapse.

Corollary 2. Under the dynamics in Equation (4), the representation vectors $(\phi_{i,t})_{i=1}^k$ cannot converge to the same vector if they are initialized differently.

With the non-collapse behavior established under the dynamics in Equation (4), we ask in hindsight what elements of the algorithm entail such a property. In addition to the faster paced optimization of the prediction matrix P_t, the semi-gradient update to Φ_t is also indispensable. Our result above provides a first theoretical justification to the "latent bootstrapping" technique in the RL case, which has been effective in empirical studies (Grill et al., 2020; Schwarzer et al., 2021; Guo et al., 2020). See Section 6 for more discussions on the relation between our analysis and prior work in the non-contrastive unsupervised learning algorithms.

We illustrate the importance of the optimality of P_t and the semi-gradient update with an empirical study. We learned representations with $k = 2$ vectors on randomly generated MDPs using three baselines: (1) using semi-gradient updates on Φ_t and with optimal predictors P_t, which strictly adheres to Algorithm 1; (2) using the optimal predictors P_t

but replacing the semi-gradient update by a full gradient on Φ_t, i.e., allowing gradients to flow into $\Phi^T y$; (3) using the semi-gradient update but corrupting the optimal predictor with some zero-mean noise at each iteration.

Figure 2 shows the absolute value of the inner product between the two (normalized) columns of Φ, also known as the cosine similarity, versus the number of iterations.

The matrix Φ is initialized to be orthogonal, so we expect the two columns in Φ to remain close to orthogonal throughout the learning dynamics (Equation (2)) with a small but finite learning rate η. Therefore, the larger values the curves take in Fig. 2, the stronger the evidence of collapse, and indeed as we claimed, when either semi-gradient or optimal predictor are removed from the learning algorithm, the representation columns start to collapse. In practice, having an optimal predictor is a stringent requirement; we carry out more extensive ablation study in Section 7. See Appendix G for more details on the tabular experiments.

Non-collapse property for a general loss function. To make our analysis simple, we focused on the squared loss function between the prediction $P^T \Phi^T x$ and target $\Phi^T y$. We note that the non-collapse property in Theorem 6 holds more generally for losses of the form $L(\Phi, P)$. Our result also applies to a slightly modified variant of the cosine similarity loss, which is more commonly used in practice (Grill et al., 2020; Chen and He, 2021; Schwarzer et al., 2021; Guo et al., 2022). See Appendix C for more details.

The non-collapse property shows that the covariance matrix $\Phi_t^T \Phi_t$, which measures the level of diversity across representation vectors $(\phi_{i,t})_{i=1}^t$ is conserved. In the next section, we will discuss the connection between the learning dynamics and spectral decomposition on the transition matrix P^π. To facilitate the discussion, we make an idealized assumption that the representation columns are initialized orthonormal.

Assumption 3. (**Orthonormal Initialization**) The representations are initialized orthonormal: $\Phi_0^T \Phi_0 = I_{k \times k}$.

Note that the assumption is approximately valid e.g., when entries of Φ_0 are sampled i.i.d. from an isotropic distribution and properly scaled, and when the state space is large relative to the representation dimension $k \ll |\mathcal{X}|$ (see, e.g., (González-Guillén et al., 2018) as a related reference). This is the case if the representation capacity is smaller than the state space (e.g., small network vs. complex image observation).

3.2. Self-predictive learning as eigenvector decomposition

For simplicity, henceforth, we assume a uniform distribution over the first-state. This assumption is made implicitly in a number of prior work on TD-learning or representation learning for RL (Parr et al., 2008; Song et al., 2016; Behzadian et al., 2019; Lyle et al., 2021).

Assumption 4. (**Uniform distribution**) The first-state distribution is uniform: $d = |\mathcal{X}|^{-1} 1_{|\mathcal{X}|}$.

For the rest of the paper, we always assume Assumption 3 and Assumption 4 to hold. As a result, the learning dynamics in Equation (4) reduces to the following:

$$P_t = \Phi_t^T P^\pi \Phi_t, \quad \dot{\Phi}_t = \left(I - \Phi_t \Phi_t^T\right) P^\pi \Phi_t (P_t)^T \tag{5}$$

We provide detailed derivations of the ODE in Appendix A.

Remarks. As a sanity check and a special case, assume the representation matrix is the identity $\Phi_t = I$, in which case $\Phi_t^T x = x$ recovers the tabular representation. In this case we have $P_t = P^\pi$ and the latent prediction recovers the original state transition matrix.

A useful property of the above dynamical system is the set of critical points where $\dot{\Phi}_t = 0$. Below we provide a characterization of such critical points.

Lemma 5. Assume P^π is real diagonalizable and let $(u_i)_{i=1}^{|\mathcal{X}|}$ be its set of $|\mathcal{X}|$ distinct eigenvectors. Let $\mathcal{C}_{P^\pi}$ be the set of critical points of Equation (5). Then $\mathcal{C}_{P^\pi}$ contains all matrices whose columns are orthonormal, and have the same span as a set of k eigenvectors.

Remarks on the critical points. Lemma 5 implies that when P^π only has eigenvectors, any matrix consisting of a subset of k eigenvectors $(u_{i_j})_{j=1}^k$ (as well as any set of k orthonormal columns with the same span) is a critical point to the self-predictive dynamics in Equation (5). However, this does not mean that $\mathcal{C}_{P^\pi}$ only consists of such critical points. As a simple example, consider the transition matrix

$$P^\pi = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \tag{6}$$

and when $k = 1$, in which case Φ_t is a 2-d vector. In addition to the two eigenvectors of P^π, there are at least four other non-eigenvector critical points (shown in Fig. 3). See Appendix D for more detailed derivations. We leave a more comprehensive study of such critical points in the general case to future work. When P^π has complex eigenvectors, the structure of the critical points to Equation (5) also becomes more complicated.

Importantly, under the assumption in Lemma 5, not all critical points are equally informative. Arguably, the top k eigenvectors of P^π with the largest absolute valued eigenvalues, should contain the most information about the matrix because they reflect the high variance directions in the one-step transition. This is the motivation behind compression algorithms such as PCA. We now show that when P^π is symmetric, intriguingly, the learning dynamics maximizes a

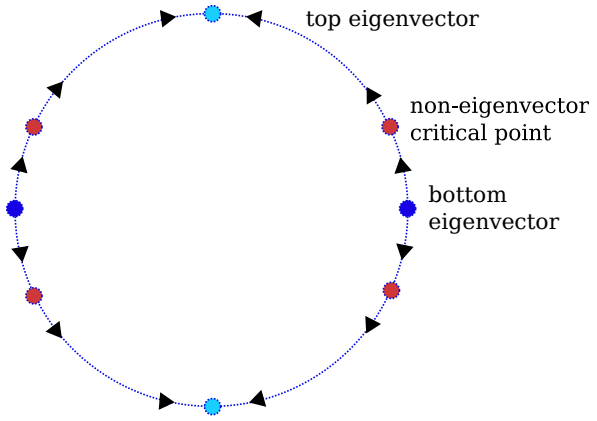


Figure 3. Critical points and local dynamics of the example MDP in Equation (6). We consider $k = 1$ so representations Φ_t are 2-d unit vectors. There are four eigenvector critical points (light and dark blue) and four non-eigenvector critical points (red) of the ODE, shown on the unit circle. The black arrows show the local update direction based on the ODE. Initialized near the bottom eigenvector, the dynamics converges to one of the four non-eigenvector critical points and not to the top eigenvector. See Appendix D for more detailed explanations.

trace objective that measures the variance information contained in P^π, and that this objective is maximized by the top k eigenvectors.

Theorem 6. If P^π is symmetric, then under Assumption 3 and learning dynamics Equation (5), the trace objective is non-decreasing $\dot{f} \geq 0$, where

$$f(\Phi_t) \coloneqq \text{Trace}\left(\left(\Phi_t^T P^\pi \Phi_t\right)^T \left(\Phi_t^T P^\pi \Phi_t\right)\right).$$

If $\Phi_t \notin \mathcal{C}_{P^\pi}$, then $\dot{f} > 0$. Under the constraint $\Phi^T\Phi = I$, the maximizer to $f(\Phi)$ is any set of k orthonormal vectors which span the principal subspace, i.e., with the same span as the k eigenvectors of P^π with top absolute eigenvalues.

To see that the trace objective f measures useful information contained in P^π, for now let us constrain the arguments to f to be the set of k eigenvectors $(u_{i_j})_{j=1}^k$ of P^π. In this case, $f([u_{i_1}...u_{i_k}]) = \sum_{j=1}^k \lambda_{i_j}^2$ is the sum of the corresponding squared eigenvalues. This implies f is a useful measure on the spectral information contained in P^π.

Theorem 6 also shows that as long as Φ_t is not at a stationary point contained in $\mathcal{C}_{P^\pi}$, the trace objective $f(\Phi_t)$ makes strict improvement over time. Equivalently, this means Φ_t has the tendency to move towards representations with high trace objective, e.g., subspaces of eigenvectors with high trace objective. In other words, we can understand the dynamics of Φ_t as principal subspace PCA on the transition matrix.

Remarks on the convergence. Thus far, there is no guarantee that Φ_t converges to the top k eigenvectors as in general there is a chance that the dynamics converges other critical points. We revisit the simple example in Equation (6), where in Fig. 3 we mark all critical points on the unit circle (four eigenvector and four non-eigenvector critical points). When initialized near the bottom eigenvector, the dynamics converges to one of the non-eigenvector critical points instead of the top eigenvector.

Nevertheless, the local improvement property of the learning dynamics can be very valuable in large-scale environments. We leave a more refined study on the convergence properties of self-predictive learning dynamics to future work.

Remarks on the case with non-symmetric P^π. Theorem 6 is obtained under the idealized assumption that P^π is symmetric. In general, when P^π is non-symmetric, the improvement in the trace objective $f(\Phi_t)$ is not necessarily monotonic. In fact, it is possible to find instances of Φ_t where the dynamics decreases the trace objective f. A plausible explanation is that since the dynamics of Φ_t can be understood as gradient-ascent based PCA on P^π, the PCA objective is only well defined when the data matrix P^π is symmetric. Motivated by the limitation of the self-predictive learning and its connection to PCA, we propose a novel self-predictive algorithm with two representations. We will introduce such a method in Section 5 and reveal how it generalizes self-predictive learning to carrying out SVD instead of PCA on P^π.

Before moving on, we empirically assess how much impact that the level of symmetry of P^π has on the trace maximization property. We carried out simulations on 100 randomly generated tabular MDPs, by unrolling the exact ODE dynamics in Equation (4) and measured the evolution of the trace objective $f_t = \text{Trace}\left((\Phi_t^T P^\pi \Phi_t)^T \Phi_t^T P^\pi \Phi_t\right)$. Figure 4 shows the ratio between the trace objective f_t and the the value of the objective for the top k eigenvectors of P^π, versus the number of training iterations t.

Fig. 4 shows that when P^π is symmetric, the trace objective smoothly improves over time, as predicted by theory. When P^π is non-symmetric, the improvement in trace objective is not guaranteed to be monotonic and, indeed, this is the case for some runs. However in our experiments, over time, the objective improved by a large margin compared to initialization, though not necessarily converging to the maximum possible values. The numerical evidence shows that the learning dynamics can still capture useful information about the transition dynamics for certain non-symmetric MDPs. It is, however, possible to design non-symmetric P^π on which self-predictive dynamics barely increases the trace objective (see Section 5). Appendix G contains additional results and more details about the experiment details.

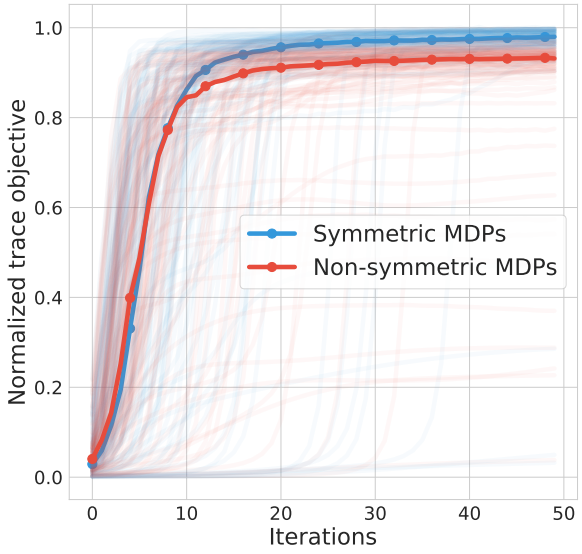


Figure 4. Ratio between the trace objective $f_t = \text{Trace}\left((\Phi_t^T P^\pi \Phi_t)^T \Phi_t^T P^\pi \Phi_t\right)$ and the value of the objective for the top k eigenvectors of P^π, versus the number of training iterations. Each light curve corresponds to one of 100 independent runs over randomly generated MDPs, and the solid curve shows the median over runs. The experiments are based on the exact ODE dynamics in Equation (5).

4. Extensions of the theoretical analysis

We have chosen to analyze the learning dynamics under arguably the simplest possible model setup. This helps elucidate important features about the self-predictive learning dynamics, but also leaves room for extensions. We discuss a few possibilities.

Additional prediction function. Practical algorithms such as SPR (Schwarzer et al., 2021) and BYOL-RL (the representation learning component of BYOL-Explore (Guo et al., 2022)) usually employ an additional prediction function, on top of the prediction function P which models the latent transition dynamics. In our framework, this can be modeled as an additional prediction matrix $Q \in \mathbb{R}^{k\times k}$ with the overall loss function as follows

$$\mathbb{E}\left[\left\|Q^T P^T \Phi^T x - \Phi^T y\right\|_2^2\right].$$

The roles of P and Q are different. P is meant to model the latent transition dynamics, while Q provides extra degrees of freedom to match the predicted latent $Q^T P^T \Phi^T x$ to the next state representation $\Phi^T y$. Though such a combination of Q, P seems redundant at the first sight, the extra flexibility entailed by the additional prediction proves very important in practice (Q is usually implemented as a MLP on top of the output of P, which is implemented as a LSTM (Schwarzer et al., 2021; Guo et al., 2020)). Our theoretical result can be extended to this case by treating the composed matrix PQ as a whole during optimization, to ensure the non-collapse of Φ.

Multi-step and action-conditional latent prediction. In practice, making multi-step predictions significantly improves the performance (Schwarzer et al., 2021; Guo et al., 2020; 2022). In our framework, this can be understood as the loss function

$$\mathbb{E}_{x_n\sim(P^\pi)^n(\cdot|x_0)}\left[\left\|P^T\Phi^T x_0 - \Phi^T x_n\right\|_2^2\right].$$

In this case, our result in Section 3 suggests that the self-prediction carries out spectral decomposition on the n-step transition $(P^\pi)^n$. Another important practical component is that latent transition models are usually action-conditional. In the one-step case, this can be understood as parameterizing multiple prediction matrices and representation matrices $(P_a, \Phi_a)_{a\in\mathcal{A}}$ and . The loss function naturally becomes

$$\mathbb{E}_{x_n\sim P^\pi(\cdot|x_0,a),a\sim\pi(\cdot|x_0)}\left[\left\|P_a^T\Phi_a^T x_0 - \Phi_a^T x_n\right\|_2^2\right].$$

The latent prediction matrix P_a and representation Φ_a effectively carry out spectral decomposition on the Markov matrix P^{π_a}, which is the transition matrix of policy π_a that takes action a in all states.

Partial observability. In many practical applications, the environment is better modeled as a partially observable MDP (POMDP; (Cassandra et al., 1994)). As a simplified setup, consider at time t the agent has access to the current history $h_t = (o_s)_{s\le t} \in \mathcal{H}$ which consists of observations $o_s \in \mathcal{O}$ in past time steps. Fixing the agent's policy $\pi : \mathcal{H} \to \mathcal{P}(\mathcal{A})$, let $\tilde{P}^\pi(h'|h)$ denote the distribution over the next observed history h' given h. Drawing direct analogy to the MDP case, one possible loss function is

$$\mathbb{E}_{h'\sim\tilde{P}^\pi(\cdot|h)}\left[\left\|P^T\Phi^T h - \Phi^T h'\right\|_2^2\right].$$

In practice, $\Phi \in \mathbb{R}^{|\mathcal{H}|\times k}$ is often implemented as a recurrent function such as LSTM (see, e.g., BYOL-RL (Guo et al., 2020) as one possible implementation and find its details in Appendix E), to avoid the explosion in the size of the set of all histories $|\mathcal{H}|$. Under certain conditions, our analysis can be extended to spectral decomposition on the history transition matrix $P^\pi(h'|h)$. However, given recent empirical advances achieved by self-predictive learning algorithm in partially observable environments (Guo et al., 2022), potentially a more refined analysis is valuable in better bridging the theory-practice gap in the POMDP case.

Finite learning rate and other factors. Our analysis and result heavily rely on the assumption of continuous time dynamics. In practice, updates are carried out on discrete time

Algorithm 2 Bidirectional self-predictive learning.

Representation matrix $\Phi_0\tilde{\Phi}_0 \in \mathbb{R}^{|\mathcal{X}|\times k}$ for $k \leq |\mathcal{X}|$
for $t = 1, 2...T$ **do**
  Compute optimal forward and backward prediction matrix $(P_t, \tilde{P}_t)$ based on Equation (8).
  Update representations $(\Phi_t, \tilde{\Phi}_t)$ based on Equation (7).
end for
Output final representation $(\Phi_T, \tilde{\Phi}_T)$.

steps with a finite learning rate. Through experiments, we observe that representations tend to partially collapse when learning rates are finite, though they still manage to capture spectral information about the transition matrix. We also study the impact of other factors such as non-optimal prediction matrix and delayed target network, see Appendix G for such ablation study. Formalizing such results in theory would be an interesting future direction.

5. Bidirectional self-predictive learning with left and right representations

Thus far, we have established a few important properties of the self-predictive learning dynamics. However, we have alluded to the fact that self-predictive learning dynamics can ill-behave in certain cases.

With insights derived from previous sections, we now introduce a novel self-predictive learning algorithm that makes use of two representations $\Phi_t, \tilde{\Phi}_t \in \mathbb{R}^{|\mathcal{X}|\times k}$ and two latent prediction matrices $P, \tilde{P} \in \mathbb{R}^{k\times k}$. We refer to Φ_t as the *left representation* and $\tilde{\Phi}_t$ the *right representation*, for reasons that will be clear shortly. With Φ_t, we make forward prediction through P, using prediction target computed from $\tilde{\Phi}_t$; with $\tilde{\Phi}_t$, we make backward prediction through $\tilde{P}$, using prediction target computed from Φ_t. Both representations follow semi-gradient updates:

$$\begin{aligned}\Phi_{t+1} &\leftarrow \Phi_t - \eta\nabla_{\Phi_t}\mathbb{E}\left[\left\|P_t^T\Phi_t^T x - \mathrm{sg}\left(\tilde{\Phi}_t^T y\right)\right\|_2^2\right],\\ \tilde{\Phi}_{t+1} &\leftarrow \tilde{\Phi}_t - \eta\nabla_{\tilde{\Phi}_t}\mathbb{E}\left[\left\|\tilde{P}_t^T\tilde{\Phi}_t^T y - \mathrm{sg}\left(\Phi_t^T x\right)\right\|_2^2\right].\end{aligned} \tag{7}$$

Similar to the analysis before, we assume $P_t, \tilde{P}_t$ are optimally adapted to the representations, by exactly solving the forward and backward least square prediction problems. This is a key requirement to ensure non-collapse in the new learning dynamics (similar to the self-predictive dynamics in Equation (5)). The pseudocode for the bidirectional self-predictive learning algorithm is in Algorithm 2.

For notational simplicity, we denote the forward and backward prediction losses as $L_\mathrm{f}(\Phi, P)$ and $L_\mathrm{b}(\tilde{\Phi}, \tilde{P})$ respectively. The continuous time ODE system for the joint variable $(P_t, \tilde{P}_t, \Phi_t, \tilde{\Phi}_t)$ is

$$\begin{aligned}P_t &\in \arg\min_P L_\mathrm{f}(\Phi_t, P),\\ \dot{\Phi}_t &= -\nabla_{\Phi_t}\mathbb{E}\left[\left\|P^T\Phi^T x - \mathrm{sg}\left(\tilde{\Phi}^T y\right)\right\|_2^2\right],\\ \tilde{P}_t &\in \arg\min_{\tilde{P}} L_\mathrm{b}(\tilde{\Phi}_t, \tilde{P}),\\ \dot{\tilde{\Phi}}_t &= -\nabla_{\tilde{\Phi}_t}\mathbb{E}\left[\left\|\tilde{P}^T\tilde{\Phi}^T y - \mathrm{sg}\left(\Phi^T x\right)\right\|_2^2\right].\end{aligned} \tag{8}$$

Similar to Theorem 1, the non-collapse property for both the left and right representations follows.

Theorem 7. Under the bidirectional self-predictive learning dynamics in Equation (8), the covariance matrices $\Phi_t^T\Phi_t \in \mathbb{R}^{k\times k}$ and $\tilde{\Phi}_t^T\tilde{\Phi}_t \in \mathbb{R}^{k\times k}$ are both constant matrices over time.

As before, to simplify the presentation, we make the assumption that both left and right representations are initialized orthonormal (cf. Assumption 3):

Assumption 8. (**Orthonormal Initialization**) The left and right representations are both initialized orthonormal $\Phi_0^T\Phi_0 = \tilde{\Phi}_0^T\tilde{\Phi}_0 = I_{k\times k}$.

5.1. Bidirectional self-predictive learning as singular value decomposition

In self-predictive learning, the forward prediction derives from the fact that the forward process $x \to y$ follows from a Markov chain. Following a similar argument, for the backward prediction to be sensible, we need to ensure that the reverse process $y \to x$ is also a Markov chain. Technically, this means we require $(P^\pi)^T$ to be a transition matrix too, which models the backward transition process. Importantly, this is a much weaker assumption than P^π be symmetric, as required by the self-predictive learning dynamics (Theorem 6).

Assumption 9. P^π is a doubly stochastic matrix, i.e., $(P^\pi)^T$ is also a transition matrix.

Under assumptions above, the learning dynamics in Equation (8) reduces to the following set of ODEs:

$$\begin{aligned}P_t &= \Phi_t^T P^\pi \tilde{\Phi}_t, \;\; \dot{\Phi}_t = \left(I - \Phi_t\Phi_t^T\right)P^\pi\tilde{\Phi}_t(P_t)^T\\ \tilde{P}_t &= \tilde{\Phi}_t^T\left(P^\pi\right)^T\Phi_t, \;\; \dot{\tilde{\Phi}}_t = \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right)\left(P^\pi\right)^T\Phi_t(\tilde{P}_t)^T\end{aligned} \tag{9}$$

Let $P^\pi = U\Sigma V^T$ be the singular value decomposition (SVD) of P^π, where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2...\sigma_{|\mathcal{X}|})$ is a diagonal matrix with non-negative diagonal entries. We call any i-th column of U and V, denoted as (u_i, v_i), a singular vector pair. As before, we start by examining the critical points of the bidirectional self-predictive dynamics.

Lemma 10. Let $\tilde{\mathcal{C}}_{P^\pi} \subset \mathbb{R}^{|\mathcal{X}|\times k} \times \mathbb{R}^{|\mathcal{X}|\times k}$ be the set of critical points to Equation (9). Then $\tilde{\mathcal{C}}_{P^\pi}$ contains any pair of matrices, whose columns are orthonormal and have the same span as k of singular vector pairs.

Lemma 10 implies that any k pairs of SVD vectors are a critical point to the learning dynamics. However, similar to the self-predictive learning dynamics, not all critical points are equally informative. We propose a SVD trace objective $\tilde{f}(\Phi_t, \tilde{\Phi}_t)$, which measures the information contained in k singular vector pairs. Interestingly, the bidirectional self-predictive learning dynamics locally improves such an objective.

Theorem 11. Under Assumption 8 and the learning dynamics in Equation (9), the following SVD trace objective is non-decreasing $\dot{\tilde{f}} \geq 0$, where

$$\tilde{f}\left(\Phi_t, \tilde{\Phi}_t\right) := \text{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)\right).$$

If $(\Phi_t, \tilde{\Phi}_t) \notin \tilde{\mathcal{C}}_{P^\pi}$, then $\dot{\tilde{f}} > 0$. Under the constraint $\Phi^T\Phi = \tilde{\Phi}^T\tilde{\Phi} = I$, the maximizer to $\tilde{f}(\Phi, \tilde{\Phi})$ is any two sets of k orthonormal vectors with the same span as the k singular vector pairs of P^π with top singular values.

To verify that the SVD trace objective $\tilde{f}$ provides an information measure on the representation vectors $(\Phi_t, \tilde{\Phi}_t)$, we constrain arguments of $\tilde{f}$ to be the set of k singular vector pairs $(u_{i_j}, v_{i_j})_{j=1}^k$ of P^π, then $\tilde{f}\left([u_{i_1}...u_{i_k}], [v_{i_1}...v_{i_k}]\right) = \sum_{j=1}^k \sigma_{i_j}^2$ is the sum of the corresponding squared singular values. The top k singular vector pairs maximize this objective, and hence contain the most information about P^π based on this measure.

Theorem 11 shows that as long as either one of the two representations are not at the critical points, i.e., $\dot{\Phi}_t \neq 0$ or $\dot{\tilde{\Phi}}_t \neq 0$, the SVD trace objective $\tilde{f}(\Phi_t, \tilde{\Phi}_t)$ is being strictly improved under the bidirectional self-predictive learning dynamics. Equivalently, this implies the left and right representations $(\Phi_t, \tilde{\Phi}_t)$ tend to move towards singular vector pairs with high SVD trace objective, i.e., seeking more information about the transition dynamics.

Two representations vs. one representation. Looking beyond the ODE analysis, we explain why having two separate representations are inherently important for representation learning in general. For a transition matrix P^π, its left and right singular vectors in general differ. bidirectional self-predictive learning provides the flexibility to learn both left and right singular vectors in parallel, without having to compromise their differences. On the other hand, a single representation will need to interpolate between left and right singular vectors, which may lead to non-monotonic behavior in the trace objective as alluded to earlier.

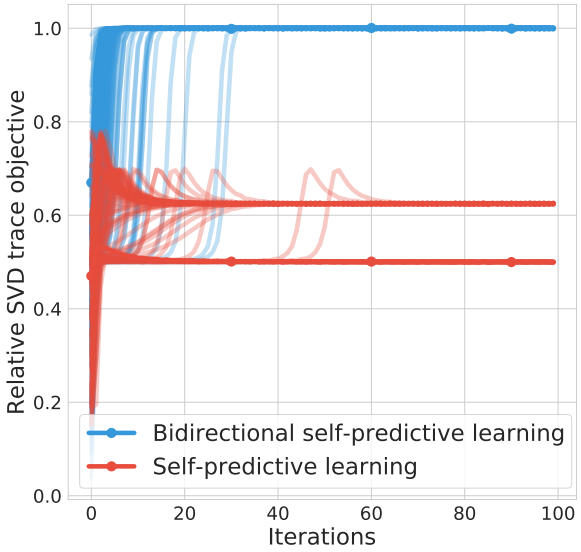


Figure 5. Ratio between the trace objectives (f_t in red for self-predictive learning, following Equation (4); and $\tilde{f}_t$ in blue for bidirectional self-predictive learning, following Equation (8)) and the value of $\tilde{f}$ for the top k singular vector pairs of P^π, versus the number of training iterations. Each light curve corresponds to one of 100 independent runs over random orthonormal initializations of Φ on the same MDP designed so that the left and right singular vectors of P^π are very different. The solid curve shows the median over runs.

Consider a very simple transition matrix with $|\mathcal{X}| = 3$ states that illustrate the failure mode of the self-predictive learning dynamics with a single representation,

$$P^\pi = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}.$$

By construction, its top left and right singular vectors differ greatly. We simulated the self-predictive learning dynamics (with a single representation, Equation (4)) and the bidirectional self-predictive learning dynamics (Equation (8)) in an MDP with this transition matrix, and measured the evolution of the two trace objectives, f_t and $\tilde{f}_t$. Figure 5 shows the ratio between the trace objectives and maximum value of $\tilde{f}$ obtained at the top k singular vector pairs, versus the number of training iterations t. The bidirectional self-predictive learning improves the objective steadily over time. In contrast, the single representation dynamics mostly halt at the initialized value, due to the limited capacity of one representation to combine two highly distinct singular vectors. See more details in Appendix F.

In addition to the improved stability shown in the example above, bidirectional self-predictive learning also captures more comprehensive information about the transition dynamics, compared to a single representation. While it is

known that left singular vectors of P^π can approximate value functions V^π with provably low errors for general transition matrix (Behzadian et al., 2019), it is challenging to learn the left singular vectors as standalone objects. Bidirectional self-predictive learning captures both left and right representations at the same time, entailing a better approximation to both left and right top singular vectors. In large-scale experiments, since bidirectional self-predictive learning consists of both forward and backward predictions, it can provide richer signals for representation learning when combined with nonlinear function approximation (Section 7).

6. Prior Work

Non-collapse mechanism of self-predictive learning dynamics. Self-prediction based RL representation learning algorithms were partly inspired from the non-contrastive unsupervised learning algorithm (Grill et al., 2020; Chen and He, 2021). A number of prior work attempt to understand the non-collapse learning dynamics of such algorithms, through the roles of semi-gradient and regularization (Tian et al., 2021), prediction head (Wen and Li, 2022) and data augmentation (Wang et al., 2021). Although our analysis is specialized to the RL case, the non-collapse mechanism (Theorem 6) is qualitatively different from prior work. Such a result is potentially useful in understanding the behavior of unsupervised learning as well.

RL representation learning via spectral decomposition. One primary line of research in representation learning for RL is via spectral decomposition of the transition matrix P^π or successor matrix $(I - \gamma P^\pi)^{-1}$. These methods are generally categorized as: (1) eigenvector-decomposition based approach, which typically assumes symmetry or real diagonizability of P^π (Mahadevan, 2005; Machado et al., 2018; Lyle et al., 2021); (2) SVD-based approach, which is more generally applicable (Behzadian et al., 2019; Ren et al., 2022) and shows theoretical benefits to downstream RL tasks. Our work draws the connections between spectral decomposition and more empirically oriented RL algorithm, such as SPR (Schwarzer et al., 2021), PBL (Guo et al., 2020) and BYOL-RL (Guo et al., 2022), and is one step in the direction of formally chracterizing high performing representation learning algorithms.

Forward-backward representations. Closely related to bidirectional self-predictive learning is the forward-backward (FB) representations (Touati and Ollivier, 2021; Blier et al., 2021). By design, the forward representation learns value functions and backward representation learns visitation distributions. This design bears close connections to the left and right singular vectors of the transition matrix P^π, which bidirectional self-predictive learning seeks to approximate. Despite the high level connection, bidirectional self-predictive learning is purely based on self-prediction, and hence has much simpler algorithmic design.

Algorithms for PCA and SVD with gradient-based update. The ODE systems in Equations (5) and (8) bear close connections to ODE systems used for studying gradient-based incremental algorithms for PCA and SVD of empirical covariance matrices in classical unsupervised learning. Example algorithms include Oja's subspace algorithm (Oja and Karhunen, 1985; Oja, 1992) and its extension to SVD (Diamantaras and Kung, 1996; Weingessel and Hornik, 1997). A primary historical motivation for such algorithms is that they entail computing top k eigenvectors or singular vectors with incremental gradient-based updates. This echoes with the observation we make in this paper, that self-predictive learning dynamics can be understood as gradient-based spectral decomposition on the transition matrix.

7. Deep RL implementation

We considered the single representation self-predictive learning dynamics as an idealized theoretical framework that aims to capture some essential aspects of a number of existing deep RL representation learning algorithms (Schwarzer et al., 2021; Guo et al., 2020). Importantly, our theoretical analysis suggests that we can get more expressive representations by leveraging the bidirectional self-predictive learning dynamics in Section 5.

Inspired by the theoretical discussions, we introduce the *deep bidirectional self-predictive learning* algorithm for representation learning for deep RL. We build the deep bidirectional self-predictive learning algorithm on top of the representation learning used in BYOL-Explore (Guo et al., 2022). While BYOL-Explore uses the prediction loss as a signal to drive exploration, we do not use such exploration bonuses in this work and focus only on the effect of representation learning.

We now provide a concise summary of how BYOL-RL works and how it is adapted for bidirectional self-predictive learning. In general partially observable environments, BYOL-RL encodes a history of observations $h_t = (f(o_s))_{s \le t}$ into its latent representation $\Phi(h_t) \in \mathbb{R}^k$ through a convnet $f : \mathcal{O} \to \mathbb{R}^k$ and a LSTM. Then, the algorithm constructs a multi-step forward prediction $p\left(\Phi(h_t), a_{t:t+n-1}\right))$ with an open loop LSTM $p : \mathbb{R}^k \times (\mathcal{A})^n \to \mathbb{R}^d$. This forward prediction is against a back-up target computed at the n-step forward future time step $f(o_{t+n})$. The bidirectional self-predictive learning algorithm hints at a backward latent self-prediction objective, i.e., predicting the past latent observations based on future representations. In a nutshell, for deep bidirectional self-predictive learning, we implement the backward prediction

in the same way as forward prediction, but with a *reversed* time axis. See Appendix E for more technical details on BYOL-RL and how it is adapted for backward predictions.

A related but different idea has been adopted in PBL (Guo et al., 2019), where they propose a reverse prediction that matches $\tilde{f}(o_{t+n})$ for some function $\tilde{f} : \mathcal{O} \to \mathbb{R}^k$ against the encoded representation $\Phi(h_{t+n})$. The design of PBL is motivated by partial observability, and does not carry out backward predictions over time. See Appendix E for details on PBL and how it differs significantly from deep bidirectional self-predictive learning.

7.1. Experiments

We compare the deep bidirectional self-predictive learning algorithm with BYOL-RL (Guo et al., 2020). BYOL-RL is built on V-MPO (Song et al., 2020), an actor-critic algorithm which shapes the representation using policy gradient, without explicit representation learning objectives.

Our testbed is DMLab-30, a collection of 30 diverse partially observable cognitive tasks in the 3D DeepMind Lab (Beattie et al., 2016). We consider the multi-task setup where the agent is required to solve all 30 tasks simultaneously. Representation learning has proved highly valuable in such a setting (Guo et al., 2019). See Fig. 11 in Appendix G for results on how BYOL-RL significantly improves over baseline RL algorithm per each game.

In Fig. 6, we show the per-game improvement of deep bidirectional self-predictive learning over BYOL-RL. In aggregate across all 30 tasks, the two algorithms perform comparably as they either under-perform or over-perform each other in a similar number of games. We measure the performance of each task with the human normalized performance $(z_i - u_i)/(h_i - u_i)$ with $1 \leq i \leq 30$, where u_i and h_i are the raw score performance of random policy and humans. A normalized score of 1 indicates that the agent performs as well as humans on the task. Interestingly, there are a number of tasks on which backward predictions significantly improve over the baseline by as much as 0.4 human normalized score. This comparison shows the promise of the deep bidirectional self-predictive learning algorithm when combined with deep RL agents.

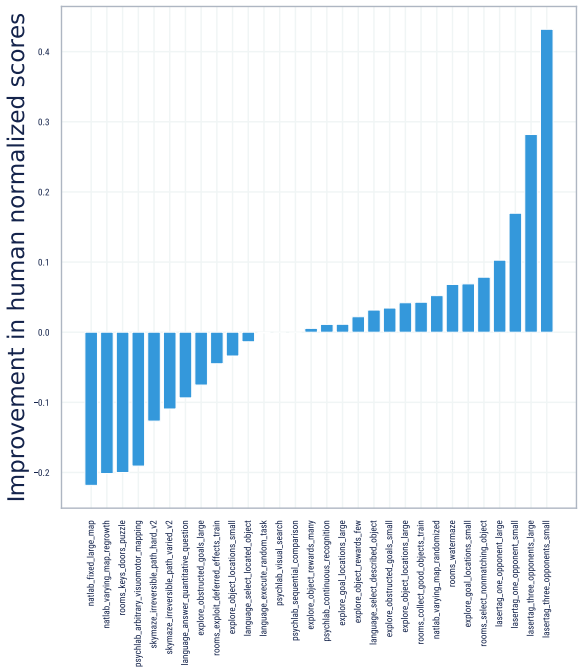


Figure 6. Per-game improvement of bidirectional self-predictive learning compared to baseline BYOL-RL, in terms of mean human normalized scores averaged across 3 seeds. The scores are obtained at the end of training. Bidirectional self-predictive learning provides noticeable gains over BYOL-RL in certain cases, while it degrades performance in some others. See Appendix G for details.

8. Conclusion

In this work, we have presented a new way to understand the learning dynamics of self-predictive learning for RL. By identifying key algorithmic components to guarantee non-collapse, and drawing close connections between self-prediction and spectral decomposition of the transition dynamics, we have provided a first justification to the practical efficacy of a few empirically motivated algorithms. Our insights also naturally led to a novel representation learning algorithm, which mirrors SVD-based algorithms and enjoys more flexibility than self-prediction with a single representation. A deep RL instantiation of the algorithm has also proved empirically comparable to state-of-the-art performance.

Our results open up many interesting avenues for future research. A few examples: in hindsight, many assumption are due to the limitations of linear models; it will be of significant interest to study non-linear latent predictions and representations. Another interesting direction would be to study how self-prediction interacts with RL algorithms, such as TD-learning or policy gradient algorithms. Lastly, our analysis also provides insights to the unsupervised learning case, and hopefully motivates more investigation along this direction in the space.

Acknowledgements. Many thanks to Jean-Bastien Grill and Csaba Szepesvári for their feedback on an earlier draft of this paper. We are appreciative of the collaborative research environment within DeepMind. The experiments in this paper were conducted using Python 3, and made heavy use of NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020a), Matplotlib (Hunter, 2007), Reverb (Cassirer et al., 2021), JAX (Bradbury et al., 2018) and the DeepMind JAX

Ecosystem (Babuschkin et al., 2020).

References

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, John Quan, George Papamakarios, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Luyu Wang, Wojciech Stokowiec, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

Bahram Behzadian, Soheil Gharatappeh, and Marek Petrik. Fast feature selection for linear value function approximation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 601–609, 2019.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *arXiv preprint arXiv:2101.07123*, 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Aaai*, volume 94, pages 1023–1028, 1994.

Albin Cassirer, Gabriel Barth-Maron, Eugene Brevdo, Sabela Ramos, Toby Boyd, Thibault Sottiaux, and Manuel Kroiss. Reverb: a framework for experience replay. *arXiv preprint arXiv:2102.04736*, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Konstantinos I Diamantaras and Sun Yuan Kung. *Principal component neural networks: theory and applications*. John Wiley & Sons, Inc., 1996.

Carlos E González-Guillén, Carlos Palazuelos, and Ignacio Villanueva. Euclidean distance between haar orthogonal and gaussian matrices. *Journal of Theoretical Probability*, 31(1):93–118, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Yijie Guo, Jongwook Choi, Marcin Moczulski, Samy Bengio, Mohammad Norouzi, and Honglak Lee. Efficient exploration with self-imitation learning via trajectory-conditioned policy. *arXiv preprint arXiv:1907.10247*, 2019.

Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Altché, Rémi Munos, and Mohammad Gheshlaghi Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. In *International Conference on Machine Learning*, pages 3875–3886. PMLR, 2020.

Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *arXiv preprint arXiv:2206.08332*, 2022.

Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2021.

Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Bk8ZcAxR-.

Sridhar Mahadevan. Proto-value functions: Developmental reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 553–560, 2005.

Erkki Oja. Principal components, minor components, and linear neural networks. *Neural networks*, 5(6):927–935, 1992.

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759, 2008.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uCQfPZwRaUu.

H. Francis Song, Abbas Abdolmaleki, Jost Tobias Springenberg, Aidan Clark, Hubert Soyer, Jack W. Rae, Seb Noury, Arun Ahuja, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Dan Belov, Martin Riedmiller, and Matthew M. Botvinick. V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylOlp4FvH.

Zhao Song, Ronald E Parr, Xuejun Liao, and Lawrence Carin. Linear feature encoding for reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020a.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3): 261–272, 2020b.

Xiang Wang, Xinlei Chen, Simon S Du, and Yuandong Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

Andreas Weingessel and Kurt Hornik. Svd algorithms: Apex-like versus subspace methods. *Neural Processing Letters*, 5(3):177–184, 1997.

Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=d-kvI4YdNu.

APPENDICES: Understanding Self-Predictive Learning for Reinforcement Learning

A. Detailed derivations of ODE systems

We provide a derivation of the ODE systems in Equations (5) and (9) below. We start with a few useful facts: recall that $x \sim d, y \sim P^\pi(\cdot|x)$ are one-hot encoding of states. Let D be a diagonal matrix with d its diagonal entries $D_{ii} = d_i, \forall 1 \leq i \leq |\mathcal{X}|$. Then, we have the following properties:

$$\mathbb{E}\left[xx^T\right] = D, \mathbb{E}\left[xy^T\right] = DP^\pi.$$

A.1. Equation (5) for self-predictive learning

Starting with Equation (4), the first-order optimality condition for P_t can be made more explicit

$$\left(\Phi_t^T \mathbb{E}\left[xx^T\right] \Phi_t\right) P_t = \Phi_t^T \mathbb{E}\left[xy^T\right] \Phi_t \Rightarrow \left(\Phi_t^T D \Phi_t\right) P_t = \Phi_t^T D P^\pi \Phi_t.$$

We can expand the dynamics for Φ_t as follows,

$$\dot{\Phi}_t = \left(D - D\Phi_t \left(\Phi^T D \Phi_t\right)^{-1} \Phi_t^T D\right) P^\pi \Phi_t (P_t)^T.$$

Under Assumptions 3 and 4, the above dynamics simplifies into

$$P_t = \Phi_t^T P^\pi \Phi_t, \;\; \dot{\Phi}_t = \left(I - \Phi_t \Phi_t^T\right) P^\pi \Phi_t (P_t)^T,$$

which is the ODE system in Equation (5).

A.2. Equation (9) for bidirectional self-predictive learning

Since the bidirectional self-predictive learning dynamics introduces least square regression from y to x, we need to calculate expectations such as $\mathbb{E}[yy^T]$ and $\mathbb{E}[yx^T]$. In general, it is challenging to express $\mathbb{E}[yy^T]$ as a function of D and P^π. When D is identity (Assumption 4) and when P^π is doubly-stochastic (Assumption 9), we have D as a stationary distribution of P^π and hence $\mathbb{E}[yy^T] = D$ and

$$\mathbb{E}\left[yx^T\right] = \mathbb{E}\left[(xy^T)^T\right] = \left(\mathbb{E}[xy^T]\right)^T = (DP^\pi)^T = (P^\pi)^T D.$$

From Equation (7), we can make explicit the form of the prediction matrix

$$\begin{aligned}
\left(\Phi_t^T \mathbb{E}\left[xx^T\right] \Phi_t\right) P_t &= \Phi_t^T \mathbb{E}\left[xy^T\right] \tilde{\Phi}_t \Rightarrow \left(\Phi_t^T D \Phi_t\right) P = \Phi_t^T D P^\pi \tilde{\Phi}_t, \\
\left(\tilde{\Phi}_t^T \mathbb{E}\left[yy^T\right] \tilde{\Phi}_t\right) \tilde{P}_t &= \tilde{\Phi}_t^T \mathbb{E}\left[yx^T\right] \Phi_t \Rightarrow \left(\tilde{\Phi}_t^T D \Phi_t\right) \tilde{P}_t = \tilde{\Phi}_t^T (P^\pi)^T D \Phi_t,
\end{aligned}$$

Next, we can expand the dynamics of Φ_t and $\tilde{\Phi}_t$ as follows

$$\begin{aligned}
\dot{\Phi}_t &= \left(D - D\Phi_t \left(\Phi^T D \Phi_t\right)^{-1} \Phi_t^T D\right) P^\pi \tilde{\Phi}_t (P_t)^T \\
\dot{\tilde{\Phi}}_t &= \left(D - D\tilde{\Phi}_t \left(\tilde{\Phi}^T D \tilde{\Phi}_t\right)^{-1} \tilde{\Phi}_t^T D\right) \underbrace{D^{-1} (P^\pi)^T D}_{\tilde{P}^\pi} \Phi_t (\tilde{P}_t)^T.
\end{aligned}$$

Interestingly, $\tilde{P}^\pi$ is also a Markov transition matrix that corresponds to the reverse Markov chain. Finally, plugging into $D = I$ (Assumption 4) and thanks to Assumption 8, we recover the dynamics in Equation (9)

$$\begin{aligned}
\dot{\Phi}_t &= \left(I - \Phi_t \Phi_t^T\right) P^\pi \tilde{\Phi}_t (P_t)^T \\
\dot{\tilde{\Phi}}_t &= \left(I - \tilde{\Phi}_t \tilde{\Phi}_t^T\right) (P^\pi)^T \Phi_t (\tilde{P}_t)^T.
\end{aligned}$$

A.3. Equivalence between assumptions in deriving Equation (9)

Now we provide a discussion on the equivalence between assumptions in deriving the ODE for bidirectional self-predictive learning dynamics. An alternative assumption to Assumption 9 is

Assumption 12. Given the sampling process $x \sim d, y \sim P^\pi(\cdot|x)$, the marginal distribution over next state y is uniform.

Our claim is that given the uniformity assumption on the first-state distribution Assumption 4, Assumption 9 and Assumption 12 are equivalent. To see why, given Assumption 9, it is straightforward to see that uniform distribution is a stationary distribution to P^π. Starting from the first-state distribution, which is uniform, the next-state distribution is also uniform, which proves the condition in Assumption 12. Now, given Assumption 12, we conclude the uniform distribution $u = |\mathcal{X}|^{-1}1_{|\mathcal{X}|}$ is a stationary distribution to P^π. By definition of the stationary distribution, this means

$$u^T P^\pi = u.$$

The above implies that each column of P^π sums to 1, and so P^π is doubly-stochastic (Assumption 9).

B. Proof of theoretical results

Theorem 1. Under the dynamics in Equation (4), the covariance matrix $\Phi_t^T \Phi_t \in \mathbb{R}^{k\times k}$ is constant over time.

Proof. Under the dynamics in Equation (4), the prediction matrix P_t optimally minimizes the loss function $L(\Phi_t, P_t)$ given the representation Φ_t. Let $A_t = \Phi_t P_t \in \mathbb{R}^{|\mathcal{X}|\times k}$ be the matrix product. The chain rule combined with the first-order optimality condition on P_t implies

$$\nabla_{P_t} L(\Phi_t, P_t) = \Phi_t^T \partial_{A_t} L(\Phi_t, P_t) = 0. \tag{10}$$

On the other hand, the semi-gradient update for Φ_t can be written as

$$\dot{\Phi}_t = -\nabla_{\Phi_t} \mathbb{E}_{x\sim d, y\sim P^\pi(\cdot|x)} \left[\left\| P_t^T \Phi_t^T x - \text{sg}\left(\Phi_t^T y\right) \right\|_2^2 \right] = -\partial_{A_t} L(\Phi_t, P_t)(P_t)^T.$$

Thanks to Equation (10), we have

$$\Phi_t^T \dot{\Phi}_t = -\Phi_t^T \partial_{A_t} L(\Phi_t, P_t)(P_t)^T = 0.$$

Then, taking time derivative on the covariance matrix

$$\frac{d}{dt}\left(\Phi_t^T \Phi_t\right) = \dot{\Phi}_t^T \Phi_t + \Phi_t^T \dot{\Phi}_t = \left(\Phi_t^T \dot{\Phi}_t\right)^T + \Phi_t^T \dot{\Phi}_t = 0,$$

which implies that the covariance matrix is constant. □

Corollary 2. Under the dynamics in Equation (4), the representation vectors $(\phi_{i,t})_{i=1}^k$ cannot converge to the same vector if they are initialized differently.

Proof. Take any two representation vectors $\phi_{i,t}$ and $\phi_{j,t}$ with $i \neq j$, which at initialization are different. This implies the cosine similarity $\langle \phi_{i,0}, \phi_{j,0} \rangle \neq 1$. Since under the dynamics in Equation (4), the covariance matrix $\Phi_t^T \Phi_t$ is preserved, this means $\phi_{i,t}^T \phi_{j,t}$, $\phi_{i,t}^T \phi_{1,t}$ and $\phi_{i,t}^T \phi_{j,t}$ are all constants over time, which implies

$$\langle \phi_{i,t}, \phi_{j,t} \rangle = \langle \phi_{i,0}, \phi_{j,0} \rangle \neq 1.$$

This means the two vectors cannot be aligned along the same direction for all time $t \geq 0$. □

Lemma 5. Assume P^π is real diagonalizable and let $(u_i)_{i=1}^{|\mathcal{X}|}$ be its set of $|\mathcal{X}|$ distinct eigenvectors. Let $\mathcal{C}_{P^\pi}$ be the set of critical points of Equation (5). Then $\mathcal{C}_{P^\pi}$ contains all matrices whose columns are orthonormal, and have the same span as a set of k eigenvectors.

Proof. Without loss of generality, consider the subset of first k right eigenvectors $U = (u_1...u_k)$. Then $P^\pi U = U\Lambda$ for some diagonal matrix $\Lambda = \text{diag}(\lambda_1...\lambda_k)$, where λ_i is the eigenvalue corresponding to u_i.

If $\Phi_t = U$, then

$$\dot{\Phi}_t = (I - UU^T)P^\pi U(P_t)^T = (I - UU^T)U\Lambda U(P_t)^T = 0.$$

Next, for any set of k orthonormal vectors with the same span as U, we can write them as $U' = UQ$ for some orthogonal matrix $Q \in \mathbb{R}^{k\times k}$. If $\Phi_t = U' = UQ$, then

$$\dot{\Phi}_t = (I - UQQ^TU)P^\pi UQ(P_t)^T = (I - UU^T)U\Lambda UQ(P_t)^T = 0,$$

which concludes the proof. □

Theorem 6. If P^π is symmetric, then under Assumption 3 and learning dynamics Equation (5), the trace objective is non-decreasing $\dot{f} \geq 0$, where

$$f(\Phi_t) \coloneqq \text{Trace}\left(\left(\Phi_t^T P^\pi \Phi_t\right)^T \left(\Phi_t^T P^\pi \Phi_t\right)\right).$$

If $\Phi_t \notin \mathcal{C}_{P^\pi}$, then $\dot{f} > 0$. Under the constraint $\Phi^T\Phi = I$, the maximizer to $f(\Phi)$ is any set of k orthonormal vectors which span the principal subspace, i.e., with the same span as the k eigenvectors of P^π with top absolute eigenvalues.

Proof. We first show that the objective is non-decreasing. We calculate

$$\begin{aligned}
\frac{d}{dt}f(\Phi_t) &= 4 \cdot \text{Trace}\left(\left(\Phi_t^T P^\pi \Phi_t\right)^T \Phi_t^T P^\pi \dot{\Phi}_t\right) \\
&=_{(a)} 4 \cdot \text{Trace}\left(P_t \Phi_t^T P^\pi \left(I - \Phi_t\Phi_t^T\right) P^\pi \Phi_t P_t^T\right) \\
&=_{(b)} 4 \cdot \text{Trace}\left(\left(P^\pi \Phi_t P_t^T\right)^T \left(I - \Phi_t\Phi_t^T\right) P^\pi \Phi_t P_t^T\right),
\end{aligned}$$

where (a) follows from $P_t = \Phi_t^T P^\pi \Phi_t$; (b) follows from the fact that P^π is symmetric and as a result P_t is symmetric. Now, let $A_t = P^\pi \Phi_t P_t^T$ and denote its column vectors as $A_t = [a_{1,t}...a_{k,t}]$. The above derivative rewrites as

$$4 \cdot \sum_{i=1}^{k} a_{i,t}^T \left(I - \Phi_t\Phi_t^T\right) a_{i,t}.$$

We remind that a projection matrix M satisfies $M^2 = M$ and $M^T = M$ and corresponds to an orthogonal projection onto certain subspace. Since $I - \Phi_t\Phi_t^T$ is a projection matrix, we have $a_{i,t}^T \left(I - \Phi_t\Phi_t^T\right) a_{i,t} \geq 0$ for any $a_{i,t} \in \mathbb{R}^{\mathcal{X}}$. Hence $\frac{d}{dt}f(\Phi_t) \geq 0$. Now, if $\Phi_t \notin \mathcal{C}_{P^\pi}$, this means there exists certain columns $a_{i,t}$ of A_t such that $a_{i,t} \notin \text{span}(\Phi_t)$. This means $a_{i,t}^T \left(I - \Phi_t\Phi_t^T\right) a_{i,t} > 0$ and therefore $\dot{f} > 0$.

Finally, we examine the maximizer to $f(\Phi)$ under the constraint $\Phi^T\Phi = I_{k\times k}$. Since $\Phi^T P^\pi \Phi$ is symmetric, there exists an orthogonal matrix Q such that

$$Q^T\Phi^T P^\pi \Phi Q = \Lambda,$$

for some diagonal matrix Λ. Note that since $(\Phi Q)^T \Phi Q = I_{k\times k}$, it is equivalent to consider the optimization problem under a stronger constraint $\Phi^T\Phi = I_{k\times k}$ and $\Phi^T P^\pi \Phi = \Lambda$ for some diagonal matrix Λ. Therefore, the optimization problem becomes

$$\max_{\Phi^T\Phi = I_{k\times k}, \Phi^T P^\pi \Phi = \Lambda} \sum_{i=1}^{k} \Lambda_{ii}^2.$$

Let $\Phi = [\phi_1, ...\phi_k]$ with column vectors $\phi_i \in \mathbb{R}^{|\mathcal{X}|}$ for all $1 \leq i \leq k$, then we have the equivalent optimization problem

$$\max_{\Phi^T\Phi = I_{k\times k}, \Phi^T P^\pi \Phi = \Lambda} \sum_{i=1}^{k} \left(\phi_i^T P^\pi \phi_i\right)^2 \leq_{(a)} \max_{\Phi^T\Phi = I_{k\times k}, \Phi^T P^\pi \Phi = \Lambda} \sum_{i=1}^{k} \|P^\pi \phi_i\|_2^2$$
$$= \max_{\Phi^T\Phi = I_{k\times k}, \Phi^T P^\pi \Phi = \Lambda} \sum_{i=1}^{k} \phi_i^T \left(P^\pi\right)^T P^\pi \phi_i.$$

Here, (a) follows from the fact that ϕ_i is a unit-length vector and the application of the inequality $a^T b \leq \|a\|_2 \|b\|_2$. It is straightforward to see that the optimal solution to the last optimization problem is the set of eigenvectors of P^π with top squared eigenvalues. Hence,

$$\max_{\Phi^T\Phi = I_{k\times k}} f(\Phi) \leq \sum_{i=1}^{k} \lambda_i^2.$$

On the other hand, the k eigenvectors of P^π with top k absolute eigenvalues is a feasible solution and therefore $\max_{\Phi^T\Phi = I_{k\times k}} f(\Phi) \geq \sum_{i=1}^{k} \lambda_i^2$. The above implies $\max_{\Phi^T\Phi = I_{k\times k}} f(\Phi) = \sum_{i=1}^{k} \lambda_i^2$, and the k eigenvectors of P^π with top k absolute eigenvalues is a maximizer to the constrained optimization problem. It is then also clear that any k orthonormal vectors with the same span as the top k eigenvectors also achieves the maximum objective. □

Theorem 7. Under the bidirectional self-predictive learning dynamics in Equation (8), the covariance matrices $\Phi_t^T \Phi_t \in \mathbb{R}^{k\times k}$ and $\tilde{\Phi}_t^T \tilde{\Phi}_t \in \mathbb{R}^{k\times k}$ are both constant matrices over time.

Proof. We consider the forward and backward loss function separately. Following the arguments in the proof of Theorem 1, we see that since P_t is computed as the optimal solution to $L_{\mathrm{f}}(P_t, \Phi_t)$, it satisfies the first-order optimality condition and as a result, $\Phi_t^T \dot{\Phi}_t = 0$. This implies $\Phi_t^T \Phi_t$ is a constant matrix over time. Applying the same set of arguments to the backward loss function $L_{\mathrm{b}}(\tilde{\Phi}_t, \tilde{P}_t)$, we conclude $\tilde{\Phi}_t^T \tilde{\Phi}_t$ is also a constant matrix over time. □

Lemma 10. Let $\tilde{\mathcal{C}}_{P^\pi} \subset \mathbb{R}^{|\mathcal{X}|\times k} \times \mathbb{R}^{|\mathcal{X}|\times k}$ be the set of critical points to Equation (9). Then $\tilde{\mathcal{C}}_{P^\pi}$ contains any pair of matrices, whose columns are orthonormal and have the same span as k of singular vector pairs.

Proof. Without loss of generality, consider the subset of k top singular vector pairs $U = (u_1...u_k), V = (v_1...v_k)$. By construction, they satisfy the equality $P^\pi V = U\Sigma$ and $\left(P^\pi\right)^T U = V\Sigma$ where $\Sigma = \mathrm{diag}(\sigma_1...\sigma_k)$ is the diagonal matrix with corresponding singular values.

Setting $\Phi_t = U, \tilde{\Phi}_t = V$, we first verify the critical conditions for $\dot{\Phi}_t = 0, \dot{\tilde{\Phi}}_t = 0$:

$$(I - \Phi_t \Phi_t^T) P^\pi \tilde{\Phi}_t (P_t)^T = (I - UU^T) P^\pi V V^T \left(P^\pi\right)^T U =_{(a)} (I - UU^T) U \Sigma^2 = 0.$$
$$\left(I - \tilde{\Phi}_t \tilde{\Phi}_t^T\right) \left(P^\pi\right)^T \Phi_t (\tilde{P}_t)^T = \left(I - VV^T\right) \left(P^\pi\right)^T U U^T P^\pi V =_{(b)} (I - VV^T) V \Sigma^2 = 0.$$

Here, (a) and (b) both follow from the property of the singular vector pairs. The above indicates that any k singular vector pairs constitute a member of $\tilde{\mathcal{C}}_{P^\pi}$.

Any orthonormal vectors with the same vector span as U, V can be expressed as UQ, VR for some orthogonal matrix $Q, R \in \mathbb{R}^{k\times k}$. Let $U' = UQ, V' = VR$, we verify the critical conditions when $\Phi_t = U', \tilde{\Phi}_t = V'$,

$$(I - \Phi_t \Phi_t^T) P^\pi \tilde{\Phi}_t (P_t)^T = (I - U'(U')^T) P^\pi V' (V')^T \left(P^\pi\right)^T U' =_{(a)} (I - UU^T) U \Sigma^2 Q = 0.$$
$$\left(I - \tilde{\Phi}_t \tilde{\Phi}_t^T\right) \left(P^\pi\right)^T \Phi_t (\tilde{P}_t)^T = \left(I - V'(V')^T\right) \left(P^\pi\right)^T U'(U')^T P^\pi V' =_{(b)} (I - VV^T) V \Sigma^2 R = 0,$$

where (a) and (b) follow from straightforward matrix operations. We have hence verified that any orthonormal vectors with the same vector span as any subset of k singular vector pairs constitute a critical point. □

Theorem 11. Under Assumption 8 and the learning dynamics in Equation (9), the following SVD trace objective is non-decreasing $\dot{\tilde{f}} \geq 0$, where

$$\tilde{f}\left(\Phi_t, \tilde{\Phi}_t\right) \coloneqq \mathrm{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)\right).$$

If $(\Phi_t, \tilde{\Phi}_t) \notin \tilde{\mathcal{C}}_{P^\pi}$, then $\dot{\tilde{f}} > 0$. Under the constraint $\Phi^T\Phi = \tilde{\Phi}^T\tilde{\Phi} = I$, the maximizer to $\tilde{f}(\Phi, \tilde{\Phi})$ is any two sets of k orthonormal vectors with the same span as the k singular vector pairs of P^π with top singular values.

Proof. We start by showing the SVD trace objective is non-decreasing on the ODE flow.

$$\frac{d}{dt}\tilde{f}\left(\Phi_t, \tilde{\Phi}_t\right) = 2 \cdot \text{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \left(\Phi_t^T P^\pi \dot{\tilde{\Phi}}_t\right)\right) + 2 \cdot \text{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \left(\dot{\Phi}_t^T P^\pi \tilde{\Phi}_t\right)\right).$$

Examining the first term on the right above, plugging in the dynamics for $\dot{\tilde{\Phi}}$,

$$\begin{aligned}\text{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \left(\Phi_t^T P^\pi \dot{\tilde{\Phi}}_t\right)\right) &= \text{Trace}\left(\left(\Phi_t^T P^\pi \tilde{\Phi}_t\right)^T \Phi_t^T P^\pi \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right)(P^\pi)^T \Phi_t(\tilde{P}_t)^T\right) \\ &=_{(a)} \text{Trace}\left(\left((P^\pi)^T \Phi_t(\tilde{P}_t)^T\right)^T \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right)(P^\pi)^T \Phi_t(\tilde{P}_t)^T.\right)\end{aligned}$$

Here, (a) follows from the form of the prediction matrix $\tilde{P}_t = \tilde{\Phi}_t^T (P^\pi)^T \Phi_t$. Now, define $\tilde{A}_t = [a_{1,t}...a_{k,t}] \coloneqq (P^\pi)^T \Phi_t(\tilde{P}_t)^T$, the above rewrites as

$$\text{Trace}(A_t^T \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right) A_t) = \sum_{i=1}^k a_{i,t}^T \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right) a_{i,t}.$$

Since $\left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right)$ is an orthogonal projection matrix, we conclude the above quantity is non-negative. Similarly, we can show that the second term on the right above is also non-negative, which concludes $\dot{\tilde{f}} \geq 0$.

Now, we assume $(\Phi_t, \tilde{\Phi}_t) \notin \tilde{\mathcal{C}}_{P^\pi}$. Without loss of generality, we assume in this case $\dot{\tilde{\Phi}}_t \neq 0$, which implies there exists certain column i such that $a_{i,t}$ is not in the span of $\tilde{\Phi}_t$. This means $a_{i,t}^T \left(I - \tilde{\Phi}_t\tilde{\Phi}_t^T\right) a_{i,t} > 0$ and subsequently $\dot{\tilde{f}} > 0$.

Finally, we show that under the constraint $\Phi^T\Phi = \tilde{\Phi}^T\tilde{\Phi} = I$, the maximizer to $\tilde{f}(\Phi, \tilde{\Phi})$ is any two set of k orthonormal vectors with the same span as the k singular vector pairs of P^π with top singular values. In general, the matrix $\Phi^T P^\pi \tilde{\Phi}$ is not diagonal. Consider its SVD

$$\Phi^T P^\pi \tilde{\Phi} = \tilde{U}\tilde{\Sigma}\tilde{V}^T,$$

then we have $\left(\Phi\tilde{U}\right)^T P^\pi \tilde{\Phi}\tilde{V} = \tilde{\Sigma}$. Consider the new representation variable $\Phi' = \Phi\tilde{U}$ and $\tilde{\Phi}' = \tilde{\Phi}\tilde{V}$ and note they also satisfy the orthonormal constraint. Under the new variable, the matrix $(\Phi')^T P^\pi \tilde{\Phi}' = \tilde{\Sigma}$ is diagonal. Since $f(\Phi, \tilde{\Phi}) = f(\Phi\tilde{U}, \tilde{\Phi}\tilde{V})$, it is equivalent to solve the optimization problem under an additional diagonal constraint

$$\max_{\Phi^T\Phi = \tilde{\Phi}^T\tilde{\Phi} = I} f(\Phi, \tilde{\Phi}) = \max_{\Phi^T\Phi = \tilde{\Phi}^T\tilde{\Phi} = I, \Phi^T P^\pi \tilde{\Phi} \text{ diagonal}} f(\Phi, \tilde{\Phi})$$

When $\Phi^T P^\pi \tilde{\Phi}$ is diagonal, the objective rewrites as

$$\sum_{i=1}^k \left(\phi_i^T P^\pi \tilde{\phi}_i\right)^2 \leq_{(a)} \sum_{i=1}^k \left\|P^\pi \tilde{\phi}_i\right\|_2^2 = \sum_{i=1}^k \tilde{\phi}_i^T (P^\pi)^T P^\pi \tilde{\phi}_i,$$

where (a) follows from the fact that ϕ_i is a unit-length vector and the application of the inequality $a^T b \leq \|a\|_2 \|b\|_2$. Now, under the constraint $\tilde{\Phi}^T\tilde{\Phi} = I$, the right hand side is upper bounded by the sum of squared top k singular values of P^π: $\sum_{i=1}^k \sigma_i^2$. Let f^* be the optimal objective of the original constrained problem. We have hence established $f^* \leq \sum_{i=1}^k \sigma_i^2$. On the other hand, if we let $\Phi, \tilde{\Phi}$ to be the top k singular vector pairs, they satisfy the constraint and this shows $f^* \geq \sum_{i=1}^k \sigma_i^2$.

In summary, we have $f^* = \sum_{i=1}^k \sigma_i^2$ and the top k singular vector pairs U, V are the maximizer. Any k orthonormal vectors with the same span as top k singular vector pairs can be expressed as $U' = UQ, V' = VR$ for some orthogonal matrix $Q, R \in \mathbb{R}^{k\times k}$. Since $(U')^T U' = (V')^T V' = I$ and $f(U, V) = f(U', V') = f^*$, they are also the maximizer to the constrained problem. □

C. Extension of non-collapse property to general loss function

Thus far, we have focused on the squared loss function for understanding the self-predictive learning dynamics,

$$\begin{aligned} &P_t \in \arg\min_P L(\Phi_t, P), \\ &\dot{\Phi}_t = -\nabla_{\Phi_t} \mathbb{E}\left[\left\| P_t^T \Phi_t^T x - \mathrm{sg}\left(\Phi_t^T y\right)\right\|_2^2\right]. \end{aligned} \tag{11}$$

We can extend the result to a more general class of loss function $L(\Phi_t, P_t)$. Such a loss function needs to be computed as an expectation over a function F of the product prediction and representation matrix $\Phi_t P_t$, used for computing the prediction; and another argument for the representation matrix Φ_t, used for computing the target. More formally, we can write $L(\Phi_t, P_t) = F\left(P_t \Phi_t, \Phi_t\right) = \mathbb{E}\left[f\left(P_t^T \Phi_t^T x, \Phi_t^T y\right)\right]$ for some function $f : \mathbb{R}^k \times \mathbb{R}^k \to k$. For the least squared case, we have $f(a, b) = \|a - b\|_2^2$. We consider the self-predictive learning dynamics with such a general loss function,

$$\begin{aligned} &P_t \in \arg\min_P F\left(\Phi_t P, \Phi_t\right), \\ &\dot{\Phi}_t = -\nabla_{\Phi_t} F\left(\Phi_t P_t, \mathrm{sg}\left(\Phi_t\right)\right). \end{aligned} \tag{12}$$

We show that the non-collapse property also holds for the above dynamics.

Theorem 13. Assume the general loss function $L(\Phi, P)$ is such that the minimizer to $L(\Phi, P)$ satisfies the first-order optimality condition, then under the dynamics in Equation (12), the covariance matrix $\Phi_t^T \Phi_t \in \mathbb{R}^{k \times k}$ is constant over time.

Proof. The proof follows closely from the proof of Theorem 1. Let $A_t = \Phi_t P_t \in \mathbb{R}^{|\mathcal{X}| \times k}$ be the matrix product, the assumption implies

$$\nabla_{P_t} F(\Phi_t P_t, \Phi_t) = \Phi_t^T \partial_{A_t} F(\Phi_t P_t, \Phi_t) = 0. \tag{13}$$

On the other hand, the semi-gradient update for Φ_t can be written as

$$\dot{\Phi}_t = -\nabla_{\Phi_t} F\left(\Phi_t P_t, \Phi_t\right) = -\partial_{A_t} L(\Phi_t, P_t)(P_t)^T.$$

Thanks to Equation (13), we have

$$\Phi_t^T \dot{\Phi}_t = -\Phi_t^T \partial_{A_t} F\left(\Phi_t P_t, \Phi_t\right)(P_t)^T = 0.$$

Then, taking time derivative on the covariance matrix

$$\frac{d}{dt}\left(\Phi_t^T \Phi_t\right) = \dot{\Phi}_t^T \Phi_t + \Phi_t^T \dot{\Phi}_t = \left(\Phi_t^T \dot{\Phi}_t\right)^T + \Phi_t^T \dot{\Phi}_t = 0,$$

which implies that the covariance matrix is constant. □

Notable examples of loss functions that satisfy the above assumptions include L_1 loss $f(a, b) = |a - b|$ and the regularized cosine similarity loss $f(a, b) = -a^T b / \left(\|a\|_2 \|b\|_2 + \epsilon\right)$, where $\epsilon > 0$ is a regularization constant.

D. Discussions on critical points to the self-predictive learning dynamics

We provide further discussions on the critical points to the self-predictive learning dynamics in Equation (5). For convenience, we recall the set of ODEs

$$P_t = \Phi_t^T P^\pi \Phi_t, \quad \dot{\Phi}_t = \left(I - \Phi_t \Phi_t^T\right) P^\pi \Phi_t (P_t)^T.$$

According to Lemma 5, under the assumption that P^π is real diagonizable, any matrix with orthonormal columns with the same span as a set of k eigenvectors constitutes a critical point to the ODE. Let $\mathcal{C}$ be the set of such matrices and recall that $\mathcal{C}_{P^\pi}$ is the set of critical points, we have $\mathcal{C} \subseteq \mathcal{C}_{P^\pi}$.

We now consider two symmetric transition matrices, under which we have $\mathcal{C} = \mathcal{C}_{P^\pi}$ and $\mathcal{C} \subsetneq \mathcal{C}_{P^\pi}$ respectively. In both cases, we have $|\mathcal{X}| = 2$ and $k = 1$ for simplicity. In this case, Φ can be expressed as a column vector Φ_t is a 2-dimensional vector and the prediction matrix P_t is now a scalar.

Case I where $\mathcal{C} = \mathcal{C}_{P^\pi}$. Consider the following transition matrix

$$P^\pi = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

The transition matrix is symmetric and has eigenvalue $\lambda_1 = 1, \lambda_2 = 0.8$. Note that

$$U = [u_1, u_2] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

is the matrix of eigenvectors. For convenience, we write

$$\Phi_t = U \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}$$

with coefficients $\alpha_t, \beta_t \in \mathbb{R}$. Assumption 3 dictates $\alpha_t^2 + \beta_t^2 = 1$. Now, we can calculate

$$P_t = \Phi_t^T P^\pi \Phi_t = \alpha_t^2 + 0.8\beta_t^2 > 0,$$

which is strictly positive and hence always non-zero. Letting $\dot{\Phi}_t = 0$, since $P_t \neq 0$, we conclude $\text{span}(P^\pi \Phi_t) \subset \text{span}(\Phi_t)$. This means Φ_t is invariant and must be a direct sum of subspaces spanned by eigenvectors. In other words, we have $\mathcal{C}_{P^\pi} \subset \mathcal{C}$ and hence the two sets are in fact equal.

Case II where $\mathcal{C} \subsetneq \mathcal{C}_{P^\pi}$. Consider the following transition matrix we mentioned in Equation (6)

$$P^\pi = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}.$$

The transition matrix is symmetric and has eigenvalue $\lambda_1 = 1, \lambda_2 = -0.8$. Note that

$$U = [u_1, u_2] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

is the matrix of eigenvectors. For convenience, we write

$$\Phi_t = U \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}$$

with coefficients $\alpha_t, \beta_t \in \mathbb{R}$. Assumption 3 dictates $\alpha_t^2 + \beta_t^2 = 1$. As before, we calculate

$$P_t = \Phi_t^T P^\pi \Phi_t = \alpha_t^2 - 0.8\beta_t^2 > 0.$$

Now, we can identify $\alpha_t = \pm\frac{1}{\sqrt{1.8}}, \beta_t = \pm\frac{\sqrt{0.8}}{\sqrt{1.8}}$ as four non-eigenvector critical points. Indeed, since $P_t = 0$, we have $\dot{\Phi}_t = 0$. However, since this critical point is a combination of two eigenvectors u_1, u_2, it does not span the same subspace as either just u_1 or u_2. In other words, we have found a critical point which does not belong to the set $\mathcal{C}$ and this implies $\mathcal{C} \subsetneq \mathcal{C}_{P^\pi}$.

Convergence of the dynamics in Case II. We replicate the diagram Fig. 3 here in Fig. 7, where we graph critical points to the self-predictive learning dynamics with the above transition matrix on a unit circle. Recall that we have $k = 1$ so that representations Φ_t are 2-d vectors. In addition to the eigenvector critical points plotted as blue dots (Lemma 5), we have also identified other four critical points in red, corresponding to $\alpha_t = \pm\frac{1}{\sqrt{1.8}}, \beta_t = \pm\frac{\sqrt{0.8}}{\sqrt{1.8}}$ in four quadrants of the 2-d plane.

The only local update dynamics consistent with the local improvement property (Theorem 6) is shown as black arrows on the unit circle. When the representation is initialized near the bottom eigenvector, it will converge to one of the four non-eigenvector critical points and not the top eigenvector.

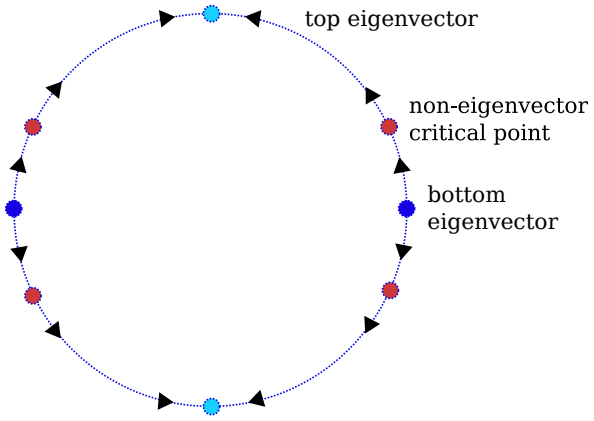


Figure 7. Critical points and local dynamics of the example MDP in Equation (6). We consider $k = 1$ so representations Φ_t are 2-d vectors. There are four eigenvector critical points (light and dark blue) and four non-eigenvector critical points (red) of the ODE, shown on the unit circle. The black arrows show the local update direction based on the ODE. Initialized near the bottom eigenvector, the dynamics converges to one of the four non-eigenvector critical points and not to the top eigenvector.

E. Algorithmic and implementation details on BYOL-RL

We provide background details on BYOL-RL (Guo et al., 2022), a deep RL agent based on which our deep bidirectional self-predictive learning algorithm is implemented. We start with a relatively high-level description of the algorithm.

BYOL-RL is designed to work in the POMDP setting, where the agent observes a sequence of observations over time $o_t \in \mathcal{O}$. Define history $h_t \in \mathcal{H}$ at time t as the combination of previous observations $h_t = (o_s)_{s \le t}$ (note here the observation o_s can contain action a_{s-1}). BYOL-RL adopts a few functions to represent the raw observation and history

- An observation embedding function $f : \mathcal{O} \to \mathbb{R}^d$ which maps the observation o_t into d-dimensional embeddings. In practice, this is implemented as a convolutional neural network.
- A recurrent embedding function $g : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^k$ which processes the observation in a recurrent way. In practice, this is implemented as the core output of a LSTM.

In POMDP, we can consider the history h_t as a proxy to the state in the MDP case. Let $\Phi(h_t) \in \mathbb{R}^k$ be the k-dimensional representation of history, we use the recurrent function to embed the history recursively

$$\Phi(h_t) = g\left(\Phi(h_{t-1}), f(o_t)\right).$$

BYOL-RL parameterizes the latent prediction function $p : \mathbb{R}^k \times (\mathcal{A})^n \to \mathbb{R}^k$, which can be understood as predicting the history representation n-step from now on, using only intermediate action sequence $a_{t:t+n-1}$

$$p\left(\Phi(h_t), a_{t:t+n-1}\right) \in \mathbb{R}^k.$$

Finally, another projection function $q : \mathbb{R}^k \to \mathbb{R}^d$ maps the predicted history representation, into the d-dimensional embedding space of the observation. Overall, the prediction objective is

$$\mathbb{E}\left[\left\|q\left(p\left(\Phi(h_t), a_{t:t+n-1}\right)\right) - \mathrm{sg}\left(f(o_{t+n})\right)\right\|_2^2\right].$$

The notation sg indicates stop-gradient on the prediction target. All parameterized functions in BYOL-RL f, g, p, q, are optimized via semi-gradient descent.

Note that there are a number of discrepancies between theory and practice, such as multi-step prediction, action-conditional prediction and partial observability. See Section 6 for some discussions on possible extensions of the theoretical model to the more general case. We refer readers to the original paper (Guo et al., 2022) for detailed description of the neural network architecture and hyper-parameters.

E.1. Details on deep bidirectional self-predictive learning with BYOL-RL

We build the deep bidirectional self-predictive learning algorithm on top of BYOL-RL. The bidirectional self-predictive learning dynamics motivates a backward prediction loss function, which we instantiate as follows in the POMDP case.

For the backward prediction, we can instantiate an observation embedding function $\tilde{f} : \mathcal{O} \to \mathbb{R}^d$ and recurrent embedding function $\tilde{g} : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^k$, analogous to the forward prediction case. We also parameterize a backward latent dynamics function $\tilde{p} : \mathbb{R}^k \times (\mathcal{A})^n \to \mathbb{R}^k$. Finally, we parameterize a projection function $\tilde{q} : \mathbb{R}^k \to \mathbb{R}^d$. The overall backward prediction objective is

$$\mathbb{E}\left[\left\|\tilde{q}\left(\tilde{p}\left(\tilde{\Phi}(h_{t+n}), a_{t:t+n-1}\right)\right) - \text{sg}\left(\tilde{f}(o_t)\right)\right\|_2^2\right],$$

where the backward recurrent representation is computed recursively as $\tilde{\Phi}(h_{t-1}) = \tilde{g}\left(\tilde{\Phi}(h_t), \tilde{f}(o_{t-1})\right)$. In other words, we can understand the backward prediction problem as almost exactly mirroring the forward prediction problem.

As a design choice, we share the observation embedding in both the forward and backward process $f = \tilde{f}$. The motivation for such a design choice is that one arguably expects the observation embedding to share many common features for both the forward and backward process, when the input observations are images; secondly, since the embedding function is usually a much larger network compared to rest of the architecture, parameter sharing helps reduce the computational cost.

E.2. Differences from PBL (Guo et al., 2019)

PBL is a representation learning algorithm based on both forward and backward predictions. In the POMDP case, PBL simply parameterizes a projection function $\tilde{q}_{\text{pbl}} : \mathbb{R}^k \to \mathbb{R}^d$. The backward prediction loss is computed as

$$\mathbb{E}\left[\left\|\tilde{q}_{\text{pbl}}\left(\tilde{f}\left(o_{t+n}\right)\right) - \text{sg}\left(\Phi(h_{t+n})\right)\right\|_2^2\right].$$

In other words, the backward prediction seeks to predict the recurrent history embedding $\Phi(h_{t+n})$ from the observation embedding $\tilde{f}(o_{t+n})$. By design, the backward prediction shares the same observation embedding function as the forward prediction $\tilde{f} = f$.

F. Transition matrix to illustrate the failure mode of self-predictive learning

To design examples that illustrate the failure mode of single representation learning dynamics, it is useful to review the intuitive interpretations of left and right singular vectors of P^π as clustering states with certain similar features. Left singular vectors cluster together states with similar outgoing distribution, i.e., states with similar rows in P^π. Meanwhile, right singular vectors cluster together states with similar incoming distributions, i.e., states with similar columns in P^π.

The example transition matrix with $|\mathcal{X}| = 3$ states is

$$P^\pi = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}$$

We can calculate the top-1 left and right singular vectors as

$$u_0 = \left[-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right], \tilde{u}_0 = \left[0, -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right],$$

which concides with the previous intuition that the top left singular vector should cluster together the first two states. Indeed, by assigning a value of $-1/\sqrt{2}$ to the first two states, the top left singular vector effectively considers the first two states as being identical. Meanwhile, the top right singular vector should cluster together the last two states.

G. Experiment details

We provide additional details on the experimental setups in the paper.

G.1. Tabular experiments

Throughout, tabular experiments are carried out on randomly generated MDPs. Instead of explicitly generating the MDPs, we generate the state transition matrix $P^\pi \in \mathbb{R}^{|\mathcal{X}\times|\mathcal{X}|}$ where $|\mathcal{X}| = 20$ by default. In general, the algorithm learns $k = 2$ representation columns.

Generating random P^π. To generate random doubly-stochastic transition matrix, we start by randomly initialize entries of P to be i.i.d. Uniform$(0, 1)$. Then we carry out column normalization and row normalization

$$P_{ij} \leftarrow P_{ij}/\sum_k P_{ik}, P_{ij} \leftarrow P_{ij}/\sum_k P_{kj}$$

until convergence. It is guaranteed that P has row sum and column sum to be both 1, and is hence doubly-stochastic. The transition matrix is computed as

$$P^\pi = \alpha P + (1-\alpha)P_{\text{perm}},$$

where P_{perm} is a randomly generated permutation matrix and $\alpha \sim$ Uniform$(0, 1)$. It is straightforward to verify P^π is doubly-stochastic. To generate symmetric transition matrix, we follow the above procedure and apply a symmetric operation $(P^\pi + (P^\pi)^T)/2$, which produces P^π as a symmetric transition matrix. In Fig. 4, we generate doubly-stochastic matrices as examples of non-symmetric matrices.

Normalized trace objective. When plotting trace objectives, we usually calculate the normalized objective. For symmetric matrix P^π, the normalizer is the sum of top-k squared eigenvalue of P^π: let λ_i be the eigenvalues of P^π ordered such that $|\lambda_i| \geq |\lambda_{i+1}|$, then normalizer is $\sum_{i=1}^k |\lambda_i|^2$. For double-stochastic matrix P^π, the normalizer is the squaresum of topd ma-$k$ximum singular singular value: let σ_i be the singular values of P^π, the normalizer is $\sum_{i=1}^k \sigma_i^2$. Such normalzers upper bound the trace objective and SVD trace objective respectively.

ODE vs. discretized update. We carry out experiments in the tabular MDPs with setups. In Figs. 4 and 5, we simulate the exact ODE dynamics using the Scipy ODE solver (Virtanen et al., 2020b). In Figs. 2 and 10, we simluate the discretized process using a finite learning rate η on the representation matrix Φ_t. This corresponds to implementing the update rule in Equation (2). By default, in discretized updates we adopt $\eta = 10^{-3}$ so that the non-coallpse property is almost satisfied.

G.1.1. Complete result to Fig. 4

We present the complete result to Fig. 4 in Fig. 8, where we display the evolution of the trace objective for a total of 100 iterations in Fig. 8(a). Note that as the simulation runs longer, in the non-symmetric MDP case, we can observe very small oscillations in the trace objective. However, overall, the trace objective improves significantly compared to the initial values.

In Fig. 2(b), we make more clear the individual runs across different MDPs. Though in most MDPs, the trace objective is improved significantly given 100 iterations, there exists MDP instances where the improvement is very slow and appear highly monotonic in the non-symmetric case. In the symmetric case, there can exist MDPs where the rate of improvement for the trace objective is slow too.

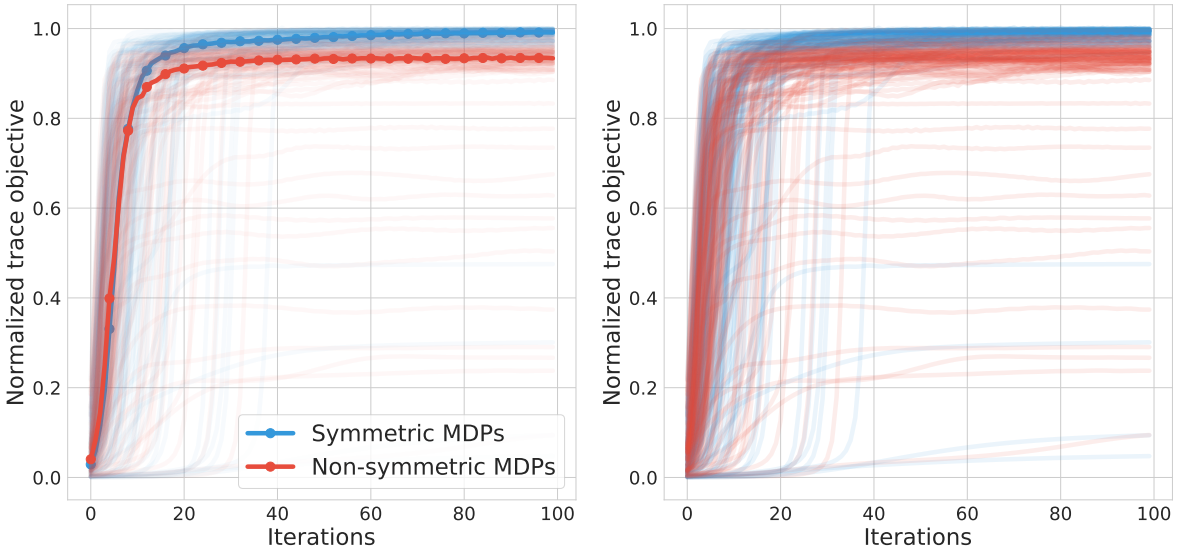
G.1.2. Effect of target network

In practice, it is common to maintain a target network for constructing prediction targets for self-prediction (Guo et al., 2019; Schwarzer et al., 2021; Guo et al., 2020). Under our framework, the prediction loss function is

$$\tilde{L}(\Phi, P, \Phi') = \mathbb{E}_{x\sim d, y\sim P^\pi(\cdot|x)}\left[\left\|P^T\Phi^T x - (\Phi')^T y\right\|_2^2\right],$$

where Φ' is the target network, or target representation matrix. The target representation matrix is updated via moving average towards the online representation matrix

$$\frac{d}{dt}\Phi'_t = \beta(\Phi_t - \Phi'_t).$$



(a) Complete run (b) Individual runs

Figure 8. Complete result for Fig. 4 where we now show the result over a large number of iterations in (a). In (b), we make more clear the individual runs across different MDPs.

The overall self-predictive learning dynamics with target representation is:

$$\begin{aligned} P_t &= \arg\min_P \tilde{L}(\Phi_t, P, \Phi'_t), \\ \dot{\Phi}'_t &= \beta(\Phi_t - \Phi'_t), \\ \dot{\Phi}_t &= -\nabla_{\Phi_t} \mathbb{E}\left[\left\| P_t^T \Phi_t^T x - \text{sg}\left((\Phi'_t)^T y\right)\right\|_2^2\right]. \end{aligned} \tag{14}$$

In Fig. 9, we carry out ablation on the effect of β. We consider the symmetric MDP case where the self-predictive learning dynamics in Equation (4) should monotonically improve the trace objective. Here, we also plot the trace objective. When $\beta = 0$, the result shows that the trace objective still improves compared to the initialization, though such an improvement is much more limited and can be very non-monotonic. When $\beta > 0$, we see that the learning dynamics behaves very similarly to Equation (4).

G.1.3. Ablation on how hyper-parameters impact non-collapse dynamics

We now present results on how a number of different hyper-parameters impact the non-collapse dynamics: finite learning rate and non-optimal predictor.

Finite learning rate. Thus far, our theory has been focused on the continuous time case, which corresponds to a infinitesimally small learning rate. With finite learning rate, we expect the non-collapse to be violated. Fig. 10(a) shows the effect of finite learning rate on the preservation of the cosine similarity between two representation vectors $\phi_{1,t}$ and $\phi_{2,t}$. The two vectors are initialized to be orthogonal, so their cosine similarity is initialized at 0. We consider a grid of learning rate $\eta \in \{0.01, 0.1, 1, 10\}$; to ensure fair comparison, for learning rate η, we showcase the cosine similarity $\langle \phi_{1,t}, \phi_{2,t} \rangle$ at iteration T/η with $T = 10000$. Interpreting η as the magnitude of the step-size, this is to ensure that at each learning rate, the result is obtained after updating for a total step-size of $\eta \cdot T/\eta = T$.

As Fig. 10(a) shows, when the learning rate increases, the cosine similarity $\langle \phi_{1,t}, \phi_{2,t} \rangle$ increases, indicating a more severe violation of the non-collapse property. As $\langle \phi_{1,t}, \phi_{2,t} \rangle \to 1$, the two representation vectors become more and more aligned with each other, and eventually coallpse to the same direction.

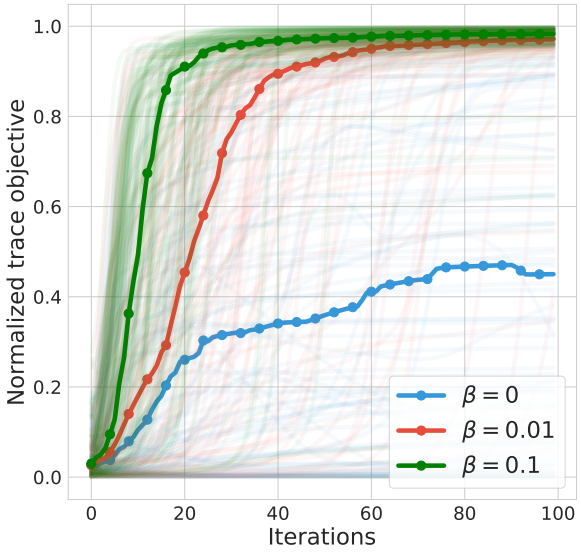


Figure 9. Impact of target representation matrix on the learning dynamics. We introduce a target representation matrix Φ'_t whose dynamics is $\frac{d}{dt}\Phi'_t = \beta(\Phi_t - \Phi'_t)$, i.e., based on a moving average update towards the main representation matrix Φ_t. With the target representation matrix in place, the trace objective still improves overall, but the improvement is likely to be non-monotonic.

Non-optimal predictor. Our theory has suggested that the optimal predictor is important for the non-collapse of the self-predictive learning dynamics. To assess how sensitive the non-collapse property is to the level of *imperfection* of the predictor, at each time step t, let P_t^* be the optimal predictor. We set the prediction matrix as a corrupted version of the optimal prediction matrix,

$$P_t = P_t^* + \epsilon,$$

where $\epsilon \in \mathbb{R}^{k\times k}$ is a noise matrix whose entries are sampled i.i.d. from Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Here, σ determines the level of noise used for corrupting the predictor. This is meant to emulate a practical setup where the prediction matrix is not learned perfectly.

In Fig. 10(b), we show the cosine similarity between $\phi_{1,t}$ and $\phi_{2,t}$ after a fixed number of iterations $T = 10000$. As the noise scale σ increases, we observe an increasing tendency to collapse.

G.2. Deep RL experiments

We provide details on the deep RL experiments.

We compare the deep bidirectional self-predictive learning algorithm, an algorithm inspired from the bidirectional self-predictive learning dynamics in Equation (8), with BYOL-RL (Guo et al., 2020). BYOL-RL can be understood as an application of self-predictive learning dynamics in the POMDP case. BYOL-RL is built on V-MPO (Song et al., 2020), an actor-critic algorithm which shapes the representation using policy gradient, without explicit representation learning objectives. See Appendix E for a more detailed description of the BYOL-RL agent and how it is adapted to allow for bidirectional self-predictive learning.

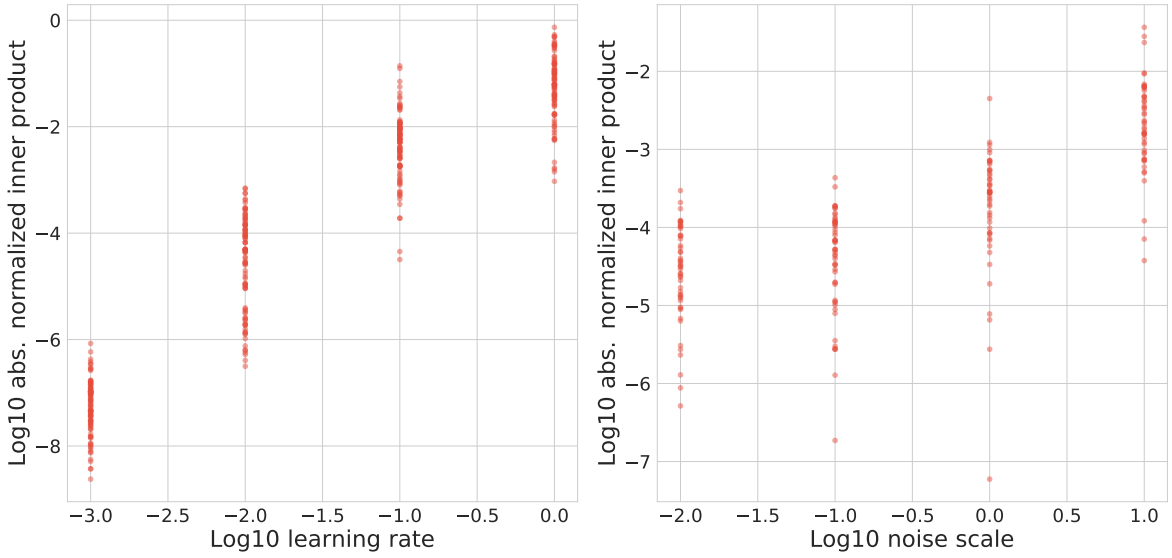
BYOL-RL implements a forward prediction loss function L_{fwd}, which is combined with V-MPO's RL loss function

$$L_{\text{BYOL-RL}} = L_{\text{rl}} + L_{\text{fwd}}.$$

The bidirectional self-predictive learning algorithm introduces a backward prediction loss function

$$L_{\text{bidirectional}} = L_{\text{rl}} + L_{\text{fwd}} + \alpha L_{\text{bwd}},$$

where $\alpha \geq 0$ is the only extra hyper-parameter we introduce. Throughout, we set $\alpha = 1$ which introduces an equal weight between the forward and backward predictions, as this most strictly adheres to the theory. All hyper-parameters and



(a) Finite learning rate

(b) Non-optimal predictor

Figure 10. Ablation experiments to assess the sensitivity of the non-collapse property to finite learning rate and non-optimal prediction matrix. Across all plots, y-axis shows the cosine similarity $\langle \phi_{1,t}, \phi_{2,t} \rangle$ after some iterations of learning. Since the two vectors are initialized to be orthogonal, as the inner product increases from 0 to 1, we expect the representations to collapse to the same direction.

architecture are shared across experiments wherever possible. We refer readers to Guo et al. (Guo et al., 2020) for complete information on the network architecture and hyper-parameters.

Test bed. Our test bed is DMLab-30, a collection of 30 diverse partially observable cognitive tasks in the 3D DeepMind Lab (Beattie et al., 2016). DMLab-30 has visual input v_t to the agent, along with the agent's previous action a_{t-1} and reward function r_{t-1}, form the observation at time t: $o_t = (v_t, a_{t-1}, r_{t-1})$.

To better illustrate the importance of representation learning, we consider the multi-task setup where the agent is required to solve all 30 tasks simultaneously. In practice, at each episode, the agent uniformly samples an environment out of the 30 tasks and generates a sequence of experience. Since the task id is not provided, the agent needs to implicitly infer the task while interacting with the sampled environment. This intuitively explains why representation learning is valuable in such a setting, as observed in prior work (Guo et al., 2019).

In Fig. 11, we compare the per-game performance of BYOL-RL with the RL baseline, measured in terms of the human normalized scores. Here, let z_i be the raw score for the i-th game, u_i the raw score of a random policy and h_i the raw score of humans, then the human normalized score is calculated as $\frac{z_i - u_i}{h_i - u_i}$. Indeed, we see that BYOL-RL significantly out-performs the RL baseline across most games.

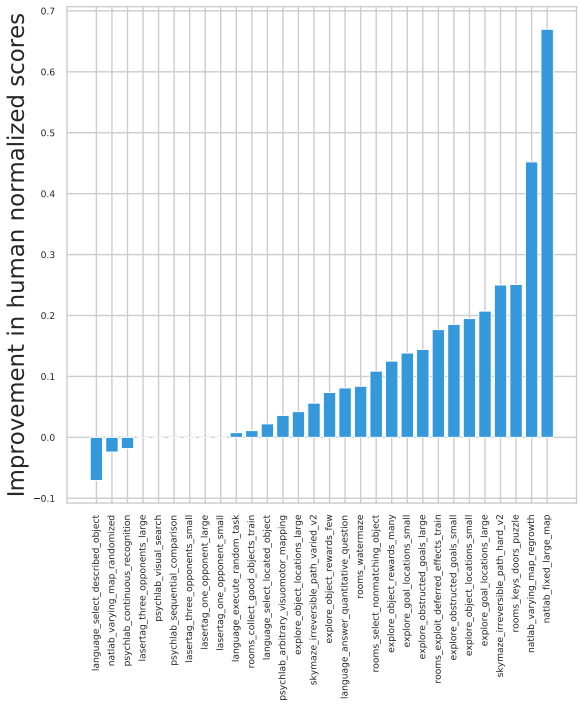


Figure 11. Per-game improvement of BYOL-RL compared to baseline RL algorithm, in terms of mean human normalized scores averaged across 3 seeds. The scores are obtained at the end of training. The improvement in performance is significant in most games.