A Unifying Framework for Action-Conditional Self-Predictive Reinforcement Learning

Khimya Khetarpal[*,†,‡] **Zhaohan Daniel Guo**[*,†] **Bernardo Avila Pires**[†]

Yunhao Tang[†] **Clare Lyle**[†] **Mark Rowland**[†] **Nicolas Heess**[†]

Diana Borsa[†] **Arthur Guez**[†] **Will Dabney**[†]

Abstract

Learning a good representation is a crucial challenge for reinforcement learning (RL) agents. Self-predictive algorithms jointly learn a latent representation and dynamics model by bootstrapping from future latent representations (BYOL). Recent work has developed theoretical insights into these algorithms by studying a continuous-time ODE model in the case of a fixed policy (BYOL-Π); this assumption is at odds with practical implementations, which explicitly condition their predictions on future actions. In this work, we take a step towards bridging the gap between theory and practice by analyzing an action-conditional self-predictive objective (BYOL-AC) using the ODE framework. Interestingly, we uncover that BYOL-Π and BYOL-AC are related through the lens of variance. We unify the study of these objectives through two complementary lenses; a model-based perspective, where each objective is related to low-rank approximation of certain dynamics, and a model-free perspective, which relates the objectives to modified value, Q-value, and Advantage functions. This mismatch with the true value functions leads to the empirical observation (in both linear and deep RL experiments) that BYOL-Π and BYOL-AC are either very similar in performance across many tasks or task-dependent.

1 Introduction

Learning a *meaningful* representation and a *useful* model of the world are among the key challenges in reinforcement learning (RL). Self-predictive learning has facilitated representation learning often by training auxiliary tasks (Lee et al., 2021), and making predictions on future observations (Schrittwieser et al., 2020) geared towards control (Jaderberg et al., 2017; Song et al., 2020). The bootstrap-your-own-latent [BYOL] framework (Grill et al., 2020) together with its RL variant (Guo et al., 2020) [BYOL-RL] offers a self-predictive paradigm for learning representations by minimizing the prediction error of its own future latent representations. Despite empirical advancements (Guo et al., 2022; Schwarzer et al., 2020), using BYOL for learning transition dynamics P in conjunction with the state representation Φ remains under-investigated from a theoretical perspective. Our work takes a step to fill in the theoretical gaps and deepen our understanding by characterizing the ODE dynamics of various BYOL objectives in the context of Markov decision processes (MDPs).

Previous work (Tang et al., 2023) provides initial important theoretical insights by considering a two-timescale, semi-gradient objective and analyzes it from an ODE perspective. A notable component

[*]Equal Contribution
[†]Google Deepmind
[‡]Correspondence to Khimya Khetarpal <khimya@google.com>.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of this objective [BYOL-Π], considers making a future prediction conditioned on a fixed policy π. This is in contrast to implementations commonly used in practice, where the future prediction is conditioned on the actions (Guo et al., 2022). Recently, Ni et al. (2024) provide analysis in the action-conditional POMDP case, but do not fully extend the analysis done by Tang et al. (2023).

In this work, we close the gap between the theoretical analysis of Tang et al. (2023) and the practical implementations of BYOL by conditioning on the action. We begin by analyzing an action-conditional BYOL loss (Eq. 1 [BYOL-AC]). We then show how the learned representations of BYOL-Π (Φ) and BYOL-AC (Φ_{ac}) satisfy a variance relation related to the eigenvalues of the dynamics (Remark 1), and we denote the variance term as BYOL-VAR (Eq. 17). We then further unify all three objectives through two complementary lenses: 1) a model-based lens where we relate each objective to a low-rank approximation of a certain dynamics matrices (Theorem 3); and 2) a model-free lens where each objective corresponds to modified value, Q-value and advantage functions (Theorem 4).

However because there is a discrepancy between the modified value functions of Theorem 4 and the true value functions, we empirical investigate the these representations in a linear setting (Appendix D.1) in how well they fit to the true value functions (Table 2), and their performance in a deep RL setting with Minigrid, and classic control domains. Surprisingly, both BYOL-Π and BYOL-AC fit the true value and Q-value functions very well, with BYOL-VAR fitting the the true advantage the best (Table 2). For deep RL, BYOL-Π and BYOL-AC again perform very similarly, with BYOL-AC slightly edging it out, while BYOL-VAR (Φ_{var}) is, as expected, a poor representation to use directly for RL since it ignores features useful for the value/Q-value function.

2 Understanding Action-Conditional BYOL (BYOL-AC)

Suppose we are the finite MDP setting. The per-action transition matrix is T_a. Assume policy π is uniform. We start with an action-conditional BYOL-AC self-predictive ODE and analyze it in a similar manner to BYOL-Π ODE (Appendix A.1), except we now have a predictor matrix P_a per action instead of a single predictor P in BYOL-Π. For brevity, assumptions and Lemmas are in the Appendix, while we summarize the main theorems here. The goal is then to minimize the following reconstruction loss in the latent space:

$$\min_{\Phi,\{\forall P_a\}} \text{BYOL-AC}(\Phi, P_{a_1}, P_{a_2}, \dots) := \mathbb{E}_{x\sim d_X, a\sim\pi(\cdot|x), y\sim T_a(\cdot|x)} \left[\left\| P_a^T \Phi^T x - \text{sg}(\Phi^T y) \right\|^2 \right] \quad (1)$$

The ODE system we consider is a two-timescale optimization, where we first solve for the optimal P_a, followed by a gradient step for Φ:

$$\forall a: \; P_a^* \in \arg\min_{P_a} \text{BYOL-AC}(\Phi, P_a), \quad \dot{\Phi} = -\nabla_\Phi \text{BYOL-AC}(\Phi, P_a)\big|_{P_a = Pa^*} \quad (2)$$

Theorem 1 (BYOL-AC ODE)**.** *Under Assumptions 1 to 6, let Φ_{ac}^* be any maximizer of the trace objective $f_{\text{BYOL-AC}}(\Phi)$:*

$$\Phi_{ac}^* \subseteq \arg\max_\Phi \; f_{\text{BYOL-AC}}(\Phi) = \arg\max_\Phi \; |A|^{-1} \textstyle\sum_a \text{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right) \quad (3)$$

Then Φ_{ac}^* is a critical point of the ODE. Furthermore, the columns of Φ_{ac}^* span the same subspace as the top-k eigenvectors of $\left(|A|^{-1} \sum_a T_a^2\right)$.

This result is similar to BYOL-Π (Appendix Theorem 5), but instead of $(T^\pi)^2$, the columns of Φ_{ac}^* span the same subspace corresponding to top-k eigenvectors of $\left(|A|^{-1} \sum_a T_a^2\right)$.

3 Unifying BYOL-Π and BYOL-AC through the lens of Variance

3.1 Variance Relation between BYOL-Π, BYOL-AC and BYOL-VAR

The learned representation for BYOL-Π (Φ^*) picks according to the eigenvalues of $(T^\pi)^2$ (Appendix Theorem 5). From Theorem 1, we know that Φ_{ac}^* picks according to $\left(|A|^{-1} \sum_a D_a^2\right)$ where D_a is the diagonal of eigenvalues of T_a. Hence these two quantities can be related through the following variance relation (for a graphical example see Figure 3).

Remark 1 (Complete Variance Relation)**.**

$$\underbrace{\mathbb{E}_{a\sim\text{Unif}}\left[D_a^2\right]}_{\textit{BYOL-AC}} = \underbrace{\left(\mathbb{E}_{a\sim\text{Unif}}\left[D_a\right]\right)^2}_{\textit{BYOL}-\Pi} + \underbrace{\text{Var}_{a\sim\text{Unif}}(D_a)}_{\textit{BYOL-VAR}}$$

We denote the variance term as BYOL-VAR. This is because BYOL-VAR also has a corresponding ODE objective, which, like in the variance equation, is a difference of the BYOL-AC and BYOL-Π objectives (BYOL-VAR = BYOL-AC − BYOL-Π). Analogous to our previous results, we derive the corresponding ODE dynamics;

$$P^* \in \arg\min_P \mathbb{E}\left[\left|\left|P^T\Phi^T x - \text{sg}(\Phi^T y)\right|\right|^2\right], \quad \forall a : P_a^* \in \arg\min_{P_a} \mathbb{E}\left[\left|\left|P_a^T\Phi^T x - \text{sg}(\Phi^T y)\right|\right|^2\right]$$
$$\dot{\Phi} = -\nabla_\Phi \text{BYOL-VAR}(\Phi, P, P_{a_1}, P_{a_2}, \dots)\big|_{P=P^*, P_a=Pa^*} \tag{4}$$

Theorem 2 (BYOL-VAR ODE)**.** *Under Assumptions 1 to 6, let $\Phi^*_{\textit{VAR}}$ be any maximizer of the trace objective $f_{\textit{BYOL-VAR}}(\Phi)$:*

$$\Phi^*_{\textit{VAR}} \subseteq \arg\max_\Phi \ |A|^{-1}\textstyle\sum_a \text{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi\right) - \text{Tr}\left(\Phi^T T^\pi \Phi\Phi^T T^\pi \Phi\right) \tag{5}$$

Then $\Phi^*_{\textit{VAR}}$ is a critical point of the ODE. Furthermore, the columns of $\Phi^*_{\textit{VAR}}$ span the same subspace as the top-k eigenvectors of $\left(|A|^{-1}\sum_a T_a^2 - (T^\pi)^2\right)$.

4 Two Unifying Perspectives: Model-Based and Model-Free

In addition to the variance relationship, we further unify the study of all three objectives through two complimentary lenses, namely, a model-based perspective and a model-free perspective.

4.1 Fitting Dynamics - A Model-Based View

From the model-base perspective, we can derive an equivalence between each of the trace objectives akin to finding a low-rank approximation of certain transition dynamics.

Theorem 3 (Unifying Model-Based View)**.** *Under Assumptions 1 to 6, the negative trace objectives of BYOL-Π, BYOL-AC, and BYOL-VAR are equivalent (up to a constant C) to the following objectives ($\|\cdot\|_F$ is the Frobenius matrix norm):*

$$-f_{\textit{BYOL-}\Pi}(\Phi) = \min_P \ \|T^\pi - \Phi P \Phi^T\|_F + \text{C} \tag{6}$$
$$-f_{\textit{BYOL-AC}}(\Phi) = |A|^{-1}\textstyle\sum_a \min_{P_a} \ \|T_a - \Phi P_a \Phi^T\|_F + \text{C} \tag{7}$$
$$-f_{\textit{BYOL-VAR}}(\Phi) = |A|^{-1}\textstyle\sum_a \min_{P_{\Delta a}} \ \|(T_a - T^\pi) - \Phi P_{\Delta a} \Phi^T\|_F + \text{C} \tag{8}$$

Therefore, maximizing the trace (over orthogonal Φ) results in BYOL-Π, BYOL-AC, and BYOL-VAR trying to fit a low-rank approximation of the dynamics matrix T^π, per-action transition matrix T_a, and the residual dynamics $(T_a - T^\pi)$ respectively.

4.2 Fitting Value Functions - A Model-Free View

Complimentary to the model-based view, we can rewrite the the maximizer to the trace objectives through a model-free lens. To do this, we assume an isotropic Gaussian reward function R i.e. $E[RR^T] = |\mathcal{X}|^{-1}I$. We can re-express the maximizers to the trace objectives as modified 1-step future reward functions over these reward functions R.

Theorem 4 (Unifying Model-Free View). *Under Assumptions 1 to 6, the negative trace objectives of BYOL-Π, BYOL-AC, and BYOL-VAR are equivalent (up to a constant* C*) to the following objectives:*

$$-f_{\text{BYOL-}\Pi}(\Phi) = |\mathcal{X}|\mathbb{E}\left[\min_{\theta,\omega} \left(\|T^\pi R - \Phi\theta\|^2 + \|T^\pi \Phi\Phi^T R - \Phi\omega\|^2\right)\right] + \mathrm{C} \tag{9}$$

$$-f_{\text{BYOL-AC}}(\Phi) = |\mathcal{X}|\mathbb{E}\left[|A|^{-1}\sum_a \min_{\theta_a,\omega_a} \left(\|T_a R - \Phi\theta_a\|^2 + \|T_a\Phi\Phi^T R - \Phi\omega_a\|^2\right)\right] + \mathrm{C} \tag{10}$$

$$-f_{\text{BYOL-VAR}}(\Phi) = |\mathcal{X}|\mathbb{E}\Big[|A|^{-1}\sum_a \min_{\theta_a,\omega_a} \left(\|(T_a R - T^\pi R) - \Phi\theta\|^2 + \|(T_a\Phi\Phi^T R - T^\pi\Phi\Phi^T R) - \Phi\omega\|^2\right)\Big] + \mathrm{C} \tag{11}$$

Note that each of these is made up of two terms, where the first term (e.g. $\|T^\pi R - \Phi\theta\|^2$) is fitting the true 1-step value, but the second term (e.g. $\|T^\pi\Phi\Phi^T R - \Phi\omega\|^2$) fits to the projected 1-step value, hence making it a modified 1-step value function. Also, it is straightforward to generalize from 1-step to the infinite-discounted case (Appendix B.2).

5 Experiments

Since our theory with assumptions is not guaranteed to translate to empirical RL performance, we perform experiments for the BYOL-objectives in both linear (Appendix Sec. D.1) and non-linear (Sec. 5.1) function approximation settings.

5.1 Deep Reinforcement Learning

We defer details on domains and hyper parameter tuning to Appendix D.2 and D.3.

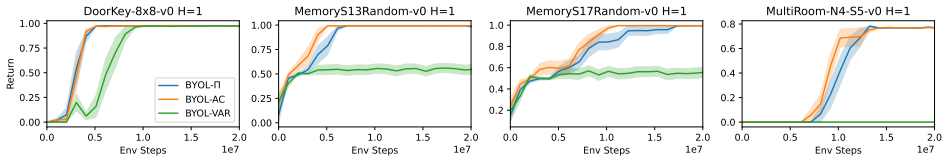

Figure 1: Comparing BYOL-Π, BYOL-AC, and BYOL-VAR in Minigrid.

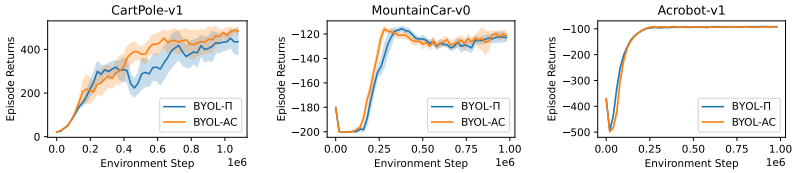

Figure 2: Comparing BYOL-Π and BYOL-AC in Classic control.

Minigrid. We consider four domains in Minigrid (Chevalier-Boisvert et al., 2023).We observe in Figure 1 that Φ_{ac} outperforms the other baselines in 3 out of 4 tasks, Φ is on par with BYOL-AC in 1 out of 4 tasks. Φ_{var} is poor in all 4 tasks; because the objective is a difference of BYOL-AC and BYOL-Π, it ends up trying to remove features that are good for BYOL-Π, resulting in a worse representation for RL. More results for multistep action predictions in Appendix Fig. 5.

Control. We consider 3 of open-AI gym's (Brockman et al., 2016) classic control environments (Sutton and Barto, 2018). We report that Φ_{ac} outperforms Φ in CartPole-v1 (leftmost) and MountainCar-v0 (center), whereas BYOL-AC performs on par with BYOL-Π in Acrobot-v1.

6 Discussion

In summary, we extended previous theoretical analysis to an action-conditional BYOL-AC objective, showing that it learns spectral information about per-action transition dynamics T_a. We discovered a variance equation that relates BYOL, BYOL-AC, and a novel variance-like objective BYOL-VAR (Remark 1) learns spectral information pertaining to the residual $(T_a - T^\pi)$. We unified the three objectives, firstly through a model-based lens, related to low-rank approximation of dynamics, and secondly through a model-free lens, establishing the connection to modified value functions. The **key takeaway** is that BYOL-Π and BYOL-AC are mostly on par with each other, with which one is better mostly being task-dependent. **Future work** further relaxing the assumptions in our theoretical analysis, with potential to generalize the theory. Also, since BYOL-VAR is concerned with features that distinguish between actions, it may be useful for learning action representations or even option discovery.

References

B. Amos, L. Dinh, S. Cabi, T. Rothörl, S. G. Colmenarejo, A. Muldal, T. Erez, Y. Tassa, N. de Freitas, and M. Denil. Learning awareness models. In *International Conference on Learning Representations*, 2018.

B. Behzadian, S. Gharatappeh, and M. Petrik. Fast feature selection for linear value function approximation. In *Proceedings of the International Conference on Automated Planning and Scheduling*, 2019.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI gym. *arXiv*, 2016.

Y. Chandak, S. Thakoor, Z. D. Guo, Y. Tang, R. Munos, W. Dabney, and D. L. Borsa. Representations and exploration for deep reinforcement learning using singular value decomposition. In *Proceedings of the International Conference on Machine Learning*, 2023.

X. Chen and K. He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

M. Chevalier-Boisvert, B. Dai, M. Towers, R. de Lazcano, L. Willems, S. Lahlou, S. Pal, P. S. Castro, and J. Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *arXiv*, 2023.

DeepMind, I. Babuschkin, K. Baumli, A. Bell, S. Bhupatiraju, J. Bruce, P. Buchlovsky, D. Budden, T. Cai, A. Clark, I. Danihelka, A. Dedieu, C. Fantacci, J. Godwin, C. Jones, R. Hemsley, T. Hennigan, M. Hessel, S. Hou, S. Kapturowski, T. Keck, I. Kemaev, M. King, M. Kunesch, L. Martens, H. Merzic, V. Mikulik, T. Norman, G. Papamakarios, J. Quan, R. Ring, F. Ruiz, A. Sanchez, L. Sartran, R. Schneider, E. Sezener, S. Spencer, S. Srinivasan, M. Stanojević, W. Stokowiec, L. Wang, G. Zhou, and F. Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

K. Ferguson and S. Mahadevan. Proto-transfer learning in Markov decision processes using spectral methods. In *ICML Workshop on Structural Knowledge Transfer for Machine Learning*, 2006.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Z. Guo, S. Thakoor, M. Pîslar, B. Avila Pires, F. Altché, C. Tallec, A. Saade, D. Calandriello, J.-B. Grill, Y. Tang, M. Valko, R. Munos, M. Gheshlaghi Azar, and B. Piot. BYOL-Explore: Exploration by bootstrapped prediction. In *Advances in neural information processing systems*, 2022.

Z. D. Guo, B. A. Pires, B. Piot, J.-B. Grill, F. Altché, R. Munos, and M. G. Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020.

J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee. Flax: A neural network library and ecosystem for JAX, 2023. URL http://github.com/google/flax.

J. Hunter and D. Dale. The matplotlib user's guide. *Matplotlib 0.90. 0 user's guide*, 2007.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *Proceedings of the International Conference on Learning Representations*, 2017.

C. L. Lan, S. Tu, M. Rowland, A. Harutyunyan, R. Agarwal, M. G. Bellemare, and W. Dabney. Bootstrapped representations in reinforcement learning. 2023.

R. T. Lange. gymnax: A JAX-based reinforcement learning environment library, 2022. URL http://github.com/RobertTLange/gymnax.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, 2021.

M. Littman and R. S. Sutton. Predictive representations of state. In *Advances in Neural Information Processing Systems*, 2001.

C. Lyle, M. Rowland, G. Ostrovski, and W. Dabney. On the effect of auxiliary tasks on representation dynamics. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021.

M. C. Machado, C. Rosenbaum, X. Guo, M. Liu, G. Tesauro, and M. Campbell. Eigenoption discovery through the deep successor representation. In *Proceedings of the International Conference on Learning Representations*, 2018.

J. Mawhin. Chapter 51 - Alexandr Mikhailovich Lyapunov, Thesis on the stability of motion (1892). In I. Grattan-Guinness, R. Cooke, L. Corry, P. Crépel, and N. Guicciardini, editors, *Landmark Writings in Western Mathematics 1640-1940*, pages 664–676. Elsevier Science, Amsterdam, 2005.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

T. Ni, B. Eysenbach, E. Seyedsalehi, M. Ma, C. Gehring, A. Mahajan, and P.-L. Bacon. Bridging state and history representations: Understanding self-predictive RL. In *Proceedings of the International Conference on Learning Representations*, 2024.

J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, 2015.

T. E. Oliphant et al. *Guide to numpy*, volume 1. Trelgol Publishing USA, 2006.

T. Ren, T. Zhang, L. Lee, J. E. Gonzalez, D. Schuurmans, and B. Dai. Spectral decomposition representation for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. Mastering Atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

M. Schwarzer, A. Anand, R. Goel, R. D. Hjelm, A. Courville, and P. Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

H. F. Song, A. Abdolmaleki, J. T. Springenberg, A. Clark, H. Soyer, J. W. Rae, S. Noury, A. Ahuja, S. Liu, D. Tirumala, et al. V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control. In *Proceedings of the International Conference on Learning Representations*, 2020.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. 2018.

Y. Tang, Z. D. Guo, P. H. Richemond, B. Á. Pires, Y. Chandak, R. Munos, M. Rowland, M. G. Azar, C. L. Lan, C. Lyle, et al. Understanding self-predictive learning for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

G. Teschl. *Ordinary differential equations and dynamical systems*. American Mathematical Soc., 2012.

E. Toledo, L. Midgley, D. Byrne, C. R. Tilbury, M. Macfarlane, C. Courtot, and A. Laterre. Flashbax: Streamlining experience replay buffers for reinforcement learning with jax, 2023. URL https://github.com/instadeepai/flashbax/.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Appendix / Supplemental Material

A Preliminaries

Reinforcement Learning. Consider an MDP $\langle \mathcal{X}, \mathcal{A}, T_a, \gamma \rangle$, where $\mathcal{X}$ is a finite set of states, $\mathcal{A}$ a finite set of actions, $x, y \in \mathbb{R}^{|\mathcal{X}|\times 1}$ assume tabular state representation where each state is a one hot vector, $T_a \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is the per-action transition dynamics defined as $(T_a)_{ij} \coloneqq p(y = j \mid x = i, a)$, and $\gamma \in [0, 1)$ the discount factor. Given $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ we let $T^\pi : \mathcal{X} \to \mathscr{P}(\mathcal{X})$ be the state transition kernel induced by the policy π, that is, $(T^\pi)_{ij} \coloneqq \sum_a \pi(a \mid x = i)(T_a)_{ij}$. Given a (deterministic) reward function of the state $R \in \mathbb{R}^{|\mathcal{X}|\times 1}$, the value function is defined as $V^\pi \coloneqq (I - \gamma T^\pi)^{-1} R$. The Q-value function is defined as $Q^\pi_a \coloneqq R + \gamma T_a V^\pi$.

Representation Learning. It can be shown under idealized conditions that value function estimation methods such as TD-learning capture the top-k subspace of the eigenbasis of T^π (Lyle et al., 2021). Moreover, the top-k eigenvectors of T^π are exactly the same as those of $(I - \gamma T^\pi)^{-1}$ (Chandak et al., 2023). Notably, when the value function is linear in the features, the basis vectors are useful features of the state.

Ordinary differential equations (ODEs) arise often in the context of describing change in the representations over time as they are being learned. Lyle et al. (2021) consider the dynamics for single-step TD learning and assume that the weight is kept fixed, which greatly simplifies the dynamics. Tang et al. (2023) consider the ODEs systems with certain dynamics that make them not the same as traditional optimization problems. To analyze such an ODE system, one constructs a Lyapunov function, which can be considered a surrogate loss function that the ODE monotonically minimizes. Lyapunov functions are also useful to show necessary and sufficient conditions for stability and convergence of the said ODE (Mawhin, 2005; Teschl, 2012).

A.1 BYOL ODE with Fixed Policy: BYOL-Π

We start with the setting considered by Tang et al. (2023). We observe triples (x, a, y) where $x \sim d_X$ is a one-hot state, $a \sim \pi(\cdot \mid x)$ is an action sampled according to a fixed policy π, and $y \sim p(\cdot \mid x, a)$ is the on-hot state observed after taking action a at state x. Letting $D_X \coloneqq \mathbb{E}[xx^T] = \text{diag}(d_X)$, we can write $\mathbb{E}[xy^T] = D_X T^\pi$. The goal is to learn a representation matrix $\Phi \in \mathbb{R}^{|\mathcal{X}|\times k}$ that embeds each state x as a k-dimensional real vector denoted by $\Phi^T x$. To model the transitions in latent space, $\Phi^T x \to \Phi^T y$, we consider a latent linear map $P \in \mathbb{R}^{k\times k}$. To learn Φ, we use a self-predictive objective that minimizes the loss in the latent space,

$$\min_{\Phi, P} \text{ BYOL-}\Pi(\Phi, P) \coloneqq \mathbb{E}_{x\sim d_X, y\sim T^\pi(\cdot|x)} \left[\left|\left| P^T \Phi^T x - \text{sg}(\Phi^T y) \right|\right|_2^2 \right] \tag{12}$$

where sg is a stop-gradient operator on the prediction target to help in avoiding degenerate solutions. The appeal of this objective is that everything is defined in the latent space, which means this can be easily extended and implemented in practice with P and Φ replaced by neural networks. However, this objective still has the trivial solution of $\Phi = 0$. To avoid this, Tang et al. (2023) formulate a two-timescale optimization process wherein we first solve the inner minimization w.r.t. (with respect to) P before taking a (semi-)gradient step (denoted as $\dot{\Phi}$) w.r.t. Φ.

$$P^* \in \arg\min_P \text{ BYOL-}\Pi(\Phi, P), \quad \dot{\Phi} = -\nabla_\Phi \text{ BYOL-}\Pi(\Phi, P)\big|_{P=P^*} \tag{13}$$

This is an ODE system for Φ with dynamics (gradient) $\dot{\Phi}$. Tang et al. (2023) makes the following simplifying assumptions to analyze it:

Assumption 1 (Orthogonal Initialization)**.** *Φ is initialized to be orthogonal i.e. $\Phi^T\Phi = I$.*

Assumption 2 (Uniform State Distribution)**.** *The state distribution d_X is uniform.*

Assumption 3 (Symmetric Dynamics)**.** *T^π is symmetric i.e. $T^\pi = (T^\pi)^T$.*

While these assumptions are quite strong and impractical, we believe the resulting theoretical insights are a useful perspective in understanding and characterizing the learned representation Φ in practice.

Lemma 1 (Non-collapse, Tang et al., 2023)**.** *Under Assumption 1, we have that $\Phi^T\dot{\Phi} = 0$, which means that $\Phi^T\Phi = I$ is preserved for all Φ throughout the ODE process.*

Intuitively, Lemma 1 suggests that because of how we set up the ODE with the semi-gradient and two-timescale optimization, an orthogonal initialization means we can avoid all trivial solutions.

Lemma 2 (BYOL Trace Objective, Tang et al., 2023)**.** *Under Assumptions 1 to 3, a Lyapunov function for the ODE is the negative of the following trace objective*

$$f_{\textit{BYOL-}\Pi}(\Phi) \coloneqq \operatorname{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) . \tag{14}$$

This means the ODE converges to some critical point.

By construction, the critical points of the ODE are also critical points of the latent loss, so Lemma 2 establishes that the ODE converges to such a non-collapsed critical point.

Theorem 5 (BYOL-Π ODE, Tang et al., 2023)**.** *Under Assumptions 1 to 3, let Φ^* be any maximizer of the trace objective $f_{\textit{BYOL-}\Pi}(\Phi)$:*

$$\Phi^* \subseteq \arg\max_\Phi \; f_{\textit{BYOL-}\Pi}(\Phi) = \arg\max_\Phi \; \operatorname{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) . \tag{15}$$

Then Φ^* is a critical point of the ODE. Furthermore, the columns of Φ^* span the same subspace as the top-k eigenvectors of $(T^\pi)^2$.

This trace objective $f_{\text{BYOL-}\Pi}(\Phi)$ is essentially a surrogate loss function that the ODE is monotonically maximizing, that also has the same critical points by construction. Thus to understand the ODE, we simply analyze the maximizer of the trace objective. In this case, the maximizer Φ^* learns important features (eigenvectors) of the transition dynamics T^π. We demonstrate this in Appendix Fig. 6, for both symmetric and non-symmetric MDPs.

B Theoretical Results

BYOL-AC. To analyze this ODE, we make a few new assumptions in addition to Assumptions 1 to 3. For additional intuition on theory, we defer the reader to Appendix F. All proofs are in Appendix G.

Assumption 4 (Uniform Policy)**.** *The (data-collection) policy π is uniform across all actions.*

We also make an analogue of Assumption 3 for the action-conditional setting:

Assumption 5 (Symmetric Per-action Dynamics)**.** *T_a is symmetric for all actions i.e. $T_a = (T_a)^T$.*

We establish analogues of Lemmas 1 and 2 and Theorem 5 for BYOL-AC, that is, the non-collapse property of BYOL-AC, a trace objective that describes a Lyapunov function for the ODE, and a main theorem that helps us understand what kind of representation BYOL-AC learns.

Lemma 3 (Non-collapse BYOL-AC)**.** *Under Assumption 1, we have that $\Phi^T \dot{\Phi} = 0$, which means that $\Phi^T \Phi = I$ is preserved for all Φ throughout the BYOL-AC ODE process.*

Lemma 4 (BYOL-AC Trace Objective)**.** *Under Assumptions 1 to 5, a Lyapunov function for the BYOL-AC ODE is the negative of the following trace objective*

$$f_{\textit{BYOL-AC}}(\Phi) \coloneqq |A|^{-1} \textstyle\sum_a \operatorname{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right) \tag{16}$$

This means the ODE converges to some critical point.

Before presenting our main theorem, we require the following assumption.

Assumption 6 (Common Eigenvectors)**.** *For all actions a, we have the eigen decomposition $T_a = Q D_a Q^T$, i.e. all T_a share the same eigenvectors.*

BYOL-VAR. Comparing BYOL and BYOL-AC, we note that we can relate their trace objective maximizers using a variance equation (Remark 1), with BYOL being the square of the first moment, and BYOL-AC being the second moment. This poses a natural question: Is there an objective corresponding to the variance term i.e. the difference between the second moment and the square of the first moment? We answer this question in the affirmative by proposing a new variance-like BYOL objective:

$$\min_\Phi \text{BYOL-VAR}(\Phi, P, P_{a_1}, P_{a_2}, \dots) \coloneqq \mathbb{E}\left[\|P_a^\top \Phi^\top x - \text{sg}(\Phi^\top y))\|^2 - \|P^\top \Phi^\top x - \text{sg}(\Phi^\top y))\|^2\right] \tag{17}$$

But now with the predictors as before, solving

$$\min_P \mathbb{E}\left[\|P^\top \Phi^\top x - \Phi^\top y\|^2\right], \qquad \text{and} \qquad \forall a : \min_{P_a} \mathbb{E}\left[\|P_a^\top \Phi^\top x - \Phi^\top y\|^2\right].$$

The BYOL-VAR objective is a difference of the BYOL-AC and BYOL-Π objectives. Analogous to our previous results, we derive the corresponding ODE dynamics, with statements about non-collapse, a Lyapunov function, and a result about what the representation captures.

$$P^* \in \arg\min_P \mathbb{E}\left[\left|\left|P^T \Phi^T x - \text{sg}(\Phi^T y)\right|\right|^2\right], \quad \forall a : P_a^* \in \arg\min_{P_a} \mathbb{E}\left[\left|\left|P_a^T \Phi^T x - \text{sg}(\Phi^T y)\right|\right|^2\right]$$

$$\dot{\Phi} = -\nabla_\Phi \text{BYOL-VAR}(\Phi, P, P_{a_1}, P_{a_2}, \dots)\big|_{P=P^*, P_a=Pa^*} \tag{18}$$

Lemma 5 (Non-collapse BYOL-VAR)**.** *Under Assumptions 1 and 4, we have that $\Phi^T \dot{\Phi} = 0$, which means that $\Phi^T \Phi = I$ is preserved for all Φ throughout the BYOL-VAR ODE process.*

Lemma 6 (BYOL-VAR Trace Objective)**.** *Under Assumptions 1 to 5, a Lyapunov function for the BYOL-VAR ODE is the negative of the following trace objective*

$$\begin{aligned} f_{\textit{BYOL-VAR}}(\Phi) &\coloneqq f_{\textit{BYOL-AC}}(\Phi) - f_{\textit{BYOL-}\Pi}(\Phi) \\ &= |A|^{-1} \sum_a \text{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right) - \text{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) \end{aligned} \tag{19}$$

This means the ODE converges to some critical point.

Intuitively, BYOL-VAR tries to learn a representation Φ^*_{var} that only captures features for distinguishing between actions. In practice, our assumptions are unlikely to be satisfied. However, the intuition behind the variance relation (Remark 1) still gives us a valuable insight: Φ^* is concerned with meaningful features for T^π; Φ^*_{ac} tries to capture meaningful features of T_a; Φ_{var} tries to only capture features that can distinguish across T_a.

B.1 Relation between BYOL-Π, BYOL-AC and BYOL-VAR

Figure 3 shows an illustrative MDP with two actions demonstrating which eigenvectors each objective converges to. In the next section, we will present two unifying perspectives for comparing between the three objectives of BYOL-Π, BYOL-AC, and BYOL-VAR.

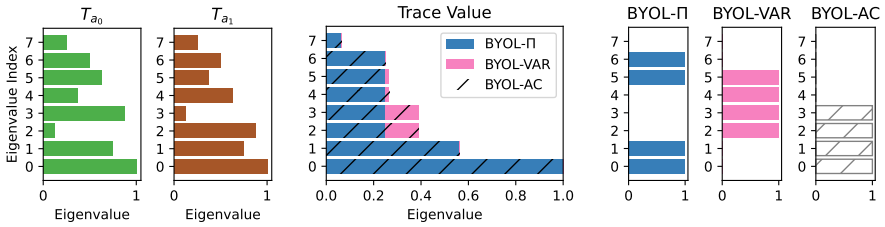

Figure 3: **On the representations across BYOL-Π, BYOL-AC, and BYOL-VAR.** We consider a simple MDP with two actions and corresponding transition functions T_{a_0}, T_{a_1}, with the eigenvalues of each action depicted in two leftmost plots. The middle plot shows a stacked bar plot of the trace objective values corresponding to each objective. The three rightmost plot shows each objective picking its top-k ($k = 4$) eigenvectors.

B.2 Generalization to Multi-Step

While our previous results consider the 1-step setting, there is a straightforward way to generalize them to the multi-step setting. To do this, it is sufficient to generalize the transition dynamics themselves to multi-step transition dynamics as follows:

$$(T^\pi)_H = \frac{1-\gamma}{1-\gamma^H} \sum_{h=1}^{H} \gamma^{h-1} (T^\pi)^h \tag{20}$$

$$(T_a)_H = \frac{1-\gamma}{1-\gamma^H} \sum_{h=1}^{H} \gamma^{h-1} T_a (T^\pi)^{h-1} \tag{21}$$

where we change the 1-step transition to a normalized, discounted mixture of h-step transitions. Note in the action-conditioned case, only the first step depends on the action and the rest of the steps follow π (like how the Q-value is defined). To implement this, we can modify the BYOL-Π objective as follows:

$$\min_{\Phi, P} \text{BYOL-}\Pi(\Phi, P) := \mathbb{E}_{x \sim d_X, y \sim (T^\pi)_H(\cdot|x)} \left[\left|\left| P^T \Phi^T x - \text{sg}(\Phi^T y) \right|\right|_2^2 \right] \tag{22}$$

where all we change is the distribution of our prediction target y to the new mixture distribution. BYOL-AC and BYOL-VAR would be similar. Note that we still satisfy the condition $(T^\pi)_H = \frac{1}{|A|} \sum_a (T_a)_H$. Consequently, all our results generalize to multi-step setting. In particular, Theorems 3 and 4 would generalize to fitting the full infinite discounted state-value (V), action-value (Q) and advantage functions as $H \to \infty$.

C Related Work

Self-Predictive Representation Learning. At the intersection of representation learning and self-supervised learning, using bootstrapped latent embeddings to train the representations has been an empirically successful approach for both image representation learning (Chen and He, 2021; Grill et al., 2020) and reinforcement learning (Guo et al., 2020).

Action-Conditioned Predictive Representations. Action-conditional predictions of the future where the prediction tasks are indicators of events on the finite observation space date back to foundational work introducing predictive state representations (PSRs) (Littman and Sutton, 2001). Deep learning approaches leveraging action-conditional predictions of the future to improve RL performance covers a broad range of ideas. For instance, interleaving an action-dependent (predictor) RNN with an observation dependent RNN (Amos et al., 2018), to predicting policies, rewards, values/logits needed for Monte Carlo tree search conditioned on actions (Schrittwieser et al., 2020) or options (Oh et al., 2015).

Understanding Predictive Representations. A key challenge in self-predictive learning is collapsing solutions. Both Grill et al. (2020) and Guo et al. (2020) propose BYOL variants where they leverage a target network to inform and train an online network for self-supervised image representation learning and reinforcement learning respectively. While Grill et al. (2020) posit the need for a momentum encoder as a key requirement for BYOL to avoid collapsing, Chen and He (2021) show that empirically the stop-gradient operation is critical to avoid collapse. Both these methods are focused on image representations in iid settings though. With a focus on RL, our work builds upon Tang et al. (2023), who identify and prove conditions for the non-collapse property of self-predictive learning through the lens of a theoretical ODE framework.

Spectral Decomposition Lens for Representations in RL. Recall that we show that BYOL-AC captures the spectral information about action transition matrices, whereas prior work shows that BYOL-Π captures spectral information about P^π. The lens of spectral decomposition of P^π or $(I - \gamma P^\pi)^{-1}$, together with eigenvector decomposition (Ferguson and Mahadevan, 2006; Lyle et al., 2021; Machado et al., 2018), and singular value decomposition (Behzadian et al., 2019; Chandak et al., 2023; Lan et al., 2023; Ren et al., 2023) greatly facilitates representation learning research for reinforcement learning.

D Additional Experiments and Details

D.1 Linear Function Approximation

First, we corroborate Theorem 4 and Theorem 3 empirically in a linear function approximation setting. We consider randomly generated MDPs with 10 states, 4 actions and symmetric per-action dynamics T_a. We learn a compressed representation with dimension 4 for each of BYOL-Π, BYOL-AC, and BYOL-VAR. Results in this section are averaged over 100 runs with 95% standard error. Table 1 shows the values of the three negative trace objectives (rows) versus the representation learned by

the three methods (columns). As predicted by the theory, we see that the smallest negative trace objective is attained by the corresponding ODE. Φ minimizes $-f_{\text{BYOL-}\Pi}$ (Eqs. (6) and (9)) i.e. **Pr(Φ is best)** is 99%, Φ_{ac} minimizes $-f_{\text{BYOL-AC}}$ (Eqs. (7) and (10)) i.e. **Pr(Φ_{ac} is the best)** 99%, whereas Φ_{var} minimizes $-f_{\text{BYOL-VAR}}$ (Eqs. (8) and (11)) i.e. **Pr(Φ_{var} is the best)** is 100%. We have additional experiments fitting value, action-value, and advantages empirically in the appendix.

Table 1: **Illustrating Theorem 3 and Theorem 4** empirically demonstrates that each method minimizes its corresponding negative trace objectives, which means BYOL-Π, BYOL-AC, and BYOL-VAR are best at capturing information about T^π, T_a, and $(T_a - T^\pi)$ respectively, and are trying to fit a certain 1-step value (V), Q-value, and Advantage function respectively.

Method	**BYOL-Π [Φ]**		**BYOL-AC [Φ_{ac}]**		**BYOL-VAR [Φ_{var}]**	
Objective		**Pr(Φ is best)**		**Pr(Φ_{ac} is best)**		**Pr(Φ_{var} is best)**
$-f_{\text{BYOL-}\Pi}$	-1.22 ± 0.00	**99%**	-1.10 ± 0.01	1%	-0.04 ± 0.00	0.%
$-f_{\text{BYOL-AC}}$	-1.31 ± 0.01	1%	-1.44 ± 0.00	**99%**	-0.55 ± 0.01	0.%
$-f_{\text{BYOL-VAR}}$	-0.09 ± 0.00	0%	-0.33 ± 0.00	0%	-0.50 ± 0.01	**100%**

Next, we consider the same three methods and fit the traditional V-MSE ($\mathbb{E}_R\left[\min_\theta \|V - \Phi\theta\|^2\right]$), Q-MSE and Advantage-MSE. Table 2 illustrates that both BYOL-Π and BYOL-AC perform competitively in fitting the state-value reporting a V-MSE of 6.32, and 6.48 respectively, while fitting an action-value suffering a Q-MSE of 8.31, and 8.01 respectively. BYOL-VAR instead learns Φ_{var} which turns out be optimal for fitting the true Advantage MSE observed to be 0.43 and **Pr(Φ_{var} is best)** to be 100%.

Table 2: **Fitting various value functions to learned representations** for Φ, Φ_{ac}, and Φ_{var}. We report both the MSE and the probability of a representation being best.

Method	**BYOL-Π [Φ]**		**BYOL-AC [Φ_{ac}]**		**BYOL-VAR [Φ_{var}]**	
Objective		**Pr(Φ is best)**		**Pr(Φ_{ac} is best)**		**Pr(Φ_{var} is best)**
V-MSE	6.32 ± 0.06	59%	6.48 ± 0.05	41%	10005.53 ± 0.05	0%
Q-MSE	8.31 ± 0.35	52%	8.01 ± 0.30	48%	10005.97 ± 0.05	0%
Advantage-MSE	0.76 ± 0.01	0%	0.61 ± 0.01	0%	0.43 ± 0.01	**100%**

Besides, we investigated the robustness of each representation to perturbations in the initial policy used to learn the representation in Appendix E. We report that Φ_{ac} learned by BYOL-AC objective is much more robust to changes in the policy compared to BYOL-Φ and BYOL-VAR.

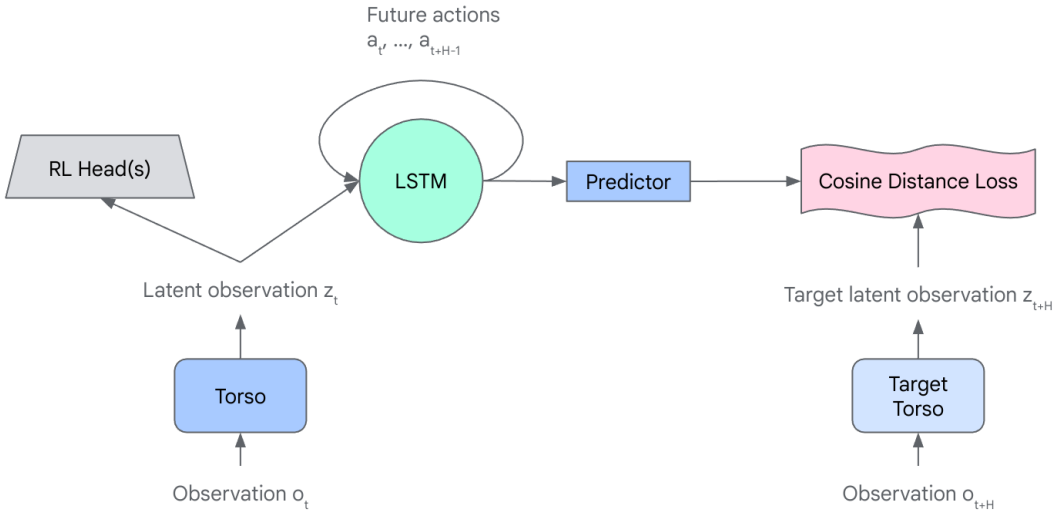

Figure 4: High Level Architecture of our RL Agent. Network details are in Appendices D.2 and D.3

D.2 V-MPO in Minigrid

We modify the implementation of a V-MPO (Song et al., 2020) agent by augmenting it with the auxiliary loss corresponding to BYOL-Π, BYOL-AC, and BYOL-VAR. Results in this section are averaged over 10 independent seeds with 95% standard error in the error bands.

In Minigrid (Chevalier-Boisvert et al., 2023), we here detail the description of each task: 1) DoorKey-8x8-v0, where the agent must pick up a key in the environment in order to unlock a door and then get to the green goal square, 2) MemoryS13Random-v0, where the agent starts in a small room with a visible object and then has to go through a narrow hallway which ends in a split. At each end of the split there is an object, one of which is the same as the object in the starting room. The agent has to remember the initial object, and go to the matching object at split, 3) MemoryS17Random-v0 is a bigger domain matching the description of the memory test in the previous environment, and 4) MultiRoom-N4-S5-v0, where the agent navigates through a series of connected rooms with doors that must be opened in order to get to the next room. The goal is to reach the final room which has the green square. Note that the 1-3 are fully observable domains, whereas 4 is a partially observable environment.

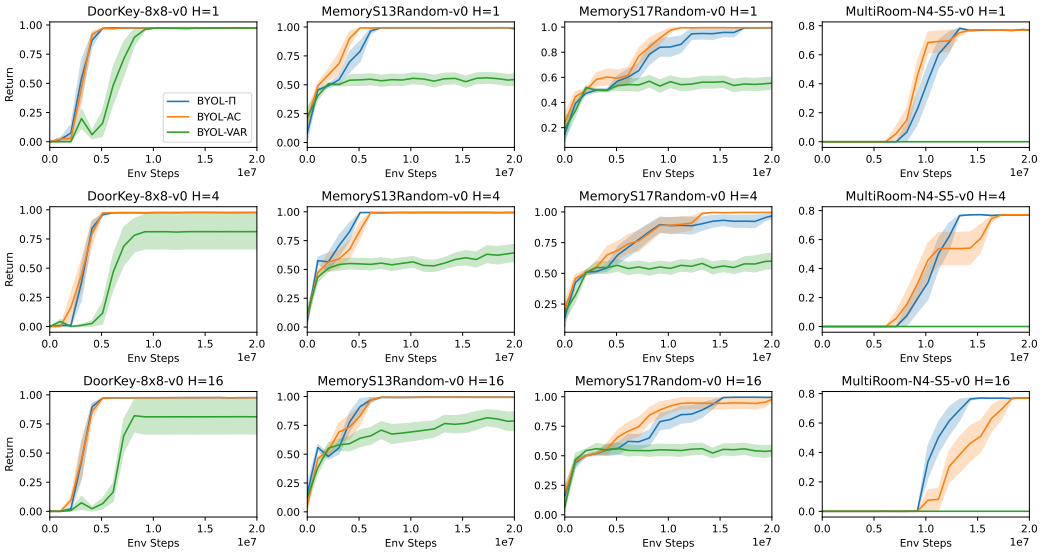

Figure 5: Comparing BYOL-Π, BYOL-AC, and BYOL-VAR on different domains in Minigrid across varying prediction horizons $H = 1, 4, 16$.

For each baseline considered in Fig. 1 and Fig 5, we first tuned the hyper-parameters of the base RL algorithm i.e. V-MPO in this case. We then tuned the BYOL-Π baseline, followed by running both BYOL-AC and BYOL-VAR for various horizons [H].

Our high level architecture is shown in Figure 4. Here are the network details:

- Torso: small ResNet
  1. Conv2D[channel=32,stride=2,kernel=3]
  2. ResBlock[channel=32,stride=1,kernel=3]
  3. ResBlock[channel=32,stride=1,kernel=3]
  4. Conv2D[channel=128,stride=2,kernel=3]
  5. ResBlock[channel=128,stride=1,kernel=3]
  6. ResBlock[channel=128,stride=1,kernel=3]
  7. Conv2D[channel=256,stride=2,kernel=3]
  8. ResBlock[channel=256,stride=1,kernel=3]
  9. ResBlock[channel=256,stride=1,kernel=3]
  10. Flatten

11. Linear[256]

- LSTM: Hidden size 256
- Predictor: MLP[128, 256, 512, 256]
- RL Head
  - Value Head: MLP[512, 1]
  - Policy Head: MLP[512, 7]

And here are our hyperparameters.

- Minibatch batch size: 48
- Minibatch sequence length: 30
- Adam Optimizer[learning_rate=1e-4, b1=0.9, b2=0.999, eps=1e-8, eps_root=1e-8]

D.3 DQN in Open-AI Gym

Next, we modify DQN's (Mnih et al., 2015) implementation from Lange (2022) by augmenting it with auxiliary losses corresponding to BYOL-Π and BYOL-AC in Figure 1 (because BYOL-VAR performed poorly in Minigrid, we do not evaluate it with DQN). Results in this section are averaged over 10 independent seeds with 95% standard error in the error bands.

For open-AI gym's (Brockman et al., 2016) classical domains (Sutton and Barto, 2018), we consider: 1) Cartpole, where the a pole is placed upright on the cart and the goal is to balance the pole by applying forces in the left and right direction on the cart, 2) Acrobot, where the goal is to apply torques on an actuated joint to swing the free end of the linear chain above a given height while starting from the initial state of hanging downwards, and 3) Mountain Car, where a car is placed stochastically at the bottom of a sinusoidal valley and the goal is to accelerate the car to reach the goal state on top of the right hill.

For both BYOL-Π and BYOL-AC here, we first tuned and fixed all the DQN specific parameters and then tuned to obtain the the best hyper-parameters for both methods.

Our high level architecture is shown in Figure 4. For these domains, we use a soft-dicretization of the observation space before passing it into our torso. We independently convert each observation dimension to a soft-one-hot encoding using a Gaussian distribution. For example, if we want to encode the value $a = 0.2$ into a soft-one-hot with 11 bins, with lower bound 0 and upper bound 1, we first compute the distance between 0.2 and the 11 bin points (b_i) (0, 0.1, 0.2, . . . , 0.9, 1.0) using $e^{-\frac{1}{2}\left(\frac{a-b_i}{\sigma}\right)^2}$. Then we normalize so that this sums to one.

CartPole Network:

- Torso: MLP[128, 64]
- LSTM: Hidden size 256
- Predictor: MLP[256, 64]
- RL Head
  - Q Head: MLP[128, 2]

And here are our hyperparameters.

- Minibatch batch size: 8
- Minibatch sequence length: 16
- Replay size 1e5
- Adam Optimizer[learning_rate=1e-4]
- Soft Discretization
  - 64 Bins
  - σ 0.1
  - Observation lower bound [-5, -5, -0.5, -5]
  - Observation lower bound [5, 5, 0.5, 5]

MountainCar Network:

- Torso: MLP[64, 64]
- LSTM: Hidden size 256
- Predictor: MLP[256, 64]
- RL Head
  - Q Head: MLP[256, 256, 2]

And here are our hyperparameters.

- Minibatch batch size: 8
- Minibatch sequence length: 16
- Replay size 1e5
- Adam Optimizer[learning_rate=5e-4]
- Soft Discretization
  - 32 Bins
  - σ 0.05
  - Observation lower bound [-1.2, -0.07]
  - Observation lower bound [0.6, 0.07]

Acrobot Network:

- Torso: MLP[128, 64]
- LSTM: Hidden size 256
- Predictor: MLP[256, 64]
- RL Head
  - Q Head: MLP[256, 256, 3]

And here are our hyperparameters.

- Minibatch batch size: 8
- Minibatch sequence length: 16
- Replay size 1e5
- Adam Optimizer[learning_rate=5e-4]
- Soft Discretization
  - 32 Bins
  - σ 0.1
  - Observation lower bound [-1., -1., -1., -1., -12.57, -28.27]
  - Observation lower bound [1., 1., 1., 1., 12.57, 28.27]

D.4 Computing and Libraries

All experiments were conducted using either TPU-V2 or L4 GPU instances. Libraries that enabled this work, include NumPy (Oliphant et al., 2006), SciPy (Virtanen et al., 2020), Matplotlib (Hunter and Dale, 2007), JAX (DeepMind et al., 2020), Gymnax (Lange, 2022), Flashbax (Toledo et al., 2023), and Flax (Heek et al., 2023).

E Measuring robustness of Φ to changes in policy

Having examined how well a representation fits to different objectives under the same policy in Table 2, we now investigate how robust each representation is to off-policy data. To establish a measure of robustness in the learned representation, we propose examining the distance in the representations with respect to the changes in policy π and therefore to changes in the induced dynamics P as follows. Formally, for a given policy π' obtained by perturbing π, we define

$$\Delta(\Phi) := d_{\text{Gr}}(\Phi^*_\pi, \Phi^*_{\pi'}) \,,$$

where $\Phi^*_\pi, \Phi^*_{\pi'}$ are limit points of the dynamics for the ODEs under π, π' respectively, with the same initialisation of Φ_0, and d_{Gr} is the Grassmann distance.

Table 3: **Stability Analysis.** For each method, we report the Δ in the representation upon perturbation in the initial policy, and $P()$ denotes the probability of a method with minimal shift in the representation compared to the other two representations. We report the standard error in the bracket corresponding to 200 independent runs over randomly initialized policy to run the ODE to obtain Φ.

Method	**BYOL-Π** $[\Phi]$		**BYOL-AC** $[\Phi_{\text{ac}}]$		**BYOL-VAR** $[\Phi_{\text{var}}]$	
Initial Policy	$\Delta(\Phi)$	$\mathbf{P}(\Delta(\Phi) \leq (\Delta(\Phi_{\text{ac}}), \Delta(\Phi_{\text{var}})))$	$\Delta(\Phi_{\text{ac}})$	$\mathbf{P}(\Delta(\Phi_{\text{ac}}) \leq (\Delta(\Phi), \Delta(\Phi_{\text{var}}))$	$\Delta(\Phi_{\text{var}})$	$\mathbf{P}(\Phi_{\text{var}} \leq (\Delta(\Phi), \Delta(\Phi_{\text{ac}}))$
ϵ-greedy ($\epsilon = 0.01$)	0.032 (0.004)	0.11	0.023 (0.007)	**0.89**	0.741 (0.03)	0.0
ϵ-greedy ($\epsilon = 0.03$)	0.042 (0.008)	0.095	0.014 (0.001)	**0.905**	0.54 (0.03)	0.0
ϵ-greedy ($\epsilon = 0.1$)	0.037 (0.008)	0.095	0.027 (0.007)	**0.84**	0.243 (0.025)	0.065
ϵ-greedy ($\epsilon = 0.25$)	0.043 (0.008)	0.08	0.035 (0.009)	**0.76**	0.167 (0.02)	0.16

F Additional Intuition on Theory

F.1 Intuition and Implications of Assumptions.

Assumption 1 (Orthogonal Initialization)**.** *Φ is initialized to be orthogonal i.e. $\Phi^T\Phi = I$.*

This assumption on the orthogonal initialization is relatively reasonable and often considered in deep RL methods as well. A random matrix where entries are taken from a unit normal distribution is highly likely to be close to orthogonal. This also suggests that a randomly initialized neural network may have a good chance to approximately satisfy this assumption depending on the input type.

Assumption 2 (Uniform State Distribution)**.** *The state distribution d_X is uniform.*

This uniform state assumption doesn't often hold in practice, and unfortunately is important for our proofs to hold. We have attempted to relax this assumption, but it requires considerable amount of work, and therefore is out of scope of this paper.

Assumption 3 (Symmetric Dynamics)**.** *T^π is symmetric i.e. $T^\pi = (T^\pi)^T$.*

We note that while our theoretical guarantees might require the transition matrices to be symmetric, Tang et al. (2023) have shown it is possible to relax this assumption. To do so, we can consider doubly stochastic matrices, together with a bi-directional algorithm that models the backward transition process in addition to the forward transition dynamics. Leveraging this insight, all our results would still hold for non-symmetric transition dynamics. We here demonstrate the evolution of trace objective with time when the assumption of symmetric MDPs is relaxed. The numerical evidence here shows that the learning dynamics can still capture useful spectral information about the corresponding transition dynamics. Notably, the assumption is stricter for BYOL-VAR which shows the trace objective might not be capturing much useful information when this assumption is relaxed.

Assumption 5 (Symmetric Per-action Dynamics)**.** *T_a is symmetric for all actions i.e. $T_a = (T_a)^T$.*

Analogous to requiring the policy induced transition matrix T^π to be symmetric, we also require the per-action matrices T_a to be symmetric, which is a strong and impractical assumption. If this assumption is violated, the trace objective is no longer a Lyapunov function. We did a simple investigation where we ran the ODE without this assumption (Figure 6) and it seems that even then the ODE does end up increasing the trace objective on average. One future avenue, where this assumption could potentially yield more interesting insights is the hierarchical RL setting. When considering options (multi-step action), this assumption can be relaxed for the primitive actions while only requiring the option-level transition matrix T_o to be symmetric.

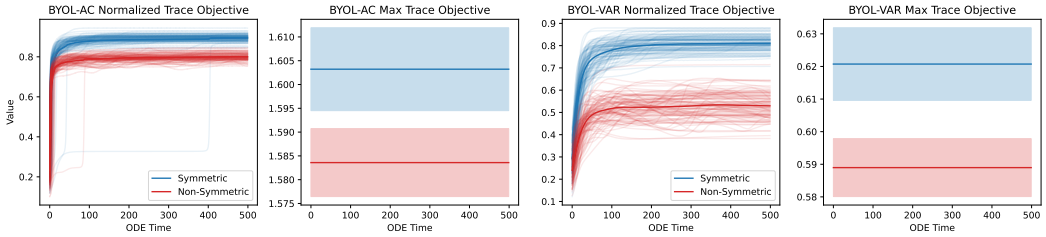

Figure 6: **Ratio between the trace objective and the value of the objective for the top k eigenvectors of T_a and $T_a - T^\pi$** corresponding to BYOL-Π and BYOL-AC respectively, versus the number of ODE training iterations. The light light curve corresponds to one of 100 independent runs over randomly generated MDPs, and the solid curve shows the median over runs.

Assumption 6 (Common Eigenvectors)**.** *For all actions a, we have the eigen decomposition $T_a = QD_aQ^T$, i.e. all T_a share the same eigenvectors.*

This common eigenspace assumption is mainly used to exactly specify what the maximizers of the trace objectives look like. Actually, the negative trace objectives being Lyapunov functions does not require this assumption. Our linear experiments in Table 1 do not satisfy this assumption. Without this assumption though it becomes very difficult to specify what the maximizer looks like, and whether it is a critical point of the ODE. Thus while we make use of this assumption in our theory, it is not strictly necessary and is mainly used to obtain easily interpretable results. Notably, our key results encompassing the two unifying views do not require this assumption, and while the variance equation does require it, the intuition behind the variance equation still holds.

Assumption 4 (Uniform Policy)**.** *The (data-collection) policy π is uniform across all actions.*

We note that the assumption 4 is more strict than necessary. It is enough for the policy π to be state-independent, i.e. the same distribution over actions at every state, and all the results would still hold with just being π-weighted. However we assume this stronger version to keep the theory simple and still retain all the important insights that we will analyze. The choice of a uniform policy is meant to simplify the presentation.

We here provide the objective with a different distribution π over the actions to obtain similar results. For example, with π instead of a uniform distribution, would result in:

$$f_{\text{BYOL-AC}}(\Phi) = \textstyle\sum_a \pi_a \operatorname{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right)$$

and

$$f_{\text{BYOL-VAR}}(\Phi) = \sum_a \pi_a \operatorname{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right) - \operatorname{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right).$$

Evidently, changing π will accordingly change the maximizers of $f_{\text{BYOL-AC}}$ and $f_{\text{BYOL-VAR}}$, as well as the outcome of representation learning.

F.2 Intuition Behind the Loss and Trace Objective.

On convergence: Our main Theorems 1, 2 and 5 focus on analyzing the maximizer to the trace objectives in Lemmas 2, 4 and 6. We show that the maximizer of the trace objective is indeed a critical point of the corresponding ODE. However there can exist other, non-maximizer, critical points of the ODE. This means it is possible that the ODE converges to a sub-optimal critical point. Because we are interested in the optimal point, we don't dive further into the properties of the sub-optimal critical points. However, it can be a useful direction for future work to investigate whether the sub-optimal critical points are stable or unstable, and if there is a way to guarantee convergence to the maximum.

G Proofs

The Assumptions, Lemmas, and Theorems of Appendix A.1 are taken (with some wording changes to be consistent with this work) from the results in Tang et al. (2023).

G.1 Proofs of Sec. 2: BYOL-AC

We first present and prove a few helper lemmas about what are P_a^* and $\dot{\Phi}$. Note as shorthand we let $\mathbb{E}[xx^T] = D_X$, which means $\mathbb{E}_{y\sim p(\cdot|x,a)}[xy^T] = D_X T_a$.

G.1.1 Finding P_a^*

Lemma 7 (Optimal P_a^*)**.** *We have the following.*

$$P_a^* \in \arg\min_{P_a} \mathbb{E}\left[\left|\left|P_a^T \Phi^T x - \text{sg}(\Phi^T y)\right|\right|_2^2\right]$$
$$\implies \left(\Phi^T D_X \Phi\right) P_a^* = \Phi^T D_X T_a \Phi$$

Proof. We first expand and rewrite the objective as a trace objective. Note that we can ignore the stop-gradient because we are only concerned with P_a.

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|P_a^T \Phi^T x - \Phi^T y\right|\right|_2^2\right] \\
&= \mathbb{E}_{x,a,y|(x,a)}\left[x^T \Phi P_a P_a^T \Phi^T x - 2y^T \Phi P_a^T \Phi^T x + y^T \Phi \Phi^T y\right] \\
&= \frac{1}{|A|}\sum_a \mathbb{E}_{x,y|(x,a)}\left[x^T \Phi P_a P_a^T \Phi^T x - 2y^T \Phi P_a^T \Phi^T x + y^T \Phi \Phi^T y\right] \quad \text{since } \pi \text{ is uniform} \\
&= \frac{1}{|A|}\sum_a \mathbb{E}_{x,y|(x,a)}\left[\text{Tr}\left(x^T \Phi P_a P_a^T \Phi^T x - 2y^T \Phi P_a^T \Phi^T x + y^T \Phi \Phi^T y\right)\right] \\
&= \frac{1}{|A|}\sum_a \text{Tr}\left(\mathbb{E}[xx^T]\Phi P_a P_a^T \Phi^T - 2\mathbb{E}[xy^T]\Phi P_a^T \Phi^T + \mathbb{E}[yy^T]\Phi\Phi^T\right)
\end{aligned}$$

The $\mathbb{E}[yy^T]$ term can be considered as just a constant since it does not depend on P_a. Thus we end up with

$$= \frac{1}{|A|}\sum_a \text{Tr}\left(D_X \Phi P_a P_a^T \Phi^T - 2D_X T_a \Phi P_a^T \Phi^T\right) + \text{Constant}$$

Next, we take the derivative w.r.t. P_a.

$$\frac{\partial}{\partial P_a}\left(\frac{1}{|A|}\sum_a \text{Tr}\left(D_X \Phi P_a P_a^T \Phi^T - 2D_X T_a \Phi P_a^T \Phi^T\right)\right) = \frac{1}{|A|}\left(2\Phi^T D_X \Phi P_a - 2\Phi^T D_X T_a \Phi\right)$$

Finally we use the fact that the derivative is zero for P_a^*:

$$0 = \frac{1}{|A|}\left(2\Phi^T D_X \Phi P_a - 2\Phi^T D_X T_a \Phi\right)$$
$$\implies \left(\Phi^T D_X \Phi\right) P_a^* = \Phi^T D_X T_a \Phi$$

□

G.1.2 Computing $\dot{\Phi}$

Lemma 8. $\dot{\Phi}$ *satisfies*

$$\begin{aligned}
\dot{\Phi} &= -\nabla_\Phi \mathbb{E}\left[\left|\left|(P_a)^T \Phi^T x - \text{sg}(\Phi^T y)\right|\right|_2^2\right]\Big|_{P_a = P_a^*} \\
&= -\frac{2}{|A|}\sum_a \left(D_X \Phi P_a^* - D_X T_a \Phi\right)(P_a^*)^T
\end{aligned}$$

Proof. We first expand out the objective into a trace objective. The steps are similar to the steps above in the proof for Lemma 7.

$$\begin{aligned}
&\mathbb{E}\left[\left|\left|(P_a)^T\Phi^T x - \operatorname{sg}(\Phi^T y)\right|\right|_2^2\right] \\
&= \frac{1}{|A|}\sum_a \operatorname{Tr}\left(D_X \Phi P_a P_a^T \Phi^T - 2D_X T_a \operatorname{sg}(\Phi) P_a^T \Phi^T + \mathbb{E}[yy^T]\operatorname{sg}(\Phi\Phi^T)\right) \\
&= \frac{1}{|A|}\sum_a \operatorname{Tr}\left(D_X \Phi P_a P_a^T \Phi^T - 2D_X T_a \operatorname{sg}(\Phi) P_a^T \Phi^T\right) + \text{Constant}
\end{aligned}$$

Next we inspect the derivative:

$$\begin{aligned}
&\frac{\partial}{\partial \Phi}\left(\frac{1}{|A|}\sum_a \operatorname{Tr}\left(D_X \Phi P_a P_a^T \Phi^T - 2D_X T_a \operatorname{sg}(\Phi) P_a^T \Phi^T\right)\right) \\
&= \frac{2}{|A|}\sum_a \left(D_X \Phi P_a - D_X T_a \Phi\right) P_a^T
\end{aligned}$$

Finally, we plug in P_a^* to obtain $\dot{\Phi}$:

$$\dot{\Phi} = -\frac{2}{|A|}\sum_a \left(D_X \Phi P_a^* - D_X T_a \Phi\right)(P_a^*)^T$$

□

G.1.3 Proof of Lemma 3 and Simplified P_a^* and $\dot{\Phi}$

Lemma 3 (Non-collapse BYOL-AC)**.** *Under Assumption 1, we have that $\Phi^T\dot{\Phi} = 0$, which means that $\Phi^T\Phi = I$ is preserved for all Φ throughout the BYOL-AC ODE process.*

Proof.

$$\begin{aligned}
\Phi^T\dot{\Phi} &= \Phi^T\left(-\frac{2}{|A|}\sum_a \left(D_X \Phi P_a^* - D_X T_a \Phi\right)(P_a^*)^T\right) \\
&= \left(-\frac{2}{|A|}\sum_a \left(\underbrace{\Phi^T D_X \Phi P_a^* - \Phi^T D_X T_a \Phi}_{=0 \text{ from definition of } P_a^*}\right)(P_a^*)^T\right) \\
&= 0.
\end{aligned}$$

Therefore, $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^T\Phi_t = 0$, that is, $\Phi_t^T\Phi_t$ is a constant for all t. Since $\Phi_0^T\Phi_0 = I$ by Assumption 1, the result follows. □

As a consequence, and in combination with Assumptions 1 to 5, we can simplify our expressions for P_a^* and $\dot{\Phi}$. Note that with a uniform distribution for D_X, we have $D_X = |\mathcal{X}|^{-1}I$.

$$P_a^* = \Phi^T T_a \Phi \tag{23}$$

$$\dot{\Phi} = \left(I - \Phi\Phi^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a T_a \Phi\Phi^T T_a \Phi\right) \tag{24}$$

G.1.4 Proof of Lemma 4

Lemma 4 (BYOL-AC Trace Objective)**.** *Under Assumptions 1 to 5, a Lyapunov function for the BYOL-AC ODE is the negative of the following trace objective*

$$f_{\text{BYOL-AC}}(\Phi) \coloneqq |A|^{-1}\textstyle\sum_a \operatorname{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi\right) \tag{16}$$

This means the ODE converges to some critical point.

Proof. To check that $-f_{\text{BYOL-AC}}$ is a Lyapunov function for the BYOL-AC ODE, we will verify that its time derivative is strictly negative (it is strictly decreasing) for all non critical points (a critical point being Φ where $\Phi_t = \Phi \Rightarrow \dot{\Phi} = 0$). By chain rule through trace we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(-f_{\text{BYOL-AC}}(\Phi_t)\right) &= -\operatorname{Tr}\left(\frac{\partial}{\partial \Phi}\left(\frac{1}{|A|}\sum_a \Phi^T T_a \Phi \Phi^T T_a \Phi\right)^T \cdot \dot{\Phi}\right) \\
&= -\operatorname{Tr}\left(\left(\frac{1}{|A|}\sum_a \Phi^T T_a \Phi \Phi^T T_a\right) \cdot \dot{\Phi}\right) \\
&= -\operatorname{Tr}\left(\left(\frac{1}{|A|}\sum_a \Phi^T T_a \Phi \Phi^T T_a\right) \cdot \left(I - \Phi\Phi^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a T_a \Phi \Phi^T T_a \Phi\right)\right)
\end{aligned}$$

Since Φ is orthogonal, $(I - \Phi\Phi^T)$ is a projection matrix i.e. $(I - \Phi\Phi^T)(I - \Phi\Phi^T) = (I - \Phi\Phi^T)$. So we can add an extra projection in.

$$\begin{aligned}
&= -\operatorname{Tr}\left(\left(\frac{1}{|A|}\sum_a \Phi^T T_a \Phi \Phi^T T_a\right)\left(I - \Phi\Phi^T\right) \cdot \left(I - \Phi\Phi^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a T_a \Phi \Phi^T T_a \Phi\right)\right) \\
&= -\frac{|\mathcal{X}|}{2}\operatorname{Tr}\left(\dot{\Phi}^T \dot{\Phi}\right),
\end{aligned}$$

which is strictly negative when $\dot{\Phi} \neq 0$. □

G.1.5 Helper Lemmas for the Proof of Theorem 1

Before we prove Theorem 1, we need to present two more helpful Lemmas. The first is the well-known Von Neumann trace inequality. The second concerns maximizing a particular constrained trace expression.

Lemma 9 (Von Neumann Trace Inequality)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ with singular values $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ respectively. Then*

$$\operatorname{Tr}(AB) \leq \sum_{i=1}^{n} \alpha_i \beta_i$$

Lemma 10 (Maximizer of Constrained Trace Expression)**.** *Let $B_a \in \mathbb{R}^{n\times n}$ be symmetric matrices for $a \in \{1, 2, \ldots, |A|\}$ and $\Phi \in \mathbb{R}^{n\times k}$ where $n \geq k$. Assume all B_a share the same eigenvectors, so $B_a = QD_aQ^T$ is an eigendecomposition of B_a and D_a is a diagonal matrix of eigenvalues $\beta_{a,1}, \beta_{a,2}, \ldots, \beta_{a,n}$. Let $Q = [Q_k\ \overline{Q}_k]$ where $Q_k \in \mathbb{R}^{n\times k}$ are the first k columns of Q (and $\overline{Q}_k$ are the rest of the columns). Let the eigenvectors in Q be sorted from largest to smallest according to $\frac{1}{|A|}\sum_a D_a^2$ i.e. Q_k is the top-k eigenvectors for $\frac{1}{|A|}\sum_a B_a^2$. Then under the constraint that Φ is orthogonal ($\Phi^T\Phi = I$), for any orthogonal matrix $C \in \mathbb{R}^{k\times k}$, we have*

$$Q_k C \in \arg\max_{\Phi}\ \frac{1}{|A|}\sum_a \operatorname{Tr}\left(\Phi^T B_a \Phi \Phi^T B_a \Phi\right).$$

In other words, a matrix whose columns span the same subspace as the columns of Q_k is a maximizer of the constrained trace expression.

Proof. We first establish an upper bound on the trace expression.

$$\begin{aligned}
0 &\leq \frac{1}{|A|}\sum_a \operatorname{Tr}\left(\underbrace{\Phi^T B_a (I - \Phi\Phi^T)(I - \Phi\Phi^T) B_a \Phi}_{\text{positive semi-definite}}\right) \\
&= \frac{1}{|A|}\sum_a \operatorname{Tr}\left(\Phi^T B_a \underbrace{(I - \Phi\Phi^T)}_{\text{projection matrix}} B_a \Phi\right)
\end{aligned}$$

where $(I - \Phi^T\Phi)$ is a projection since Φ is orthogonal. Then we have

$$\begin{aligned}
&\frac{1}{|A|}\sum_a \text{Tr}\left(\Phi^T B_a \Phi\Phi^T B_a \Phi\right) \\
&\leq \frac{1}{|A|}\sum_a \text{Tr}\left(\Phi^T B_a \Phi\Phi^T B_a \Phi\right) + \frac{1}{|A|}\sum_a \text{Tr}\left(\Phi^T B_a (I - \Phi\Phi^T) B_a \Phi\right) \\
&= \frac{1}{|A|}\sum_a \text{Tr}\left(\Phi^T B_a^2 \Phi\right) \\
&= \text{Tr}\left(\left(\frac{1}{|A|}\sum_a B_a^2\right)\Phi\Phi^T\right) \quad \text{cyclic property of trace}
\end{aligned}$$

Since Φ is orthogonal, we also have that $\Phi\Phi^T$ is a projection matrix. $\Phi\Phi^T$ is also symmetric, which means that it has an eigendecomposition. Due to being a projection, its eigenvalues must either be 0 or 1. More specifically, since Φ has rank k, it must be that k of its eigenvalues 1 and the rest are 0. Then by the Von Neumann Trace Inequality (Lemma 9), we can bound

$$\text{Tr}\left(\left(\frac{1}{|A|}\sum_a B_a^2\right)\Phi\Phi^T\right) \leq \sum_{i=1}^{k}\left(\frac{1}{|A|}\sum_a \beta_{a,i}^2\right)$$

i.e. it is bounded above by the sum of the top-k eigenvalues of $\left(\frac{1}{|A|}\sum_a B_a^2\right)$, since the eigenvalues of $\Phi\Phi^T$ essentially act as a filter, picking out k eigenvalues of $\left(\frac{1}{|A|}\sum_a B_a^2\right)$. Thus, to summarize what we have so far, we have the following upper bound

$$\frac{1}{|A|}\sum_a \text{Tr}\left(\Phi^T B_a \Phi\Phi^T B_a \Phi\right) \leq \sum_{i=1}^{k}\left(\frac{1}{|A|}\sum_a \beta_{a,i}^2\right) \tag{25}$$

Note that is a global upper bound since it holds for any Φ. Next we will show that $\Phi = Q_k C$ actually attains the upper bound, and is thus a maximizer. We plug this into the trace expression.

$$\begin{aligned}
&\frac{1}{|A|}\sum_a \text{Tr}\left((Q_kC)^T B_a (Q_kC)(Q_kC)^T B_a (Q_kC)\right) \\
&= \frac{1}{|A|}\sum_a \text{Tr}\left(Q_k^T B_a Q_k Q_k^T B_a Q_k\right) \\
&= \frac{1}{|A|}\sum_a \text{Tr}\left(Q_k^T (QD_aQ^T) Q_k Q_k^T (QD_aQ^T) Q_k\right) \qquad \text{substitute eigendecomposition of } A \\
&= \frac{1}{|A|}\sum_a \text{Tr}\left([I_k\ 0] D_a [I_k\ 0]^T [I_k\ 0] D_a [I_k\ 0]^T\right) \qquad \text{since } Q_k^T Q = Q_k^T[Q_k\ \overline{Q}_k] = [I_k\ 0] \\
&= \text{Tr}\left([I_k\ 0]^T [I_k\ 0]\left(\frac{1}{|A|}\sum_a D_a^2\right)\right) \qquad \text{since diagonal matrices commute} \\
&= \sum_{i=1}^{k}\left(\frac{1}{|A|}\sum_a \beta_{a,i}^2\right)
\end{aligned}$$

Thus $\Phi = Q_k C$ results in the trace expression attaining its global upper bound, so it is a maximizer. □

G.1.6 Proof of Theorem 1

Theorem 1 (BYOL-AC ODE)**.** *Under Assumptions 1 to 6, let Φ_{ac}^* be any maximizer of the trace objective $f_{\text{BYOL-AC}}(\Phi)$:*

$$\Phi_{ac}^* \subseteq \arg\max_\Phi \ f_{\text{BYOL-AC}}(\Phi) = \arg\max_\Phi \ |A|^{-1}\textstyle\sum_a \text{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi\right) \tag{3}$$

Then Φ_{ac}^* is a critical point of the ODE. Furthermore, the columns of Φ_{ac}^* span the same subspace as the top-k eigenvectors of $\left(|A|^{-1}\sum_a T_a^2\right)$.

Proof. Let the eigendecomposition of T_a be $T_a = QD_aQ^T$ for all actions a. By Assumption 6, Q are the shared eigenvectors across all T_a. Let the eigenvectors of Q be sorted from largest to smallest according to $\frac{1}{|A|}\sum_a D_a^2$. Let $Q = [Q_k\ \overline{Q}_k]$ so that Q_k are the top-k eigenvectors. Let $C \in \mathbb{R}^{k\times k}$ be an arbitrary orthogonal matrix. From Lemma 10 we have that $\Phi^*_{\text{ac}} = Q_kC$ is a maximizer of the trace objective. Because we are mutiplying Q_k by C on the right, Φ^*_{ac} columns span the same subspace as the columns of Q_k i.e. the same subspace as the top-k eigenvectors of $\left(\frac{1}{|A|}\sum_a T_a^2\right)$, as we wanted to show.

To finish the proof, it remains to show that Φ^*_{ac} is a critical point of the BYOL-AC ODE, that is, $\Phi_t = \Phi^*_{\text{ac}} \Rightarrow \dot{\Phi} = 0$. We plug $\Phi^*_{\text{ac}} = Q_kC$ into $\dot{\Phi}$ (Eq. 24):

$$
\begin{aligned}
&\left(I - (Q_kC)(Q_kC)^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a T_a(Q_kC)(Q_kC)^T T_a(Q_kC)\right)\\
&= \left(I - Q_kQ_k^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a T_aQ_kQ_k^TT_aQ_kC\right)\\
&= \left(I - Q_kQ_k^T\right)\left(\frac{2}{|A||\mathcal{X}|}\sum_a (QD_aQ^T)Q_kQ_k^T(QD_aQ^T)Q_kC\right) \quad \text{substitute } T_a = QD_aQ^T\\
&= \frac{2}{|A||\mathcal{X}|}\sum_a \left(I - Q_kQ_k^T\right)QD_a[I_k\ 0]^T[I_k\ 0]D_a[I_k\ 0]^TC \quad \text{where } Q_k^TQ = Q_k^T[Q_k\ \overline{Q}_k] = [I_k\ 0]\\
&= \frac{2}{|A||\mathcal{X}|}\sum_a \left(I - Q_kQ_k^T\right)Q[I_k\ 0]^T[I_k\ 0]D_a^2[I_k\ 0]^TC \quad \text{since diagonal matrices commute}\\
&= \frac{2}{|A||\mathcal{X}|}\sum_a \left(I - Q_kQ_k^T\right)Q_kD_a^2[I_k\ 0]^TC\\
&= \frac{2}{|A||\mathcal{X}|}\sum_a \left(Q_k - Q_k\right)D_a^2[I_k\ 0]^TC\\
&= 0
\end{aligned}
$$

Thus Φ^*_{ac} is a critical point. □

G.2 Proofs of Sec. 3: BYOL-VAR

We already know what P_a^* is from Lemma 7, and through a similar argument we also know what P^* is. Now we focus on first computing $\dot{\Phi}$ for BYOL-VAR.

G.2.1 Computing $\dot{\Phi}$

Lemma 11. *Under the uniform policy assumption Assumption 4. We have the following.*

$$
\begin{aligned}
\dot{\Phi} &= -\nabla_\Phi \mathbb{E}\left[\|P_a^\top\Phi^\top x - \text{sg}(\Phi^\top y))\|_2^2 - \|P^\top\Phi^\top x - \text{sg}(\Phi^\top y))\|_2^2\right]\big|_{P=P^*, P_a=Pa^*}\\
&= -\frac{2}{|A|}\sum_a \left(D_X\Phi P_a^* - D_XT_a\Phi\right)(P_a^*)^T + 2\left(D_X\Phi P^* - D_XT^\pi\Phi\right)(P^*)^T
\end{aligned}
$$

Proof. We know the following from Lemma 7 (the same steps apply for P^* but we replace T_a with T^π).

$$
\begin{aligned}
\left(\Phi^TD_X\Phi\right)P_a^* &= \Phi^TD_XT_a\Phi\\
\left(\Phi^TD_X\Phi\right)P^* &= \Phi^TD_XT^\pi\Phi
\end{aligned}
$$

Note that because π is uniform, we have $T^\pi = \frac{1}{|A|}\sum_a T_a$. This also implies $P^* = \frac{1}{|A|}\sum_a P_a^*$. Then we just apply the steps in Lemma 8 on the two terms in $\dot{\Phi}$ to get the difference. □

G.2.2 Proof of Lemma 5 and Simplified P^*, P_a^* and $\dot{\Phi}$

Lemma 5 (Non-collapse BYOL-VAR)**.** *Under Assumptions 1 and 4, we have that $\Phi^T\dot{\Phi} = 0$, which means that $\Phi^T\Phi = I$ is preserved for all Φ throughout the BYOL-VAR ODE process.*

Proof.

$$\begin{aligned}\Phi^T\dot{\Phi} &= \Phi^T\left(-\frac{2}{|A|}\sum_a \left(D_X\Phi P_a^* - D_X T_a\Phi\right)(P_a^*)^T + 2\left(D_X\Phi P^* - D_X T^\pi\Phi\right)(P^*)^T\right)\\ &= -\frac{2}{|A|}\sum_a\left(\underbrace{\Phi^T D_X\Phi P_a^* - \Phi^T D_X T_a\Phi}_{=0 \text{ by definition of } P_a^*}\right)(P_a^*)^T + 2\left(\underbrace{\Phi^T D_X\Phi P^* - \Phi^T D_X T^\pi\Phi}_{=0 \text{ by definition of } P^*}\right)(P^*)^T\\ &= 0.\end{aligned}$$

Therefore, $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^T\Phi_t = 0$, that is, $\Phi_t^T\Phi_t$ is a constant for all t. Since $\Phi_0^T\Phi_0 = I$ by Assumption 1, the result follows. □

As a consequence, and in combination with Assumptions 1 to 5, we can simplify our expressions for $P*$, P_a^* and $\dot{\Phi}$. Note that with a uniform distribution for D_X, we have $D_X = |\mathcal{X}|^{-1}I$.

$$P* = \Phi^T T^\pi\Phi \tag{26}$$

$$P_a^* = \Phi^T T_a\Phi \tag{27}$$

$$\dot{\Phi} = 2\left(I - \Phi\Phi^T\right)\left(\frac{1}{|A||\mathcal{X}|}\sum_a T_a\Phi\Phi^T T_a\Phi - T^\pi\Phi\Phi^T T^\pi\Phi\right) \tag{28}$$

G.2.3 Proof of Lemma 6

Lemma 6 (BYOL-VAR Trace Objective)**.** *Under Assumptions 1 to 5, a Lyapunov function for the BYOL-VAR ODE is the negative of the following trace objective*

$$\begin{aligned}f_{\textit{BYOL-VAR}}(\Phi) &\coloneqq f_{\textit{BYOL-AC}}(\Phi) - f_{\textit{BYOL-}\Pi}(\Phi)\\ &= |A|^{-1}\sum_a \mathrm{Tr}\left(\Phi^T T_a\Phi\Phi^T T_a\Phi\right) - \mathrm{Tr}\left(\Phi^T T^\pi\Phi\Phi^T T^\pi\Phi\right)\end{aligned} \tag{19}$$

This means the ODE converges to some critical point.

Proof. To check that $-f_{\text{BYOL-VAR}}$ is a Lyapunov function for the BYOL-VAR ODE, we will verify that its time derivative is strictly negative (it is strictly decreasing) for all non critical points (a critical point being Φ where $\Phi_t = \Phi \Rightarrow \dot{\Phi} = 0$). By chain rule through trace we have

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\left(-f_{\text{BYOL-VAR}}(\Phi_t)\right)\\ &= -\mathrm{Tr}\left(\frac{\partial}{\partial\Phi}\left(\mathrm{Trace}\left(\frac{1}{|A|}\sum_a\left(\Phi^T T_a\Phi\right)^T\Phi^T T_a\Phi - \left(\Phi^T T^\pi\Phi\right)^T\Phi^T T^\pi\Phi\right)\right)^T\cdot\dot{\Phi}\right)\\ &= -\mathrm{Tr}\left(\mathrm{Trace}\left(\frac{1}{|A|}\sum_a\Phi^T T_a\Phi\Phi^T T_a - \Phi^T T^\pi\Phi\Phi^T T^\pi\right)\cdot\dot{\Phi}\right)\end{aligned}$$

Since Φ is orthogonal, $(I - \Phi\Phi^T)$ is a projection matrix i.e. $(I - \Phi\Phi^T)(I - \Phi\Phi^T) = (I - \Phi\Phi^T)$. So we can add an extra projection in front of $\dot{\Phi}$

$$\begin{aligned}&= -\mathrm{Tr}\left(\mathrm{Trace}\left(\frac{1}{|A|}\sum_a\Phi^T T_a\Phi\Phi^T T_a - \Phi^T T^\pi\Phi\Phi^T T^\pi\right)(I - \Phi\Phi^T)\cdot\dot{\Phi}\right)\\ &= -\frac{|\mathcal{X}|}{2}\mathrm{Tr}\left(\dot{\Phi}^T\dot{\Phi}\right)\end{aligned}$$

Therefore this is strictly negative when $\dot{\Phi} \neq 0$. □

G.2.4 Proof of Theorem 2

Theorem 2 (BYOL-VAR ODE). *Under Assumptions 1 to 6, let Φ^*_{VAR} be any maximizer of the trace objective $f_{BYOL\text{-}VAR}(\Phi)$:*

$$\Phi^*_{VAR} \subseteq \arg\max_{\Phi} \; |A|^{-1} \sum_a \text{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right) - \text{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) \tag{5}$$

Then Φ^*_{VAR} is a critical point of the ODE. Furthermore, the columns of Φ^*_{VAR} span the same subspace as the top-k eigenvectors of $\left(|A|^{-1} \sum_a T_a^2 - (T^\pi)^2\right)$.

Proof. The first thing we do is rewrite the trace objective so that we can apply Lemma 10 more directly.

$$\begin{aligned}
&\text{Trace}\left(\frac{1}{|A|}\sum_a \Phi^T T_a \Phi \Phi^T T_a \Phi - \Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) \\
&= \text{Trace}\left(\mathbb{E}\left[\Phi^T T_a \Phi \Phi^T T_a \Phi\right] - \mathbb{E}[\Phi^T T_a \Phi]\mathbb{E}[\Phi^T T_a \Phi]\right)
\end{aligned}$$

We first rewrite the sums and T^π as (pointwise) expectations over the uniform action. Notice how we now have the difference of the expectation of a square with the square of the expectation. This means we can re-express this as a (pointwise) variance term.

$$\begin{aligned}
&\text{Trace}\left(\mathbb{E}\left[\Phi^T T_a \Phi \Phi^T T_a \Phi\right] - \mathbb{E}[\Phi^T T_a \Phi]\mathbb{E}[\Phi^T T_a \Phi]\right) \\
&= \text{Trace}\left(\mathbb{E}\left[\left(\Phi^T T_a \Phi - \mathbb{E}[\Phi^T T_a \Phi]\right)^2\right]\right) \\
&= \text{Trace}\left(\mathbb{E}\left[\Phi^T \left(T_a - T^\pi\right) \Phi \Phi^T \left(T_a - T^\pi\right) \Phi\right]\right) \\
&= \text{Trace}\left(\frac{1}{|A|}\sum_a \Phi^T \left(T_a - T^\pi\right) \Phi \Phi^T \left(T_a - T^\pi\right) \Phi\right)
\end{aligned}$$

Now in this new form, we are ready to apply Lemma 10.

Let the eigendecomposition of T_a be $T_a = Q D_a Q^T$ for all actions a. So Q are the shared eigenvectors across all T_a (Assumption 6). We also know $T^\pi = \frac{1}{|A|}\sum_a T_a$, which means we also have the eigendecomposition $T^\pi = Q\left(\frac{1}{|A|}\sum_a D_a\right) Q^T$. In other words, T^π also shares the same eigenvectors. This means that $\frac{1}{|A|}\sum_a \left(T_a - T^\pi\right)^2$ also has the same eigenvectors.

Let the eigenvectors of Q be sorted from largest to smallest according to $\frac{1}{|A|}\sum_a \left(T_a - T^\pi\right)^2$. Let $Q = [Q_k \; \overline{Q}_k]$ so that Q_k are the top-k eigenvectors. Let $C \in \mathbb{R}^{k \times k}$ be an arbitrary orthogonal matrix. From Lemma 10 we have that $\Phi^*_{\text{var}} = Q_k C$ is a maximizer of the trace objective. Because we are mutiplying Q_k by C on the right, Φ^*_{var} columns span the same subspace as the columns of Q_k i.e. the same subspace as the top-k eigenvectors of $\frac{1}{|A|}\sum_a \left(T_a - T^\pi\right)^2$. We also have (variance relationship)

$$\begin{aligned}
\frac{1}{|A|}\sum_a \left(T_a - T^\pi\right)^2 &= \frac{1}{|A|}\sum_a \left(T_a^2 + (T^\pi)^2 - 2 T_a T^\pi\right) \\
&= \frac{1}{|A|}\sum_a T_a^2 + (T^\pi)^2 - 2 T^\pi T^\pi \\
&= \frac{1}{|A|}\sum_a T_a^2 - (T^\pi)^2
\end{aligned}$$

Thus, equivalently, Φ^*_{var} columns span the same subspace as the columns of the top-k eigenvectors of $\frac{1}{|A|}\sum_a T_a^2 - (T^\pi)^2$, as we wanted to show.

To finish the proof, we show that Φ^*_{var} is a critical point of the BYOL-VAR ODE, that is, $\Phi_t = \Phi^*_{\text{var}} \Rightarrow \dot{\Phi} = 0$. We plug $\Phi^*_{\text{var}} = Q_k C$ into $\dot{\Phi}$ (Eq. 28).

$$2\left(I - \Phi\Phi^T\right)\left(\frac{1}{|A||\mathcal{X}|}\sum_a T_a \Phi \Phi^T T_a \Phi - T^\pi \Phi \Phi^T T^\pi \Phi\right)$$

$$= 2\left(I - Q_k Q_k^T\right)\left(\frac{1}{|A||\mathcal{X}|}\sum_a T_a Q_k Q_k^T T_a - T^\pi Q_k Q_k^T T^\pi\right) Q_k$$

$$= 2\left(I - Q_k Q_k^T\right)\left(\frac{1}{|A||\mathcal{X}|}\sum_a T_a Q_k Q_k^T T_a - \frac{1}{|\mathcal{X}||A|^2}\sum_{a,a'} T_a Q_k Q_k^T T_{a'}\right) Q_k$$

To help simplify further we examine more closely the following term.

$$\begin{aligned}
&\left(I - Q_k Q_k^T\right)\left(T_a Q_k Q_k^T T_{a'}\right) Q_k \\
&= \left(I - Q_k Q_k^T\right)\left((Q D_a Q^T) Q_k Q_k^T (Q D_{a'} Q^T)\right) Q_k \quad \text{substitute } T_a = Q D_a Q^T \\
&= \left(I - Q_k Q_k^T\right)\left(Q D_a [I_k\ 0]^T [I_k\ 0] D_{a'} Q^T\right) Q_k \quad \text{where } Q_k^T Q = Q_k^T [Q_k\ \overline{Q}_k] = [I_k\ 0] \\
&= \left(I - Q_k Q_k^T\right)\left(Q [I_k\ 0]^T [I_k\ 0] D_a D_{a'} Q^T\right) Q_k \quad \text{since diagonal matrices commute} \\
&= \left(I - Q_k Q_k^T\right)\left(Q_k D_a D_{a'} Q^T\right) Q_k \\
&= \left(Q_k - Q_k\right)\left(D_a D_{a'} Q^T\right) Q_k \\
&= 0
\end{aligned}$$

Thus we have $\dot{\Phi} = 0$:

$$2\left(I - Q_k Q_k^T\right)\left(\frac{1}{|A||\mathcal{X}|}\sum_a T_a Q_k Q_k^T T_a - \frac{1}{|\mathcal{X}||A|^2}\sum_{a,a'} T_a Q_k Q_k^T T_{a'}\right) Q_k = 0$$

Thus Φ^*_{var} is a critical point. □

G.3 Proofs of Sec. 4: Unifying Perspectives

G.3.1 Proof of Theorem 3

Theorem 3 (Unifying Model-Based View)**.** *Under Assumptions 1 to 6, the negative trace objectives of BYOL-Π, BYOL-AC, and BYOL-VAR are equivalent (up to a constant* C*) to the following objectives ($\|\cdot\|_F$ is the Frobenius matrix norm):*

$$-f_{\textit{BYOL-}\Pi}(\Phi) = \min_P \ \|T^\pi - \Phi P \Phi^T\|_F + \mathrm{C} \tag{6}$$

$$-f_{\textit{BYOL-AC}}(\Phi) = |A|^{-1}\textstyle\sum_a \min_{P_a} \ \|T_a - \Phi P_a \Phi^T\|_F + \mathrm{C} \tag{7}$$

$$-f_{\textit{BYOL-VAR}}(\Phi) = |A|^{-1}\textstyle\sum_a \min_{P_{\Delta a}} \ \|(T_a - T^\pi) - \Phi P_{\Delta a} \Phi^T\|_F + \mathrm{C} \tag{8}$$

Proof. We start with the proof of Eq. 6. We first expand out the Frobenius norm on the right-hand side. Note that from Assumption 5 we know that T^π and T_a are symmetric, and from Lemma 1 we know Φ is orthogonal.

$$\begin{aligned}
\|T^\pi - \Phi P \Phi^T\|_F &= \mathrm{Tr}\left(\left(T^\pi - \Phi P \Phi^T\right)^T \left(T^\pi - \Phi P \Phi^T\right)\right) \\
&= \mathrm{Tr}\left(T^\pi T^\pi - 2\Phi P^T \Phi^T T^\pi + \Phi P^T P \Phi^T\right)
\end{aligned}$$

To minimize w.r.t. P, we compute the matrix derivative w.r.t. P.

$$\frac{\partial}{\partial P}\mathrm{Tr}\left(T^\pi T^\pi - 2\Phi P^T \Phi^T T^\pi + \Phi P^T P \Phi^T\right) = -2\Phi^T T^\pi \Phi + 2P$$

Then setting this to zero, we solve for the minimizer P^*.

$$\begin{aligned}
0 &= -2\Phi^T T^\pi \Phi + 2P^* \\
\implies P^* &= \Phi^T T^\pi \Phi
\end{aligned}$$

Plugging this back in Eq. 6, we get

$$\begin{aligned}
\mathrm{Tr}\left(T^\pi T^\pi - 2\Phi (P^*)^T \Phi^T T^\pi + \Phi (P^*)^T P^* \Phi^T\right) &= \mathrm{Tr}\left(T^\pi T^\pi\right) - \mathrm{Tr}\left(\Phi^T T^\pi \Phi \Phi^T T^\pi \Phi\right) \\
&= \mathrm{Tr}\left(T^\pi T^\pi\right) - f_{\text{BYOL-}\Pi}(\Phi)
\end{aligned}$$

Thus we have Eq. 6 where the constant term is $\mathrm{Tr}\left(T^\pi T^\pi\right)$.

Next to prove Eq. 7, we follow the same steps, except we substitute T_a for T^π. This results in

$$\begin{aligned}\frac{1}{|A|}\sum_a \min_{P_a} \|T_a - \Phi P_a \Phi^T\|_F &= \frac{1}{|A|}\sum_a \left(\mathrm{Tr}\left(T_a T_a\right) - \mathrm{Tr}\left(\Phi^T T_a \Phi \Phi^T T_a \Phi\right)\right) \\ &= \left(\frac{1}{|A|}\sum_a \mathrm{Tr}\left(T_a T_a\right)\right) - f_{\text{BYOL-AC}}(\Phi)\end{aligned}$$

Finally, we do the same again for Eq. 8 but with $(T_a - T^\pi)$.

$$\begin{aligned}&\frac{1}{|A|}\sum_a \min_{P_{\Delta a}} \|(T_a - T^\pi) - \Phi P_{\Delta a} \Phi^T\|_F \\ &= \frac{1}{|A|}\sum_a \left(\mathrm{Tr}\left((T_a - T^\pi)^2\right) - \mathrm{Tr}\left(\Phi^T (T_a - T^\pi)\Phi\Phi^T (T_a - T^\pi)\Phi\right)\right) \\ &= \left(\frac{1}{|A|}\sum_a \mathrm{Tr}\left((T_a - T^\pi)^2\right)\right) - \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi - \Phi^T T^\pi \Phi\Phi^T T^\pi \Phi\right) \\ &= \left(\frac{1}{|A|}\sum_a \mathrm{Tr}\left((T_a - T^\pi)^2\right)\right) - f_{\text{BYOL-VAR}}(\Phi)\end{aligned}$$

where the second-last step uses the fact that $T^\pi = \frac{1}{|A|}\sum_a T_a$, which holds since π is uniform in accordance with Assumption 4. □

G.3.2 Proof of Theorem 4

Theorem 4 (Unifying Model-Free View)**.** *Under Assumptions 1 to 6, the negative trace objectives of BYOL-Π, BYOL-AC, and BYOL-VAR are equivalent (up to a constant* C*) to the following objectives:*

$$-f_{\textit{BYOL-}\Pi}(\Phi) = |\mathcal{X}|\mathbb{E}\left[\min_{\theta,\omega} \left(\|T^\pi R - \Phi\theta\|^2 + \|T^\pi \Phi\Phi^T R - \Phi\omega\|^2\right)\right] + \mathrm{C} \tag{9}$$

$$-f_{\textit{BYOL-AC}}(\Phi) = |\mathcal{X}|\mathbb{E}\left[|A|^{-1}\textstyle\sum_a \min_{\theta_a,\omega_a} \left(\|T_a R - \Phi\theta_a\|^2 + \|T_a \Phi\Phi^T R - \Phi\omega_a\|^2\right)\right] + \mathrm{C} \tag{10}$$

$$\begin{aligned}-f_{\textit{BYOL-VAR}}(\Phi) = |\mathcal{X}|\mathbb{E}\big[|A|^{-1}\textstyle\sum_a \min_{\theta_a,\omega_a} &\left(\|(T_a R - T^\pi R) - \Phi\theta\|^2\right. \\ &\left.+ \|(T_a \Phi\Phi^T R - T^\pi \Phi\Phi^T R) - \Phi\omega\|^2\right)\big] + \mathrm{C}\end{aligned} \tag{11}$$

Proof. We start with proving the first equation (Eq. 9). First, we solve the inner minimization w.r.t. θ, resulting in:

$$\min_\theta \|T^\pi R - \Phi\theta\|^2$$

Given its form of a standard linear least squares equation ($\|A\theta - B\|^2$), the solution for θ is:

$$\begin{aligned}\theta^* &= \left(\Phi^T\Phi\right)^{-1}\Phi^T T^\pi R \\ &= \Phi^T T^\pi R \qquad \text{since } \Phi \text{ is orthogonal}\end{aligned}$$

Through the same argument except substituting $T^\pi\Phi\Phi^T$ in for T^π, the solution for ω is:

$$\omega^* = \Phi^T T^\pi \Phi\Phi^T R$$

Substituting θ^* and ω^* back in Eq. 9, and recalling that $|\mathcal{X}|\mathbb{E}[RR^T] = I$, we get:

$$\begin{aligned}&|\mathcal{X}|\mathbb{E}_R\left[\|T^\pi R - \Phi\theta^*\|^2 + \|T^\pi\Phi\Phi^T R - \Phi\omega^*\|^2\right] \\ &= |\mathcal{X}|\mathbb{E}_R\left[\|T^\pi R - \Phi\Phi^T T^\pi R\|^2 + \|T^\pi\Phi\Phi^T R - \Phi\Phi^T T^\pi\Phi\Phi^T R\|^2\right] \\ &= |\mathcal{X}|\mathbb{E}_R\left[\|(I - \Phi\Phi^T)T^\pi R\|^2 + \|(I - \Phi\Phi^T)T^\pi\Phi\Phi^T R\|^2\right] \\ &= |\mathcal{X}|\mathbb{E}_R\left[R^T T^\pi (I - \Phi\Phi^T)(I - \Phi\Phi^T)T^\pi R + R^T\Phi\Phi^T T^\pi (I - \Phi\Phi^T)(I - \Phi\Phi^T)T^\pi\Phi\Phi^T R\right]\end{aligned}$$

$$
\begin{aligned}
&= |\mathcal{X}|\mathbb{E}_R\left[R^T T^\pi (I - \Phi\Phi^T)T^\pi R + R^T \Phi\Phi^T T^\pi (I - \Phi\Phi^T)T^\pi \Phi\Phi^T R\right] \\
&= |\mathcal{X}|\mathbb{E}_R\left[\mathrm{Tr}\left(R^T T^\pi (I - \Phi\Phi^T)T^\pi R\right) + \mathrm{Tr}\left(R^T \Phi\Phi^T T^\pi (I - \Phi\Phi^T)T^\pi \Phi\Phi^T R\right)\right] \\
&= \mathrm{Tr}\left(|\mathcal{X}|\mathbb{E}[RR^T]T^\pi (I - \Phi\Phi^T)T^\pi\right) + \mathrm{Tr}\left(|\mathcal{X}|\mathbb{E}[RR^T]\Phi\Phi^T T^\pi (I - \Phi\Phi^T)T^\pi \Phi\Phi^T\right) \\
&= \mathrm{Tr}\left(T^\pi (I - \Phi\Phi^T)T^\pi\right) + \mathrm{Tr}\left(\Phi\Phi^T T^\pi (I - \Phi\Phi^T)T^\pi \Phi\Phi^T\right) \\
&= \mathrm{Tr}\left(T^\pi (I - \Phi\Phi^T)T^\pi\right) + \mathrm{Tr}\left(T^\pi (I - \Phi\Phi^T)T^\pi \Phi\Phi^T\right) \\
&= \mathrm{Tr}\left(T^\pi T^\pi\right) - \mathrm{Tr}\left(\Phi^T T^\pi \Phi\Phi^T T^\pi \Phi\right) \\
&= C - f_{\text{BYOL-}\Pi}(\Phi)
\end{aligned}
$$

Thus we have proved Eq. 9.

Next, for Eq. 10, the steps are very similar. We first solve the inner minimization for θ_a and ω_a, which are the same as for θ and ω except we substitute T_a for T^π.

$$
\begin{aligned}
\theta_a^* &= \Phi^T T_a R \\
\omega_a^* &= \Phi^T T_a \Phi\Phi^T R
\end{aligned}
$$

Then substituting $\theta_a^\star$ and $\omega_a^\star$ back in Eq. 10, we have.

$$
\begin{aligned}
&|\mathcal{X}|\mathbb{E}_R\left[\frac{1}{|A|}\sum_a \left(\|T_a R - \Phi\theta_a^*\|^2 + \|T_a\Phi\Phi^T R - \Phi\omega_a^*\|^2\right)\right] \\
&= \frac{1}{|A|}\sum_a |\mathcal{X}|\mathbb{E}_R\left[\left(\|T_a R - \Phi\theta_a^*\|^2 + \|T_a\Phi\Phi^T R - \Phi\omega_a^*\|^2\right)\right] \\
&= \frac{1}{|A|}\sum_a \left[\mathrm{Tr}\left(T_a T_a\right) - \mathrm{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi\right)\right]
\end{aligned}
$$

We get the last line above by following the same steps we just did before when substituting in θ^* and ω^*. Thus

$$
\begin{aligned}
&= \frac{1}{|A|}\sum_a \mathrm{Tr}\left(T_a T_a\right) - \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T T_a \Phi\Phi^T T_a \Phi\right) \\
&= C - f_{\text{BYOL-AC}}(\Phi)
\end{aligned}
$$

This completes the proof of Eq. 10.

Finally we prove Eq. 11. We follow the same steps as for the proof for Eq. 10, except we use $(T_a - T^\pi)$ in the place of T_a. This means we have

$$
\begin{aligned}
&|\mathcal{X}|\mathbb{E}_R\left[\frac{1}{|A|}\sum_a \min_{\theta_a,\omega_a}\left(\|(T_a R - T^\pi R) - \Phi\theta\|^2 + \|(T_a\Phi\Phi^T R - T^\pi\Phi\Phi^T R) - \Phi\omega\|^2\right)\right] \\
&= \frac{1}{|A|}\sum_a \mathrm{Tr}\left((T_a - T^\pi)(T_a - T^\pi)\right) - \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T(T_a - T^\pi)\Phi\Phi^T(T_a - T^\pi)\Phi\right) \\
&= C - \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T(T_a - T^\pi)\Phi\Phi^T(T_a - T^\pi)\Phi\right)
\end{aligned}
$$

To further simplify this, we note that because we assume a uniform policy (Assumption 4), we have $T^\pi = \frac{1}{|A|}\sum_a T_a$. So expanding it out

$$
\begin{aligned}
&\frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T(T_a - T^\pi)\Phi\Phi^T(T_a - T^\pi)\Phi\right) \\
&= \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T T_a\Phi\Phi^T T_a\Phi - \Phi^T T_a\Phi\Phi^T T^\pi\Phi - \Phi^T T^\pi\Phi\Phi^T T_a\Phi + \Phi^T T^\pi\Phi\Phi^T T^\pi\Phi\right) \\
&= \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T T_a\Phi\Phi^T T_a\Phi\right) - 2\,\mathrm{Tr}\left(\Phi^T T^\pi\Phi\Phi^T T^\pi\Phi\right) + \mathrm{Tr}\left(\Phi^T T^\pi\Phi\Phi^T T^\pi\Phi\right) \\
&= \frac{1}{|A|}\sum_a \mathrm{Tr}\left(\Phi^T T_a\Phi\Phi^T T_a\Phi\right) - \mathrm{Tr}\left(\Phi^T T^\pi\Phi\Phi^T T^\pi\Phi\right) \\
&= f_{\text{BYOL-VAR}}(\Phi)
\end{aligned}
$$

This completes the proof of Eq. 11. □