When does Self-Prediction help? Understanding Auxiliary Tasks in Reinforcement Learning

Claas Voelcker
University of Toronto
Vector Institute, Toronto
cvoelcker@cs.toronto.edu

Tyler Kastner
University of Toronto
Vector Institute, Toronto
tkastner@cs.toronto.edu

Igor Gilitschenski
University of Toronto
Vector Institute, Toronto
gilitschenski@cs.toronto.edu

Amir-massoud Farahmand
University of Toronto
farahmand@cs.toronto.edu

Abstract

We investigate the impact of auxiliary learning tasks such as observation reconstruction and latent self-prediction on the representation learning problem in reinforcement learning. We also study how they interact with distractions and observation functions in the MDP. We provide a theoretical analysis of the learning dynamics of observation reconstruction, latent self-prediction, and TD learning in the presence of distractions and observation functions under linear model assumptions. With this formalization, we are able to explain why latent-self prediction is a helpful *auxiliary task*, while observation reconstruction can provide more useful features when used in isolation. Our empirical analysis shows that the insights obtained from our learning dynamics framework predicts the behavior of these loss functions beyond the linear model assumption in non-linear neural networks. This reinforces the usefulness of the linear model framework not only for theoretical analysis, but also practical benefit for applied problems.

1 Introduction

Since the emergence of deep learning, techniques for deep supervised learning have been successfully incorporated into reinforcement learning (RL) agents (Mnih et al., 2013; Lillicrap et al., 2016). However, the RL setting contains additional complications such as non-stationary optimization target and the reliance on bootstrapping. These hurdles generally add instability to the RL training process, and recent work has identified the *failure to learn good features* as a central problem in deep RL (Lyle et al., 2022; Nikishin et al., 2022b; Kumar et al., 2021).

To mitigate this failure, one common approach is to add auxiliary tasks to the learning objective (Jaderberg et al., 2017). Popular examples include predicting next state observations (Jaderberg et al., 2016) and predicting functions of the next state (Schwarzer et al., 2021; Ni et al., 2024). To understand the performance of these approaches, recent literature (Tang et al., 2022; Le Lan et al., 2023) considers the *learning dynamics* of auxiliary task learning in simple linear surrogate models (Saxe et al., 2014). One hypothesis in the literature is that observation reconstruction should provide better features than latent self-prediction (Behzadian et al., 2019; Tang et al., 2022). However, this causes a theory-practice gap as empirical work has found that latent self-prediction outperforms observation reconstruction across many benchmarks (Schwarzer et al., 2021; Ni et al., 2024).

To address this gap, we pose two questions: *(a) How do auxiliary losses behave when combined with a TD loss?* Tang et al. (2022); Le Lan et al. (2022); Tang & Munos (2023) have studied the learning dynamics of auxiliary tasks alone, evaluating their performance without addressing

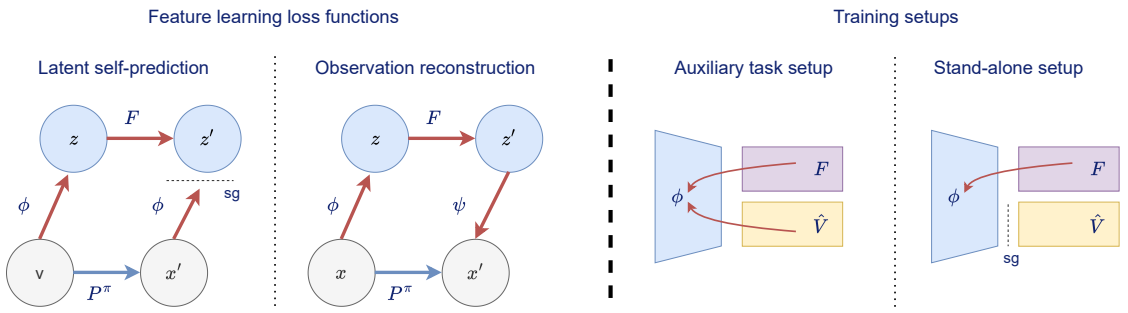

Figure 1: Diagram of the considered loss functions and the different use cases. In latent self-prediction, the aim is to predict next state *features* given by a embedding function $\phi(x')$ using the states features $\phi(x)$ and a latent prediction model F. In observation reconstruction, the aim is to match next state *ground truth observations* x' via the use of a decoder function $\psi(F(\phi(x)))$. In the *auxiliary task setup*, both the gradients from the feature learning loss and value function learning are propagated to the encoder, while in the *stand-alone scenario*, only the gradients from the feature learning loss are used to update Φ.

the interaction between the auxiliary task and the main goal, to learn a (correct) value function. *(b) How can we describe the behavior of auxiliary losses in the presence of distractions (states and transition dynamics irrelevant for the reward) and observation functions (different ways to measure the underlying state)?* MDP structures like distractions have been hypothesized to lead to differing performance between different auxiliary tasks (Ni et al., 2024), but to our knowledge no theoretical study has been established.

In Section 3, we present a formalization of distractions and observation functions. We use the framework of *factored MDPs* (Boutilier et al., 2000) with Kronecker products (Mahadevan, 2009) to represent a common class of distractions. To model observation functions, we use *linear reparametrization* as a tractable way to go beyond one-hot representations.

In Section 4 and Section 5 we analyze the features learned with observation prediction and latent self-prediction alone and in combination with TD learning. We also show how these stationary features change with the introduction of distractions and observation functions. From this analysis we find that latent self-prediction is a strong *auxiliary task*, while observation prediction is a strong feature learning method when used *alone*. The differences are highlighted in Figure 1. This bridges one of the biggest gaps between previous analysis of learning dynamics and empirical results.

In Section 6 we test the predictions derived from our theoretical framework by evaluating feature learning losses in the MinAtar suite (Young & Tian, 2019). [1] We design ablations that mirror both our formalization and previous approaches to test distraction robustness in empirical environments. The theory partially predicts the performance differences in the test suite, validating that the insights we obtain from the simple linear surrogate models used for analysis are useful for practitioners. However, we also find surprising deviations from our predictions on some environments, suggesting that there is need for additional research to fully bridge the theory-practice gap.

2 Background

We briefly introduce the standard formalism of reinforcement learning and linear value function approximations to clarify the notation used. Following this, we briefly introduce the training dynamics framework of Tang et al. (2022); Le Lan et al. (2023). We frame all losses analyzed in this work with the same models to highlight similarities and differences.

[1] All code for our experiments is available at https://github.com/adaptive-agents-lab/understanding_auxiliary_tasks.

Reinforcement Learning. We consider a discounted Markov decision process (MDP) (Puterman, 1994; Sutton & Barto, 2018) $(\mathcal{X}, \mathcal{A}, \mathcal{P}, r, \gamma)$, with state space $\mathcal{X}$, action space $\mathcal{A}$, transition kernel $\mathcal{P} : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, reward function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1)$. Given a policy $\pi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the value function is defined as the expected return conditioned on a state x

$$V^\pi(x) = \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t r_t | x_0 = x \right],$$

where $r_t = r(x_t, a_t)$ is the reward at time t. The goal of an agent is to maximize its value at each state, using the (approximate) value function of its policy V^π.

In finite state spaces ($|\mathcal{X}| = n$ for some integer n), the policy induced transition kernel $\mathcal{P}^\pi(x'|x) = \int \mathcal{P}(x'|x, a) \mathrm{d}\pi(a|x)$ can be represented as a stochastic matrix P^π, and the reward function as a vector r^π. This permits the compact notation $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$. We review several additional properties of stochastic matrices in Subsection C.1.

Value function approximation. A finite state value function can always be expressed as a table or vector, but it is often infeasible to do so, due to limited storage capacity or resources once the state space becomes sufficiently large. Therefore, function approximations need to be introduced, which commonly take the form of a feature function $\phi(x)$ and a weight vector $\hat{V}$, with $V^\pi(x) \approx \phi(x)^\top \hat{V}$. These features can be pre-specified or learned, i.e. using neural networks. Since we are interested in finite state spaces and linear models, we will assume that $\phi(x) : \mathbb{R}^n \to \mathbb{R}^k$ with $k < n$ represented by a matrix $\Phi \in \mathbb{R}^{n \times k}$ and $\hat{V} \in \mathbb{R}^k$.

2.1 Two-layer linear networks as analytical models for training dynamics

Rigorously analyzing the effect of different loss functions on neural networks is challenging due to non-linearities in the networks, shifting data distributions, and policy updates. Therefore, we have to resort to studying simplified models to obtain quantitative and qualitative results, and only consider the fixed policy case in our analysis[2]. Studying feature learning dynamics using linear networks was popularized by Saxe et al. (2014) and has proven to be a valuable tool to analyze diverse objectives such as TD learning (Tang & Munos, 2023; Le Lan et al., 2023), latent self-prediction (Tian et al., 2021; Tang et al., 2022), and linear autoencoders (Pretorius et al., 2018; Bao et al., 2020).

We rewrite the feature learning algorithms using two to three matrices in lieu of more complex functions. Furthermore, we use several assumptions throughout the paper that are listed here for clarity.

Assumption 1. *Let $\Phi \in \mathbb{R}^{n \times k}$ be an* encoder *mapping to a k dimensional embedding space, $F \in \mathbb{R}^{k \times k}$ a* latent model *mapping to the next state's latent embedding, $\hat{V} \in \mathbb{R}^k$ and $\hat{r} \in \mathbb{R}^k$* value and reward weights*, and $\Psi \in \mathbb{R}^{k \times n}$ a* decoder*. Let the sampling distribution of state samples $\mathcal{D}$ be uniform and fixed throughout learning.*

Using this notation, we study four important loss functions for RL: observation reconstruction, where the aim is to fit the next state observation $x^\top \Phi F \Psi \approx x'$, latent reconstruction $x^\top \Phi F \approx x' \Phi$, where the aim is to predict learned features of the next state, and TD learning $x^\top \Phi \hat{V} \approx x^\top (r^\pi + \gamma x'^\top \phi \hat{V})$. To clarify the differences, we show a diagram explaining the losses and training setups in Figure 1. Following common notation, $[\cdot]_{\text{sg}}$ signifies a *stop-gradient* operation; no gradient is taken with regard to terms in the parenthesis.

[2]We discuss these and other assumptions and their implications in detail in Appendix G.

Formally, these are written as

$$\begin{aligned}
\text{Reconstruction:} \qquad & L_{\text{rec}}(\Phi, F, \Psi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \Phi F \Psi - x^\top P^\pi\right\|_2^2\right], \\
\text{Latent self-prediction:} \qquad & L_{\text{lat}}(\Phi, F) = \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \Phi F - \left[x^\top P^\pi \Phi\right]_{\text{sg}}\right\|_2^2\right], \\
\text{TD Learning:} \qquad & L_{\text{td}}(\Phi, \hat{V}) = \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \Phi \hat{V} - \left[x^\top \left(r^\pi + \gamma P^\pi \Phi \hat{V}\right)\right]_{\text{sg}}\right\|_2^2\right].
\end{aligned}$$

Following the nomenclature of Farahmand et al. (2017), we call the first two losses as *decision-agnostic*, and TD-learning as *decision-aware* as it depends on information specific to decision problem. To analyze the discrete-time learning dynamics, we study the analogous continuous-time gradient flow, which allows us to use the toolkit of dynamical systems theory. Writing α_Φ and α_F for the representation and model learning rates, i.e., the latent self-prediction dynamics are

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t &= -\alpha_\Phi \nabla_{\Phi_t} L_{\text{lat}}(\Phi_t, F_t) = -2\alpha_\Phi(\Phi_t F_t - P^\pi \Phi_t)F_t^\top, \\
\frac{\mathrm{d}}{\mathrm{d}t}F_t &= -\alpha_F \nabla_{F_t} L_{\text{lat}}(\Phi_t, F_t) = -2\alpha_F \Phi_t^\top(\Phi_t F_t - P^\pi \Phi_t).
\end{aligned}$$

We primarily consider the *two-timescale* regime, under the assumption that $\alpha_F \to \infty$ (Tang et al., 2022). Intuitively, this describes a learning setup in which the latent model is learned "much faster" than the latent mapping. This results in the following dynamics for self-predictive learning:

$$F_t^* = \left(\Phi_t^\top \Phi_t\right)^{-1} \Phi_t^\top P^\pi \Phi_t, \quad \frac{d}{dt}\Phi_t = \left(I - \Phi_t \left(\Phi_t^\top \Phi_t\right)^{-1} \Phi_t^\top\right) P^\pi \Phi_t {F_t^*}^\top.$$

3 Formalizing the impact of distractions and observation functions

To bridge the theory-practice gap in feature learning, we formalize two structures found in MDPs found in Deep RL benchmarks that have, to the best of our knowledge, not appeared in work analyzing feature learning: observation functions and distractions. To allow a close comparison with previous work, our changes to the formalism are minimal on purpose, while still highlighting the important role these changes play in different loss functions.

3.1 Observation functions

Previous literature (Tang et al., 2022; Tang & Munos, 2023; Le Lan et al., 2022) has eschewed the underlying observation of states in their analysis of representation dynamics. The correctness of the dynamical system in Assumption 2.1 hinges on the fact that $\mathbb{E}[xx^\top] = I$, which implies uncorrelated state representations for each state x and a uniform distribution over states. The simplest form of such a representation would be a *one-hot* vector, a representation for the i-th state in which all entries are 0, except the i-th, which is 1.

This leads to an assumption that the features for the underlying states can be learnt independently. With a one-hot representations, the features of x are simply $x^\top \Phi = \Phi[i]$, the i-th row of Φ. A more realistic setting, which we focus on, is considering observation functions acting on the underlying states. This allows for representing systems where some states have correlated observations, which may be helpful *or* harmful for the RL problem. We provide a motivating example and a more extensive discussion regarding the effect observation functions on the learning process in Subsection A.1.

We briefly state this for reference later.

Assumption 2. *Let an* observation function *for a finite state space MDP be an invertible matrix* $\mathcal{O} \in \mathbb{R}^{n\times n}$.

Formalization: To introduce an observation function while remaining in the regime of analyzing linear networks and finite state problems, each one of n states is mapped to a unique $n-$dimensional observation vector by an invertible *observation matrix* $\mathcal{O} \in \mathbb{R}^{n \times n}$.[3] Invertibility is assumed to ensure the Markov property with linear function approximation.

This change from one-hot vectors to arbitrary vectors allows us to account for similarity. For example, if two states have almost identical observation vectors, they will be mapped to similar points in the latent representation space unless the features directly counteract this. We study the impact of changing the observations with a linear reparameterization in Subsection 4.2 by replacing x with $\bar{x} = \mathcal{O}^T x$.

3.2 Distracting state dynamics

In addition to observation functions, another common problem that many reinforcement learning algorithms face are distractions. While distractions have been a focus of empirical work studying the relative efficacy of different auxiliary tasks (Ni et al., 2024), a simple formalism whose effect on learning dynamics can be analyzed has not been presented. We propose to model an MDP with distractions using factored MDPs (Boutilier et al., 2000).

Definition 1 (A factored MDP model of a distraction)**.** *Let $M = (\mathcal{M}, P_M, R_M)$, $N = (\mathcal{N}, P_N, R_N)$ be a pair of Markov decision processes. The product process $M \otimes N$ is a MDP with state space $\mathcal{M} \times \mathcal{N}$, transition kernel $P_M \otimes P_N$ (where $\otimes$ signifies the Kronecker product), and reward function $R_M \otimes \mathbf{1} + \mathbf{1} \otimes R_N^\top$.*

If $R_N = 0$, we refer to N as a distracting process, as it does not contribute to the reward.

This process models a common occurrence: two non-interacting processes unfold simultaneously, with the states being a combination of the two. Such a process can model a well-studied form of distraction, the background distractions in Stone et al. (2021) or the random observation dimension in Nikishin et al. (2022a); Voelcker et al. (2022). In this case, the foreground process M is assumed to carry the reward information, while the reward vector of the background process N is 0. We review important properties of the Kronecker product in Subsection C.1.

Note that our formalizations of observation functions and distracting processes is distinct from the assumptions in *linear MDPs* (Jin et al., 2020). Concretely we do not assume that the processes are low-rank compressible, just that they are factorizable.

4 Reconstruction and self-prediction losses

In this section and the next, we present an analysis of the stability conditions of reconstruction and self-prediction losses with linear networks. Using this analysis, we obtain several *insights*, qualitative predictions about how we expect the studied losses to behave in more complicated scenarios. These *insights* present the basis for our empirical comparison in Section 6.

4.1 Case 1: Orthogonal state representations

Tang et al. (2022) show that for symmetric MDPs, latent self-prediction converges to subspaces spanned by eigenvectors of P^π. We extend this result in the following sense: if P^π has positive real eigenvalues, invariant sub-spaces which are not spanned by the top-k eigenvectors are unstable for gradient descent.[4] It is interesting to note that the resulting features are identical to those obtained using the multi-reward approach described by Le Lan et al. (2023) (albeit under slightly different

[3] We could also consider projection into higher dimensional spaces $\mathcal{O} \in \mathbb{R}^{n \times d}$ with $d > n$, without violating the Markov assumption, but this leads to additional complications (working with pseudo-inverses instead of inverses) which do not contribute meaningfully to the insights in this work.

[4] The assumption of real *positive* eigenvalues is both more and less restrictive than the symmetry assumption made by Tang et al. (2022). An extension of our result to negative eigenvalues is presented in Appendix D, together with an extended comparison to the results obtained by Tang et al. (2022) and Le Lan et al. (2023).

technical conditions), which highlights the close connection between the self-predictive approach and bootstrapped generalized value function learning. This furthermore suggests that using random rewards as auxiliary objectives (Farebrother et al., 2023) could result in very similar features as using self-prediction, which presentes an interesting avnue for further empirical study.

Assumption 3. *Assume a two-timescale scenario and F_0 being initialized with full rank, and hence the non-collapse property ($\Phi_t^\top \Phi_t = \Phi_0^\top \Phi_0$) (Tang et al., 2022) holds.*

When referring to eigenvectors and singular vectors, we mean the *right* vectors of the corresponding matrices unless stated otherwise. We now present our first theoretical result.

Proposition 1 (Stationary points of latent self-prediction)**.** *Assume Assumption 1 and Assumption 2 hold. Furthermore, suppose P^π is real diagonalizable. If the columns of Φ_t span an invariant subspace of P^π, Φ_t is a stationary point of the dynamical system. Furthermore, if P^π is real-diagonalizable with positive eigenvalues, all invariant subspaces not spanned by the top-k eigenvectors sorted by eigenvalue are asymptotically unstable for gradient descent.*

This implies that even without the assumptions of symmetry of P^π required by Tang et al. (2022), the dynamics of latent self-prediction will tend to converge to invariant subspaces spanned by eigenvectors with large eigenvalues as other invariant subspaces are unstable. This is important as we expect these to be more important for representing potential reward functions in the environment (Le Lan et al., 2023).

We can contrast this with the features learned by a reconstruction loss. We write $\text{span}(A)$ for both the span of the column vectors of A or for the span of a set of vectors A, depending on context.

Proposition 2 (Stationary points of reconstruction)**.** *Assume Assumption 1 and Assumption 2 hold. Write $(u_1, \ldots, u_n)$, $(v_1, \ldots, v_n)$ for the left and right singular vectors of P^π sorted in descending order by singular value. Any stationary point (Φ^*, F^*, Ψ^*) of L_{rec} under the two timescale scenario satisfies* $\text{span}(\Phi^*) = \text{span}(\{u_1, \ldots, u_k\})$, $\text{span}({\Psi^*}^\top) = \text{span}(\{v_1, \ldots, v_k\})$.

Features of this form have been studied extensively and the convergence properties of linear auto-encoders are well understood (Baldi & Hornik, 1989; Pretorius et al., 2018; Bao et al., 2020).

Proposition 1 and Proposition 2 together show that there is a subtle but important difference between latent self-prediction and observation reconstruction: the features will converge to eigenspaces in the former case, and to singular space in the latter case. Note that if P^π is a symmetric matrix, then the singular spaces and the eigen-spaces coincide and latent self-prediction and reconstruction converge to the same features (Tang et al., 2022).

Behzadian et al. (2019) show that top k singular vectors are optimal low-rank linear features when making no assumptions on the reward, meaning observation reconstruction should lead to the best features when considering every possible bounded (or unknown) reward. Behzadian et al. (2019) and Le Lan et al. (2023) both highlight that if eigenvectors and singular vectors differ, singular vectors often lead to better performing features.

Insight 1 (Optimality of observation prediction)**.** *The features learned by observation prediction are in general superior to those of latent self-prediction, when using solely one of these as the loss function.*

4.2 Case 2: Observation function dependence

Recall that the gradient dynamics presented in (2.1) and analyzed in Proposition 1 and Proposition 2 rely on the assumption that $\mathbb{E}[xx^\top] = I$ (see Assumption 1). We now introduce the observation matrix $\mathcal{O}$, which leads to correlations between different features. To do this, we simply replace every occurence of $x^\top$ in the losses presented in Subsection 2.1 with $x^\top \mathcal{O}$. We assume that x is a one-hot vector as discussed before and the coverage is still uniform (Assumption 1 still holds), so all correlation between states arise as $\mathbb{E}[\mathcal{O}^\top xx^\top \mathcal{O}] = \mathcal{O}^\top \mathcal{O}$.

It is important to note that for BYOL and TD this rewriting leads to a linear basis change of Φ compared to the original loss, as each occurrence of $\mathcal{O}$ is multiplied by Φ. The only loss for which this is not the case is the reconstruction approach.

Proposition 3. *Assume Assumption 1, Assumption 2, and Assumption 3 hold. Let* $\{\Phi^*_{\text{lat/td}}\}$ *be the set of critical points of* L_{lat} *or* L_{td} *respectively. Then* $\mathcal{O}^{-1}\Phi^*_{\text{lat/td}}$ *are stationary points for the reparameterized losses* $L^{\mathcal{O}}_{\text{lat}}$ *and* $L^{\mathcal{O}}_{\text{td}}$.[5] *Furthermore, if* $\Phi^*_{\text{lat/td}}$ *is an asymptotically stable point of* $L_{\text{lat/td}}$ *that has a Jacobian with all negative eigenvalues,* $\mathcal{O}^{-1}\Phi^*_{\text{lat/td}}$ *is an asymptotically stable point of* $L^{\mathcal{O}}_{\text{lat/td}}$.

Note that while the stationary points and asymptotic stability conditions of the gradient flow might be unaffected by the introduction of observation distortions, the same might not be true for the dynamics of descent with finite step sizes. The numerical conditioning of the involved matrices change depending on $\mathcal{O}$ and so the impact of discretization due to finite step sizes changes the resulting dynamical system.

Proposition 4. *Assume Assumption 1, Assumption 2, and Assumption 3 hold. Let* $(u_1, \ldots, u_n)$, $(v_1, \ldots, v_n)$ *be the left and right singular vectors of* $\mathcal{O}^{-1}P^\pi\mathcal{O}$. *Any stationary point* (Φ^*, F^*, Ψ^*) *of* $L^{\mathcal{O}}_{rec}$ *satisfies* $\text{span}\,(\Phi^*) = \text{span}\,(\{u_1, \ldots, u_k\})$, $\text{span}\,({\Psi^*}^\top) = \text{span}\,(\{v_1, \ldots, v_k\})$.

The singular value decomposition of $\mathcal{O}^{-1}P^\pi\mathcal{O}$ will in general not have a clearly interpretable relationship to that of P^π and r, so the optimality result obtained by Behzadian et al. (2019) do not hold in this case. However this does not mean that different observation functions will always harm the ability of the reconstruction loss to obtain good features. Consider for example an observation transformation that maps states directly to value and reward function. This would clearly be an example of a helpful observation transformation. However, in general we conjecture that arbitrarily changing the observation function will harm the reconstruction loss approach. A more detailed analysis involving the reward function of the problem being solved and its connections to the observation model are an exciting avenue for future work.

Insight 2 (Observation dependence of autoencoder models)**.** *Due to the invariance properties of latent self-prediction, we expect the performance of latent self-prediction to suffer less than the performance of observation reconstruction when perturbing the observation space arbitrarily.*

5 Understanding the effects on value function learning

The optimality of a representation for value estimation depends non-trivially on the structure of the reward structure of the MDP. Previous works (Behzadian et al., 2019; Bellemare et al., 2019; Le Lan et al., 2022) attempted to reason about the optimality of representations without relying on the reward structure, by arguing that certain subspaces (such as the span of top-k eigenvectors or top-k singular vectors) are optimal given reward agnosticism.

Now we take a differing approach by taking the value function structure into account. Furthermore, we argue that the top-k eigenspaces (resp. singular spaces) are not always optimal. Indeed, we demonstrate in Appendix A.2 that with distractions, these subspaces can be particularly poor.

We begin by formalizing the reward function structure we will analyze. Let us write $w_1, \ldots, w_n$ for the eigenvectors of P^π. We will assume that r^π has a low-dimensional structure in the following sense:

Assumption 4. $\exists i_1, \ldots, i_m \in \{1, \ldots, |\mathcal{X}|\}$ *such that* $r^\pi \in \text{span}(w_{i_1}, \ldots, w_{i_m})$, *and* $m \leq k$. *Let furthermore* $\{w_{i1}, \ldots, w_{im}\}$ *be a minimal basis in the sense that* $w_{in}{w_{in}}^\top r^\pi \neq 0$ *for all* n.

We now write the summed losses

$$L_{\text{rec+td}}(\Phi, F, \psi, \bar{r}) = L_{\text{rec}}(\Phi, F, \psi) + L_{\text{td}}(\Phi, \bar{r})$$

[5] Due to space constraints, we present the full equations in the proof.

and

$$L_{\text{lat+td}}(\Phi, F, \bar{r}) = L_{\text{lat}}(\Phi, F) + L_{\text{td}}(\Phi, \bar{r}).$$

Proposition 5. *Suppose that P^π is real diagonalizable, and that Assumption 1, Assumption 3, and Assumption 4 hold. There exists a non-trivial critical point Φ^* of the two-timescale TD loss L_{td} such that $\text{span}(r^\pi) \subseteq \text{span}(\Phi^*)$. Furthermore, Φ^* is a critical point of the two-timescale joint loss L_{td+lat}. Therefore combining L_{TD} and L_{Lat} does not exclude the existence of a stationary point with 0 value function approximation error.*

We leave the extension of the stability result for TD learning to the joint loss case open for future work. Note that without the addition of TD learning, the latent loss would stabilize the top-k eigenspace representation, but we hypothesize that this behavior changes when combining the losses.

Proposition 6. *Let Assumption 1, Assumption 2, and Assumption 4 hold. If the reward spanning eigenvectors do not lie within the span of the top-k singular vectors, $\text{span}(w_{i_1}, \ldots, w_{i_m}) \not\subseteq \text{span}(u_1, \ldots, u_k)$, the critical points of the two-timescale joint loss L_{td+rec} are guaranteed to not be minimizers of the value function approximation error.*

Contrasting these propositions suggests that when combined with TD learning losses, latent self-prediction can be a more helpful auxiliary task. Indeed, when combining it with TD learning we can still guarantee that there exists an optimal combined solution. This does not hold for the reconstruction loss, where we can construct cases in which the joint loss leads to worse TD error.

Insight 3 (Latent self-prediction as an auxiliary task)**.** *For good performance across a wide variety of tasks, latent self-prediction needs to be combined with TD learning as an auxiliary task. It is a preferable auxiliary task to observation prediction in most scenarios, but especially in scenarios with distracting processes.*

6 Empirical study of theoretical results in deep learning based settings

We aim to empirically verify the statements marked as *"Insights"* throughout the paper: superiority of observation prediction as a standalone feature learning loss (Insight 1), impact of the observation function on the different loss functions (Insight 2), and the relative strength of latent-self prediction as an auxiliary loss compared to reconstruction (Insight 3). As our theory addresses the simplified setting of policy evaluation with linear models, we seek to test if the insights transfer to the more common setting of control with neural networks. Across all experiments, we report mean performance over 30 seeds and shaded 95 bootstrapped confidence interval.

To test these hypotheses, we use the MinAtar suite of five Atari inspired videogames (Young & Tian, 2019) and the DMC 15 suite (Tunyasuvunakool et al., 2020).[6] Both are small enough to perform thorough investigations, while providing non-trivial observation spaces and dynamics. Detailed information about the implementation and hyperparameters can be found in Appendix E.

Auxiliary task learning vs general purpose feature learning (Figure 2 and Figure 3): First, we compare both the auxiliary task and stand-alone feature learning scenarios. As expected from prior work (Jaderberg et al., 2016; Schwarzer et al., 2021; Farebrother et al., 2023), in all cases using an auxiliary loss performs no worse (and often better) than vanilla DQN. We find that as expected from Insight 3, latent self-prediction is a stronger auxiliary loss function than observation reconstruction in three out of five environments. However, when using the decision-agnostic losses alone, we clearly see observation reconstruction performing significantly better than latent self-prediction, which fails to learn any relevant features in several cases. This verifies Insight 1.

Curiously, in the case of the Seaquest environment, we find that using observation prediction alone outperforms using it as an auxilliary task strongly, and performs on par with the auxiliary task

[6] DMC experiments are presented in the appendix due to space constraints.

variant of the latent self prediction loss. Seaquest also has the sparsest reward structure in the test suite, which can make it a challenging environment for DQN. In this case, features based purely on the observed transition might allow for better policy learning.

This highlights that no algorithm is clearly superior in all settings and the reward and observation structure is very relevant for the performance of each loss.

Observation space distortions (Figure 4): To test the impact of changing the observation function, we sample a random binary matrix and multiply it to the flattened observation vector. We then reshape the observation to the original shape again.

All algorithms show themselves to be strongly impacted by this random observation distortions, which suggests that our claim of invariance of self-prediction relies too strongly on the linear gradient-flow limit. This can in part be explained by the use of a convolutional layer in the standard baseline implementation of DQN which we adapted. However, we still find that at least on two environments (Seaquest and Freeway), the latent self-prediction auxiliary task is able to recover more of the original performance than either observation prediction or DQN. Interestingly, the DQN baseline seems to suffer the most from the introduction of the observation change, which suggests that correlations of the existing observation space play an important role in learning correct value function prediction.

Overall, we find that Insight 2 does not fully translate to the more complex test setting. In part, this may be explained by the fact that the original observation spaces of the test environments already violate our assumptions for the one-hot encoding. In addition, introducing linear correlation might not impact non-linear model learning in the same way it would impact linear models. This highlights the need for more in-depth research on the interplay between given observation space and feature learning.

Distractions (Figure 5 and Figure 6): As our results are dependent on the spectral structure of the environment, different distraction models can be assumed to have differing impact on the efficacy of the tested losses. This behavior is dependent on the structure of the noise. If the distraction does not strongly change the top-k singular or eigenspaces, it will be less problematic for the auxiliary tasks, especially for observation reconstruction. Testing the impact of different distraction models on the top-k spaces is out of scope for this work, but we conjecture that fully random noise has less structure than distractions following clear patterns.

Therefore, we consider two simple distraction models in our experiments. The distractions are concatenated to the original observations along the channel dimension. First, we use random noise sampled independently for each state from a Bernoulli distribution. As there is no predictable structure in this noise, we expect all algorithms to be able to deal with this distraction better. Second, we choose one of the environments at random (Freeway-v1) and concatenate two copies to the observation space of each environment. The dynamics are obtained by sampling a random action at each timestep independent from the policy and stepping the distraction environment with it.

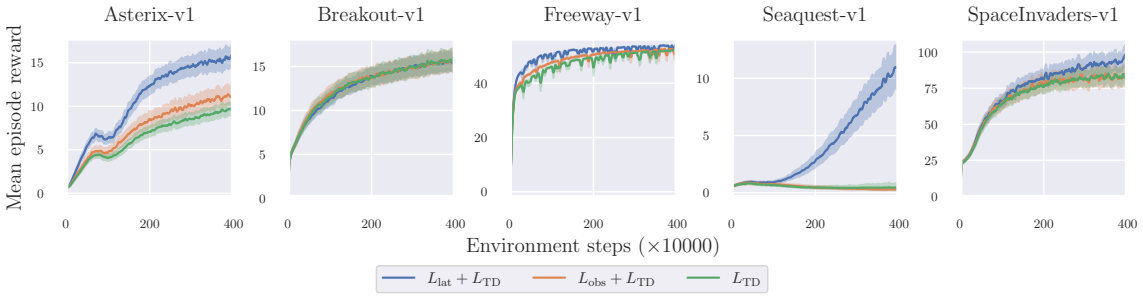

Figure 2: Auxiliary task setup: Performance of all losses on the observation space as given without changes to the environment.

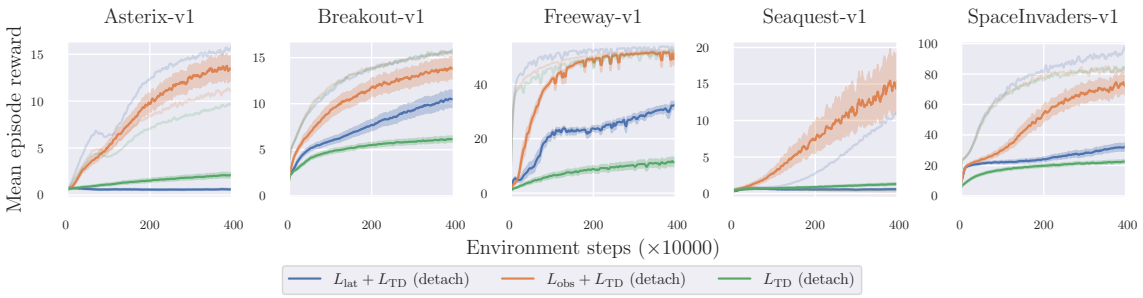

Figure 3: Stand-alone setup: Performance of all losses on the observation space as given without changes to the environment. The DQN baseline is using random features, which are not updated, to verify that learning features is indeed superior to a random feature baseline.

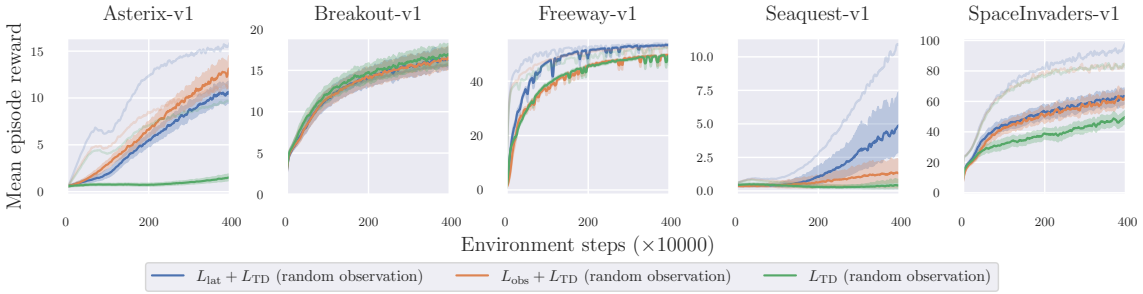

Figure 4: Distorted observation function with a random transformation.

We find a small advantage in some environments to using the latent self-prediction loss and using random noise, and no clear advantage from any algorithm in the structured noise case. Structured noise poses a much larger challenge to most algorithms, completely preventing learning in several cases. This partially validates that not only the presence or absence of noise matters, but also how it changes relevant quantities, e.g. the eigenvalues of the transition kernel.

In the continuous control experiments presented in Appendix F, we find that the self-prediction loss performs generally better than in the MinAtar suite. As the observation and reward structure differs between these two benchmark suite, this obvservation lends more credibility to our claims that the observation structure impacts the performance of algorithms.

7 Conclusions

When choosing an RL approach to use, practitioners are overwhelmed with a variety of loss function choices, without clear indication which one will be preferable in what scenario. In our work, we introduce analytically tractable notions of distractions and observation functions. With these we predict the performance of latent self-prediction and observation reconstruction as stand-along feature learning methods and auxiliary tasks, and study our theoretical insights empirically. Our evaluation lends credibility to the use of simple surrogate models to obtain practically relevant insights into algorithmic performance. However, in several cases we also find deviations between our predictions and more complex benchmarks. Therefore, while we claim that our experiments have the ability to guide the choice of algorithms for applied settings, there is still a sizeable gap between theory and practice that remains to be bridged in future work.

We also note that our experiments showed substantial differences in behavior of auxiliary losses both within and across benchmarks and different noise distractions. Previous work that studied the effect of distraction (Nikishin et al., 2022a; Voelcker et al., 2022; Ni et al., 2024) did not discuss their

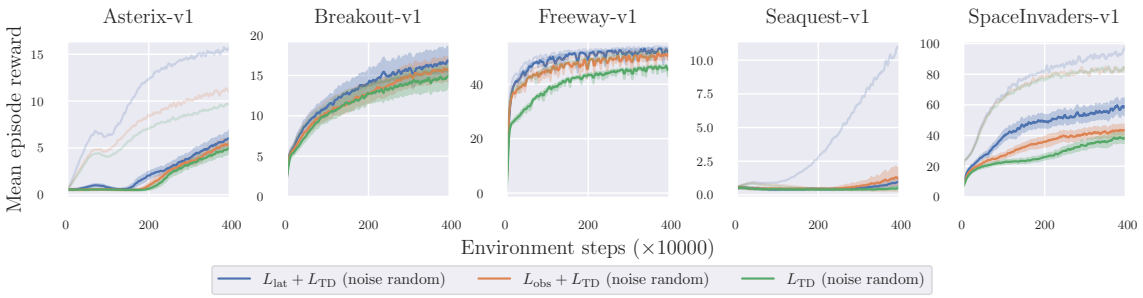

Figure 5: Appending random noise channels.

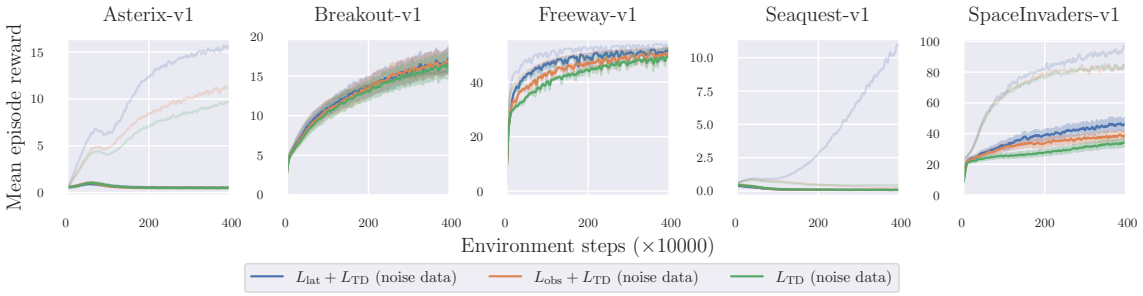

Figure 6: Appending structured noise channels using the Freeway environment.

distraction model in further details. In light of our results, we suggest that empirical research should be careful about the choice of benchmark and experimental setup and discuss the implications of the empirical setup explicitely.

One of the most important gaps between the work presented in this paper and the behavior of online algorithms is the restrictive assumption of the fixed policy evaluation case. Therefore one of the most exciting avenues for future work is analysing policy improvement, where the underlying dynamics of the environment change due to the policy updates. On the empirical side the surprising effectiveness of the observation prediction loss on the Seaquest environment highlights the fact that even within benchmark suites, differences in the reward functions and observation models can lead to differing rankings between algorithms. This further highlights the necessity of studying the structure of MDPs and to design algorithms that are robust to different structures, or adapt to them automatically.

Acknowledgments

We would like to thank the members of the Adaptive Agents Lab and the Toronto Intelligent Systems Lab for their help and feedback on the ideas and the draft. CV thanks Charline Le Lan and Anvith Thudi for help with the stability proofs and other mathematical details. AMF acknowledges the funding from the Canada CIFAR AI Chairs program, as well as the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. *Advances in Neural Information Processing Systems*,

2020.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 2017.

Bahram Behzadian, Soheil Gharatappeh, and Marek Petrik. Fast feature selection for linear value function approximation. In *International Conference on Automated Planning and Scheduling*, 2019.

Marc G Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. *Advances in neural information processing systems*, 2019.

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial intelligence*, 2000.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 1993.

Amir-massoud Farahmand, André Barreto, and Daniel Nikovski. Value-Aware Loss Function for Model-based Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Jesse Farebrother, Joshua Greaves, Rishabh Agarwal, Charline Le Lan, Ross Goroshin, Pablo Samuel Castro, and Marc G Bellemare. Proto-value networks: Scaling representation learning with auxiliary tasks. In *International Conference on Learning Representations*, 2023.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, 2018.

Dibya Ghosh and Marc G Bellemare. Representations for stable off-policy reinforcement learning. In *International Conference on Machine Learning*, 2020.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 1975.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *International Conference on Learning Representations*, 2016.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021.

Charline Le Lan, Marc G Bellemare, and Pablo Samuel Castro. Metrics and continuity in reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pp. 8261–8269, 2021.

Charline Le Lan, Stephen Tu, Adam Oberman, Rishabh Agarwal, and Marc G Bellemare. On the generalization of representations in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Charline Le Lan, Stephen Tu, Mark Rowland, Anna Harutyunyan, Rishabh Agarwal, Marc G Bellemare, and Will Dabney. Bootstrapped representations in reinforcement learning. In *International Conference on Machine Learning*, 2023.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *International Conference on Learning Representations*, 2016.

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*, 2021.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Marlos C Machado, Andre Barreto, Doina Precup, and Michael Bowling. Temporal abstraction in reinforcement learning with the successor representation. *Journal of Machine Learning Research*, 24(80):1–69, 2023.

Sridhar Mahadevan. Learning representation and control in markov decision processes: New frontiers. *Foundations and Trends® in Machine Learning*, 2009.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NeurIPS Deep Learning Workshop*. 2013.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. *To appear in International Conference on Learning Representations*, 2024.

Evgenii Nikishin, Romina Abachi, Rishabh Agarwal, and Pierre-Luc Bacon. Control-oriented model-based reinforcement learning with implicit differentiation. In *AAAI Conference on Artificial Intelligence*, 2022a.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning*, 2022b.

Marek Petrik. An analysis of laplacian methods for value function approximation in mdps. In *IJCAI*, pp. 2574–2579, 2007.

Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning*, 2018.

Martin L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. In *Wiley Series in Probability and Statistics*, 1994.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

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.

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite - a challenging benchmark for reinforcement learning from pixels. *ArXiv*, abs/2101.02722, 2021.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

Yunhao Tang and Rémi Munos. Towards a better understanding of representation dynamics under td-learning. In *International Conference on Machine Learning*, 2023.

Yunhao Tang, Zhaohan Daniel Guo, Pierre Harvey Richemond, Bernardo Ávila Pires, Yash Chandak, Rémi Munos, Mark Rowland, Mohammad Gheshlaghi Azar, Charline Le Lan, Clare Lyle, et al. Understanding self-predictive learning for reinforcement learning. *International Conference on Machine Learning*, 2022.

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

Manan Tomar, Utkarsh Aashu Mishra, Amy Zhang, and Matthew E. Taylor. Learning representations for pixel-based control: What matters and why? *Transactions on Machine Learning Research*, 2023.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6, 2020.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI Conference on Artificial Intelligence*, 2016.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. *International Conference on Learning Representations*, 2022.

Yifan Wu, George Tucker, and Ofir Nachum. The laplacian in RL: Learning representations with efficient approximations. In *International Conference on Learning Representations*, 2019.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

A Motivating examples

A.1 Motivating example for observation spaces

To further motivate our focus on observation models, consider the common problem of representing rotational angles. As these are continuous values, discretized one-hot representations may introduce errors that may harm efficient control. When choosing a continuous representation, designers are faced with the choice between representing the angle as $\omega \in [-\pi, \pi)$ (the exclusion of the right endpoint is arbitrary), or by a decomposition into $[\sin(\omega), \cos(\omega)] \in [0, 1]^2$. The choice of representation has measurable impact on the ease of learning: the former is 1D instead of 2D, which can reduce the size of networks needed when dealing with many angles. However a complication with the first representation is it is discontinuous at the right endpoint, in the sense that $\lim_{x\to\pi} x = -\pi$ (this is generally due to the structure of $\mathbb{R}/2\pi\mathbb{Z}$). This creates a peculiar continuity condition for functions on this representation space: for a function f to be continuous, it must be continuous and also satisfy the boundary condition $\lim_{x\to\pi} f(x) = f(-\pi)$. Without explicitly enforcing this, the majority of functions learned on this domain (such as estimated value functions) will be discontinous, leading to additional difficulties in the learning process. On the other hand, this issue does not exist for the second representation, potentially making learning much easier.

The important question is that of *similarity* and *continuity*: in our pendulum example, the more compact representation breaks the intuitive notion that states that behave similariy should be mapped to close representations.

A.2 Distractions

The optimality of features can be measured in how close the projection of the true value function onto their span is to the ground truth, i.e. in the L_2 norm. This raises the question under what conditions the top k eigenvectors or singular vectors would not span the value function well. For this, we turn to our notion of an MDP with distractions.

Proposition 7 (Suboptimality of top k eigenspaces with distractions)**.** *Assume an MDP with distraction composed of two independent processes $\mathcal{M}$ and $\mathcal{N}$ according to Definition 1. Let $v_1, \ldots, v_n$ and $u_1, \ldots, u_m$ be the eigenvectors of $\mathcal{M}$ and $\mathcal{N}$ respectively, with associated real eigenvalues λ_j and μ_i ordered. Assume $\forall i < k : \mu_i > \lambda_2$ and $r_{\mathcal{N}} = 0$. Let U_k be an orthonormal basis for the top-k eigenspace of $\mathcal{M} \otimes \mathcal{N}$. Then* $\operatorname{span}(U_k) = \operatorname{span}(\{\mathbf{1} \otimes v_i | \forall i \leq k\})$ *and* $\operatorname{proj}_{U_k}(r_{\mathcal{M}} \otimes \mathbf{1}) = (\sum_{i=0}^{m} r_i/n)\mathbf{1}_{n\cdot m}$.

This means that there is a natural notion of a process in which the top-k eigenvectors do not span the reward function well. In this case, the distracting process has larger eigenvalues than the reward-relevant process. As the eigenvector basis is composed of Kronecker products of the individual eigenvectors, the top-k eigenspace contains redundant copies of the reward relevant processes eigenvectors. By Lemma 5, this directly implies suboptimal value function approximation. Note that our assumptions here are restrictive as we consider a *worst case* distraction for clarity, but the problem emerges whenever the distracting process has several large eigenvalues.

B Related work

In this work, we take inspiration from the work of Tang et al. (2022) and Le Lan et al. (2023), which present analyses of representation dynamics under various losses. The main purpose of our work is to take a closer look at the implicit assumptions on the MDP structure underlying previous work to obtain a better picture of the empirical behavior of the analyzed algorithms.

Bellemare et al. (2019) shows how the space of linear features relates to optimal value functions. Building on this, Le Lan et al. (2022) analyzes the optimality of, while Le Lan et al. (2021) focus on the topological properties of representation functions. Our work expands on these by focusing on the properties of features learned by common approaches instead of reasoning about optimal features.

Lyle et al. (2021) and Farebrother et al. (2023) analyse the impact of random cumulants and auxiliary value function prediction on the loss dynamics. As Le Lan et al. (2023) show, the induced gradients are similar to those we analyze for latent self-prediction, adding another avenue for understanding the design of auxiliary tasks. As latent self-prediction and auxiliary tasks lead to similar stationary points, it is an exciting avenue for future work to use the spectral properties of existing reward functions to design better auxiliary rewards, i.e. in environments with distractions.

Tomar et al. (2023) asks a closely related question to ours ("What matters for reinforcement learning from pixels?") and provide an extensive empirical study. Many of their findings are corroborated in our experiments and strengthened by our theoretical analysis. They finding only limited benefit for latent self-prediction, which is different from our finding, however they focused on more complex visual observation spaces, which might highlight that the benefits of this approach are more strongly observed in smaller representation spaces.

Linear features have also been studied in the context of the graph Laplacian (Petrik, 2007; Mahadevan, 2009; Wu et al., 2019) and successor features Dayan (1993); Barreto et al. (2017), which result in feature construction methods that closely resemble those learned by current methods. Our work directly relates modern deep learning methods to these classic approaches for feature construction. Furthermore, Machado et al. (2023) uses explicit graph drawing objectives to extract eigenvector representations from transition matrices. Comparing the stability and synergy of these approaches in terms of the training dynamics with auxiliary objectives such as those discussed here is a fruitful direction for future work.

Finally, Ni et al. (2024) recently published an analysis on the relative performance of different auxiliary tasks in deep reinforcement learning. Our analysis is orthogonal and complimentary to theirs. While they focus on highlighting the benefits of latent self-prediction in the idealized limit of perfect predictions, we analyse the case where trade-offs have to be made due to compression. In addition, while they highlight the noise robustness of latent self-prediction, they show no formalization or theoretical justification. Our analysis on the other hand shows the strong impact that the form of the distracting dimensions have on the problem solution.

C Helpful definitions and lemmata

This section contains helpful lemmata that are used for our proofs. Where we took these from existing work, we provide references, otherwise the proofs are our own, although probably also known in the literature.

C.1 Linear Algebra

As before, for matrices V we write $\text{span}(V)$ to denote the span of their column vectors.

Definition 2 (Top-k singular vectors)**.** *Let $P^\pi = U^\top \Sigma V$ be the singular value decomposition, and assume that the diagonal of Σ is arranged in decreasing order. The first k rows of $U^\top$ and the first k columns of V are called the top-k left and right singular vectors, respectively.*

Lemma 1 (Spectrum of Kronecker product matrix)**.** *Consider two non-singular matrices M and N. Let λ_i be the eigenvalues of M and μ_j be the eigenvalues of N, with eigenvectors u_i and v_j respectively. Then the eigenvalues of $M \otimes N$ are $\lambda_i \mu_j$ with eigenvectors $u_i \otimes v_j$ respectively.*

Proof. For any eigenvector u_i of M and v_j of N, we have

$$(M \otimes N)(u_i \otimes v_j) = (Mu_i) \otimes (Nv_j) = (\lambda_i u_i) \otimes (\mu_j v_j) = \lambda_i \mu_j (u_i \otimes v_j).$$

So $u_i \otimes v_j$ is an eigenvector of $M \otimes N$ with eigenvalue $\lambda_i \mu_j$ □

Lemma 2 (Orthogonalized bases and the Kronecker product)**.** *Let $V \in \mathbb{R}^{n \times m}$ be a matrix. Let $\text{orth}(V)$ be a matrix of any orthonormal basis vectors for the column vectors of V. Then $\text{span}(V \otimes \mathbf{1}) = \text{span}(\text{orth}(V) \otimes \mathbf{1})$. Furthermore $(1/k)\text{orth}(V) \otimes \mathbf{1}$ is an orthonormal basis of $\text{span}(V \otimes \mathbf{1})$.*

Proof. Let $V_i \otimes \mathbf{1}^k$ be any column vector of $V \otimes \mathbf{1}^k$. Let $\alpha_1, \ldots, \alpha_m$ be coefficients so that $V_i = \sum_{j=1}^m \alpha_j \text{orth}(V)_j$. Then

$$\sum_{j=1}^m \alpha_j \left(\text{orth}(V)_j \otimes \mathbf{1}^k\right) = \left(\sum_{j=1}^m \alpha_j \text{orth}(V)_j\right) \otimes \mathbf{1}^k = V_i \otimes \mathbf{1}^k,$$

following the standard associative and distributive properties of the Kronecker product.

As every vector in the original span can be represented in the orthogonalized span, the two are equivalent.

Finally note that

$$\frac{1}{k}\left(\text{orth}(V) \otimes \mathbf{1}\right)_i^\top \left(\text{orth}(V) \otimes \mathbf{1}\right)_i = \frac{1}{k}\sum_{j=1}^k \left(\text{orth}(V)_i^\top \text{orth}(V)_i^\top\right) = 1, \text{ and}$$

$$\frac{1}{k}\left(\text{orth}(V) \otimes \mathbf{1}\right)_j^\top \left(\text{orth}(V) \otimes \mathbf{1}\right)_i = \frac{1}{k}\sum_{t=1}^k \left(\text{orth}(V)_j^\top \text{orth}(V)_i^\top\right) = 0.$$

□

Lemma 3 (Reduced rank regression)**.** *Let $C, D \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times k}$, and $B \in \mathbb{R}^{k\times n}$ be full rank matrices with $n \geq k$. Let $A^*, B^* = \min_{A,B} \|CAB - DC\|_F^2$. Let $u_1, \ldots, u_k$ and $v_1 \ldots, v_k$ be the top-k singular vectors of $C^{-1}DC$ according to Definition 2. Then* $\text{span}(A^*) = \text{span}(u_1, \ldots, u_k)$ *and* $\text{span}(B) = \text{span}(v_1, \ldots, v_k)$.

This is a reduced rank regression problem (Izenman, 1975) or low-rank matrix approximation problem.

The unconstrained solution to the problem is given by $\hat{A}\hat{B} = C^{-1}DC$. From the Eckhart-Young theorem, we know that the optimal low-rank approximation to $C^{-1}DC$ is given by the top-k singular vectors. Therefore A and B span top-k left and right singular vectors respectively. For a more extensive proof, please refer to Izenman (1975).

C.2 Stochastic matrices

Lemma 4 (Spectrum of a resolvent matrix)**.** *Let A be an invertible matrix with unique real eigenvalues and $-1 \leq \lambda_{\min} \leq \lambda_{\max} \leq 1$. Let $\gamma \in (0, 1)$. The matrices A and $(I - \gamma A)^{-1}$ have the same eigenvectors and the ordering of the corresponding eigenvalues remains the same.*

Proof. Let e be an eigenvector of A and λ the corresponding eigenvalue. Then

$$(I - \gamma A)^{-1} e = \sum_{i=0}^n \gamma^n A^n e = \sum_{i=0}^n (\gamma\lambda)^n e = \frac{1}{1-\gamma\lambda} e.$$

As $(1 - \gamma\lambda)^{-1}$ is a monotonic function for $-1 \leq \lambda \leq 1$ and $\gamma \in [0, 1]$, the ordering of the eigenvalues remains the same. □

Lemma 5 (Basis equivalence of linear reward and value function)**.** *Let r be the vector representation of the reward function and V of the value function respectively for an MDP with fixed policy and transition matrix P^π. Let $U = \{u_1, \ldots, u_n\}$ be the set of eigenvectors of P^π, and let $U^r \subseteq U$ be a minimal set of eigenvectors so that $r \in \text{span}(U^r)$. Then $V \in \text{span}(U^r)$.*

Proof. Let $r = \sum_{i=1}^k \alpha_i u_i^r$ be the reward representation in the basis U^r. Then, by Lemma 4

$$V^\pi = (I - \gamma P^\pi)^{-1} r = \sum_{i=1}^k \frac{\alpha}{1 - \lambda_i \gamma} u_i^r.$$

□

C.3 Ordinary differential equations

For describing the stability of ODEs, we use the notion of *asymptotic stability*, with the condition $\Re(\lambda_i) < 0$ for all eigenvalues λ_i of the Jacobian at a critical point.

Lemma 6 (Linear reparameterization of an autonomous ODE)**.** *Let $y' = f(y)$ be an autonomous ordinary differential equation. Let y^* be any critical point for which $f(y^*) = 0$. Let furthermore $x' = f(Ax)$ be a reparameterized autonomous ODE for any invertible matrix A. Then the x^* is a critical point with $f(Ax) = 0$ iff $x^* = A^{-1}y^*$. Furthermore, the eigenvalues of the Jacobian of y' at y^* are equal to the eigenvalues of the Jacobian of x' at $x^* = A^{-1}y^*$.*

Proof. The direction $x = A^{-1}y^* \Rightarrow f(Ax) = 0$ is clear by direct evaluation. We now focus on the direction $f(Ax) = 0 \Rightarrow x = A^{-1}y^*$. Assume that $f(Ax)$ is 0 and $x = A^{-1}y$ for a y which is not a point satisfying $f(y) = 0$. But then $0 = f(AA^{-1}y) = f(y) \neq 0$, a contradiction.

For stability, note that

$$\begin{aligned}
& & \frac{\mathrm{d}}{\mathrm{d}t}y &= f(y) \\
&\implies & \frac{\mathrm{d}}{\mathrm{d}t}x &= \frac{\mathrm{d}}{\mathrm{d}y}x\frac{\mathrm{d}}{\mathrm{d}t}y \\
& & &= A^{-\top}f(y) \\
&\implies & \frac{\mathrm{d}}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}t}x &= A^{-\top}\frac{\mathrm{d}}{\mathrm{d}x}f(Ax) \\
& & &= A^{-\top}\frac{\mathrm{d}}{\mathrm{d}y}f(y)\frac{\mathrm{d}}{\mathrm{d}x}Ax \\
& & &= A^{-\top}\frac{\mathrm{d}}{\mathrm{d}y}f(y)A^{\top}.
\end{aligned}$$

This shows that the Jacobian $\frac{\mathrm{d}}{\mathrm{d}x}f(Ax)|_{x=A^{-1}y_0}$ is similar to the Jacobian $\frac{\mathrm{d}}{\mathrm{d}y}f(y)|_{y=y_0}$, which means their spectra are identical. □

C.4 MDP representation and TD learning

Lemma 7 (Lemma 5 of Tang et al. (2022))**.** *Suppose P^π is real-diagonalizable, and write $u_1, \ldots, u_n$ for its eigenvectors. Then any orthonormal matrix Φ which has the same span as a set of k eigenvectors is a minimizer of L_{lat}.*

The next three statements are taken from Ghosh & Bellemare (2020). They address the TD loss wrt to the learned weights $\hat{V}$ and fixed Φ (compare Subsection 2.1). We changed the notation of the statements to fit our notation here, we have denoted the diagonal matrix of the state distribution as D and assume that $D = I$ in Assumption 1, while Ghosh & Bellemare (2020) uses Ξ. They use θ for the value function weights while we use $\hat{V}$. They also uses $\mathrm{Spec}(A)$ to denote the spectrum, the set of all eigenvalues of a matrix A.

The notion of stability used in Ghosh & Bellemare (2020) is that of convergence to the unique fixed point of the projected Bellman update of the linear ODE induced by the L_{TD} loss when fixing Φ. In the two-timescale scenario considered in this paper, this corresponds to the "inner" ODE over $\hat{V}$.

Lemma 8 (Proposition 3.1 of Ghosh & Bellemare (2020))**.** *TD(0) is stable if and only if the eigenvalues of the implied iteration matrix $A_\Phi = \Phi^\top D(I - \gamma P^\pi)\Phi$ have positive real components, that is*

$$Spec\,(A_\Phi) \subset \mathbb{C}_+ := \{z : Re(z) > 0\}.$$

We say that a particular choice of representation Φ is stable for (P^π, γ, D) when A_Φ satisfies the above condition.

Lemma 9 (Proposition 3.2 of Ghosh & Bellemare (2020))**.** *An orthogonal representation Φ is stable if and only if the real part of the eigenvaluse of the induced transition matrix $\Pi P^\pi \Pi$ where $\Pi = \Phi\Phi^\top$ is bounded above, according to*

$$Spec\,(\Pi P^\pi \Pi) \subset \{z \in \mathbb{C} : Re(z) < \frac{1}{\gamma}\}.$$

In particular, Φ is stable if $\rho(\Pi P^\pi \Pi) < \frac{1}{\gamma}$.

Lemma 10 (Theorem 4.1 of Ghosh & Bellemare (2020))**.** *An orthogonal invariant representation Φ (meaning $span(P^\pi\Phi) \subseteq span(\Phi)$) satisfies*

$$Spec\,(\Pi P^\pi \Pi) \subseteq Spec(P^\pi) \cup \{0\}$$

and is therefore stable.

As a corollary of their proof we have that

Lemma 11 (Corollary of Ghosh & Bellemare (2020))**.** *Let Φ be an orthogonal (but not necessarily square) invariant representation of P^π. Then the spectral radius $\rho(\Phi^\top P^\pi \Phi) \leq 1$.*

The proof follows directly from Lemma 10 by the cyclicality of the spectrum.

The following two results are our own, although closely related results exist in the literature.

Lemma 12 (Lossless approximation of V)**.** *Let Φ be an orthonormal basis of an invariant subspace of P^π and let Assumption 4 hold. Let $V = (I - \gamma P^\pi)^{-1} r$ be the value function of P^π and r^π. Then*

$$\Phi(I - \gamma\Phi^\top P^\pi \Phi)^{-1}\Phi^\top r^\pi = V.$$

Proof. By Assumption 4 we can find a matrix A so that $r^\pi = \Phi A$ and $\Phi\Phi^\top r^\pi = r^\pi$, and by the invariant subspace assumption, we can find a matrix B so that $P^\pi\Phi = \Phi B$. Writing the inverted matrix as an infinite sum, which is valid as the spectrum of $\Phi^\top P^\pi \Phi$ is bounded by 1 following from Lemma 11 and Carl Neumann's theorem over power series, we obtain

$$\begin{aligned}
\Phi(I - \gamma\Phi^\top P^\pi \Phi)^{-1}\Phi r^\pi &= \Phi \sum_{n=0}^{\infty} \gamma^n (\Phi^\top P^\pi \Phi)^n \Phi r^\pi \\
&= \Phi \sum_{n=0}^{\infty} \gamma^n B^n \Phi^\top r^\pi && (P^\pi\Phi = \Phi B) \text{ and } (\Phi^\top\Phi = I) \\
&= \sum_{n=0}^{\infty} \gamma^n \Phi B^n \Phi^\top r^\pi \\
&= \sum_{n=0}^{\infty} \gamma^n {P^\pi}^n \Phi\Phi^\top r^\pi && (\Phi B = P^\pi\Phi) \text{ iterated} \\
&= \sum_{n=0}^{\infty} \gamma^n {P^\pi}^n r^\pi && (\Phi\Phi^\top r^\pi = r^\pi) \\
&= V
\end{aligned}$$

□

Lemma 13 (Critical points of L_{TD})**.** *Let Assumption 1, Assumption 3, and Assumption 4 hold. Assume $\Phi^* \in \mathbb{R}^{n\times k}$ is an orthonormal invariant representation for P^π in the sense that ${\Phi^*}^\top\Phi^* = I$ and $span(\Phi^*) = span(P^\pi\Phi^*)$. Let furthermore $r^\pi \in span(\Phi^*)$ and let V be the corresponding value function. Then Φ^* is a critical point of L_{TD} and for the corresponding weights $\hat{V}^*$ we have $\Phi^*\hat{V}^* = V$.*

Proof. By Lemma 8, Lemma 9, and Lemma 10 and the stated assumptions the weights $\hat{V}$ converge. Therefore, we can analyze the induced dynamical system with $\hat{V}*$.

Find $\hat{V}^*$ as the stationary point of $\nabla_{\hat{V}} L_{\text{TD}}$

$$\begin{aligned}
& \nabla_{\hat{V}} \left\| \Phi^* \hat{V} - \left[r^\pi + \gamma P \Phi^* \hat{V} \right]_{\text{sg}} \right\|_2^2 \Big|_{\hat{V}=\hat{V}^*} = {\Phi^*}^\top \left(\Phi^* \hat{V}^* - \left[r^\pi + \gamma P \Phi^* \hat{V}^* \right] \right) = 0 \\
\Leftrightarrow \quad & {\Phi^*}^\top \Phi^* \hat{V}^* - {\Phi^*}^T r^\pi - \gamma {\Phi^*}^\top \Phi^* B \hat{V}^* = ({\Phi^*}^\top \Phi^* - \gamma {\Phi^*}^\top P^\pi \Phi^*) \hat{V}^* - {\Phi^*}^\top r^\pi = 0 \\
\Leftrightarrow \quad & \hat{V}^* = ({\Phi^*}^\top \Phi^* - \gamma {\Phi^*}^\top P^\pi \Phi^*)^{-1} {\Phi^*}^\top r^\pi = (I - \gamma {\Phi^*}^\top P^\pi \Phi^*)^{-1} {\Phi^*}^\top r^\pi
\end{aligned}$$

The invertibility of $(I - \gamma \Phi^\top P^\pi \Phi) = \Phi^T D (I - \gamma P^\pi) \Phi = A_\Phi$ is guaranteed as all eigenvectors are nonzero.

We now show that Φ^* is a stationary point by showing that

$$\nabla_\Phi L_{\text{TD}} |_{\Phi=\Phi^*} = 0.$$

We note that as Φ^* spans an invariant subspace of P^π there exists an invertible matrix B so that $P^\pi \Phi^* \Phi^* B$. Therefore $(I - \gamma {\Phi^*}^\top P^\pi \Phi^*) = (I - \gamma B)$.

$$\begin{aligned}
\nabla_\Phi L_{\text{TD}} |_{\Phi=\Phi^*} &= \nabla_\Phi \left\| \Phi \hat{V}^* - \left[r^\pi + \gamma P \Phi \hat{V}^* \right]_{\text{sg}} \right\|_2^2 \Big|_{\Phi=\Phi^*} \\
&= \left(\Phi \hat{V}^* - r^\pi - \gamma P \Phi \hat{V}^* \right) (\hat{V}^*)^\top \Big|_{\Phi=\Phi^*} \\
&= \left(\Phi^* (I - \gamma B) \hat{V}^* - r^\pi \right) (\hat{V}^*)^\top \qquad \text{(substitute first occurrence of) } \hat{V}^* \\
&= \left(\Phi^* {\Phi^*}^\top r^\pi - r^\pi \right) (\hat{V}^*)^\top \\
&= \underbrace{(r^\pi - r^\pi)}_{=0} (\hat{V}^*)^\top = 0
\end{aligned}$$

The final line is due to the fact that the columns of Φ^* are orthonormal, which means that $\Phi^* {\Phi^*}^\top$ is an orthogonal projection onto the span of Φ^*. To verify, note that

$$\left(\Phi^* {\Phi^*}^\top \right)^2 = \left(\Phi^* \underbrace{{\Phi^*}^\top \Phi^*}_{=I} {\Phi^*}^\top \right) = \left(\Phi^* {\Phi^*}^\top \right).$$

Furthermore, by Assumption 4, $r^\pi \in \text{span}(\Phi^*)$, which means $\Phi^* {\Phi^*}^\top r^\pi = r^\pi$ for an orthogonal projection $\Phi^* {\Phi^*}^\top$. Moreover, by Lemma 12, $\Phi^* \hat{V}^* = V$, which concludes the proof. □

This proof closely follows related statements by Ghosh & Bellemare (2020), Tang et al. (2022), and Le Lan et al. (2022). We repeated the argument here for easier legibility with all assumptions necessary for our work.

D Proofs of main results

Proofs for Section 4

Proposition 1 (Stationary points of latent self-prediction)**.** *Assume Assumption 1 and Assumption 2 hold. Furthermore, suppose P^π is real diagonalizable. If the columns of Φ_t span an invariant subspace of P^π, Φ_t is a stationary point of the dynamical system. Furthermore, if P^π is real-diagonalizable with positive eigenvalues, all invariant subspaces not spanned by the top-k eigenvectors sorted by eigenvalue are asymptotically unstable for gradient descent.*

Proof. At any stationary point the gradient $\frac{d}{dt}\Phi_t$ must be equal to 0, which from Section 2.1 means that we must have $\left(\Phi_t\left(\Phi_t^\top\Phi\right)^{-1}\Phi_t^\top - I\right)P^\pi\Phi_t\Phi_t^\top {P^\pi}^\top\Phi_t\left(\Phi_t^\top\Phi\right)^{-\top} = 0$.

Assume that the column vectors of Φ^* spans an invariant subspace of P^π. This implies that there exists a full rank matrix A so that $P^\pi\Phi^* = \Phi^* A$. Then

$$\left(\Phi^*\left({\Phi^*}^\top\Phi^*\right)^{-1}{\Phi^*}^\top - I\right)P^\pi\Phi^*F^* = \left(\Phi^*\underbrace{\left({\Phi^*}^\top\Phi^*\right)^{-1}{\Phi^*}^\top\Phi^*}_{=I} - \Phi^*\right)AF^* = 0.$$

This proves the first part of the proposition.

There are additional critical points of the differential equation, as discussed by Tang et al. (2022). In the analysis of stability, we first show the case of critical points corresponding to the claim in the proposition. We then briefly discuss other cases after the proof.

Case 1: Φ_t spans an invariant subspace of P^π Invariant subspaces correspond to subspaces spanned by right eigenvectors of P^π.

We write P for P^π to reduce notational clutter. Let $e_1, \ldots, e_k$ be the eigenvectors corresponding to the k largest eigenvalues of P. Let Φ^* correspond to any set of k eigenvectors of P. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^* = -\left(\Phi^*F^* - P\Phi^*\right){F^*}^\top = 0.$$

To show that all non top-k eigenspaces are asymptotically unstable critical points of the differential equation defined by the gradient flow of Φ. To show this, we aim to show that there exists an eigenvector of the Jacobian with an eigenvalue larger than 0. For this, we construct the directional derivative at the critical point. The directional derivative is the Jacobian vector product, which allows us to circumvent the need to work with higher order tensor derivatives. We then proceed to show that there exists a direction which corresponds to the eigenvector of the Jacobian with positive eigenvalue. This concludes the proof. This technique closely follows the one used by Le Lan et al. (2023).

Assume $\mathrm{span}\{\Phi^*\} \neq \mathrm{span}\{e_1, \ldots, e_k\}$. This implies that there exists at least one eigenvector $e_j \in \{e_1, \ldots, e_k\}$ and $e_j \notin \mathrm{span}\{\Phi^*\}$, with corresponding eigenvalues λ_j.

Let D_Δ be the directional derivative of $\frac{\mathrm{d}}{\mathrm{d}t}\Phi|_{\Phi=\Phi^*}$ in the direction Δ. We construct the directional derivative using the product rule (terms colored for ease of reading),

$$\begin{aligned} D_\Delta\frac{\mathrm{d}}{\mathrm{d}t}\Phi|_{\Phi=\Phi^*} &= -D_\Delta\left(\left(\Phi F^* - P\Phi\right){F^*}^\top\right)|_{\Phi=\Phi^*} \\ &= -D_\Delta\left(\Phi F^* - P\Phi\right)|_{\Phi=\Phi^*}{F^*}^\top - \left(\Phi^*F^* - P\Phi^*\right)D_\Delta F^*|_{\Phi=\Phi^*}{}^\top. \end{aligned}$$

For the directional derivative, we only consider directions that are orthogonal to Φ^*, so ${\Phi^*}^\top\Delta = 0$. Then $\underbrace{P\Phi^* = \Phi^* A}_{\text{subspace condition}} \implies \Delta^\top P\Phi^* = 0$. For the derivative with regard to F^*, we have

$$\begin{aligned} D_\Delta F^*|_{\Phi=\Phi^*} =& D_\Delta\left(\Phi^\top\Phi\right)^{-1}\Phi^\top P\Phi \\ =& \left(D_\Delta\left(\Phi^\top\Phi\right)^{-1}\right){\Phi^*}^\top P\Phi^* + \left({\Phi^*}^\top\Phi^*\right)^{-1}\left(D_\Delta\Phi^\top P\Phi\right) \\ =& -\underbrace{\left(\Delta^\top\Phi^* + {\Phi^*}^\top\Delta\right)}_{=0}\left({\Phi^*}^\top\Phi^*\right)^{-2}{\Phi^*}^\top P\Phi^* + \left({\Phi^*}^\top\Phi^*\right)^{-1}\left(\underbrace{\Delta^\top P\Phi^*}_{=0} + {\Phi^*}^\top P\Delta\right) \\ =& \left({\Phi^*}^\top\Phi^*\right)^{-1}\left({\Phi^*}^\top P\Delta\right). \end{aligned}$$

Therefore, the first term in the second line is dropped, as well as the first term of the final summand.

Note that $F^* = \left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top}P^\pi\Phi^* = \underbrace{\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top}\Phi^*}_{=I}\operatorname{diag}(\Lambda_i) = \operatorname{diag}(\Lambda_i)$, where Λ_i is the set of eigenvalues corresponding to the eigenvectors in Φ^* and $\operatorname{diag}(\Lambda_i)$ is the diagonal matrix of eigenvalues corresponding to those eigenvectors spanned by Φ^*.

This allows us to compute the remaining derivative,

$$\begin{aligned}
&D_\Delta\left(\Phi F^* - P\Phi\right)|_{\Phi=\Phi^*}\operatorname{diag}(\Lambda_i)\\
=&\left(\Delta\operatorname{diag}(\Lambda_i) + \Phi^* D_\Delta F^* - P\Delta\right)\operatorname{diag}(\Lambda_i)\\
=&\left(\Delta\operatorname{diag}(\Lambda_i) + \Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top}P\Delta - P\Delta\right)\operatorname{diag}(\Lambda_i),
\end{aligned}$$

where we use the fact that $D_\Delta P\Phi^* = PD_\Delta\Phi^* = P\Delta$.

Finally, as $\Phi^*F^* = \Phi^*\operatorname{diag}(\Lambda_i) = P^\pi\Phi^*$, we obtain

$$D_\Delta\frac{\mathrm{d}}{\mathrm{d}t}\Phi|_{\Phi=\Phi^*} = -\left(\Delta\operatorname{diag}(\Lambda_i) + \Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top}P\Delta - P\Delta\right)\operatorname{diag}(\Lambda_i) - \underbrace{\left(\Phi^*F^* - P\Phi^*\right)}_{=0\text{ as shown}}\left(\Phi^{*\top}P\Delta\right)^\top.$$

By the definition of the directional derivative as the Jacobian-vector product, we can now assert

$$\left(\frac{\mathrm{d}}{\mathrm{d}\Phi}\frac{\mathrm{d}}{\mathrm{d}t}\Phi|_{\Phi=\Phi^*}\right)\Delta = -\left(\Delta\operatorname{diag}(\Lambda_i) + \left(\Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top} - I\right)P\Delta\right)\operatorname{diag}(\Lambda_i).$$

What remains to be shown is that there exist a direction which corresponds to a positive eigenvalue of the Jacobian of the dynamics.

Choose $\Delta = v_j u^\top$. Let v_j be an eigenvector not in the span of Φ^* but in the top-k eigenvectors. Let λ_j be the corresponding eigenvalue. By our assumption before, there exist at least one eigenvalue $\lambda_i \in \Lambda_i$ so that $\lambda_j > \lambda_i$.

Note that $\left(\Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top} - I\right)Pv_j = \left(\Phi\left(\Phi^{*\top}\Phi^*\right)^{-1}\underbrace{\Phi^{*\top}v_j}_{0\text{ by construction}} - Iv_j\right)\lambda_j = -\lambda_j v_j$ and therefore $\left(\Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top} - I\right)P\Delta = -\Delta\lambda_j I$.

To simplify notation, we will write Λ_i for $\operatorname{diag}(\Lambda_i)$ from now on as there is no risk of confusion.

$$\begin{aligned}
-\left(\Delta\Lambda_i + \left(\Phi^*\left(\Phi^{*\top}\Phi^*\right)^{-1}\Phi^{*\top} - I\right)P\Delta\right)\Lambda_i =& -\Delta\left(\Lambda_i - \lambda_j I\right)\Lambda_i\\
=& -\Delta\left(\Lambda_i^2 - \lambda_j\Lambda_i\right)\\
=&\Delta\left(\lambda_j\Lambda_i - \Lambda_i^2\right).
\end{aligned}$$

We can now choose u so that it is any eigenvector of $\left(\lambda_j\Lambda_i - \Lambda_i^2\right)$. As this is a diagonal matrix, it is easy to see that if $\lambda_j > \lambda_i$ for any λ_i, the matrix will have a positive eigenvalue, meaning there exists a direction in which the critical point is unstable.

□

Case 2: Non-invariant subspace cases: Not all critical points lie in invariant subspaces. One such an alternative critical point is the case of $\Phi^\top P\Phi = 0$, and more generally, for each set of column vectors ϕ_i in Φ, they needs to either be mapped to an invariant or an orthogonal subspace

by P^π to be stable. In the orthogonal case, the Jacobian at the critical point becomes 0, meaning no conclusion about stability can be drawn from this analysis.

We leave further analysis of non invariant subspace critical points open for future work. We do however conjecture that the non invariant subspace critical points are also saddle-points or unstable solutions of the ODE, following the experimental analysis by Tang et al. (2022).

Negative eigenvalues: In case the matrix has negative eigenvalues, the stability conditions in the final step of the proof change. The matrix $\lambda_j \Lambda_i - \Lambda_i^2$ will not have negative eigenvalues if λ_i is negative but λ_j is positive. The ranking of stable points follows this slightly un-intuitive ordering: all negative eigenvalues sorted by absolute value followed by all positive eigenvalues sorted by absolute value.

Proposition 2 (Stationary points of reconstruction)**.** *Assume Assumption 1 and Assumption 2 hold. Write $(u_1, \dots, u_n)$, $(v_1, \dots, v_n)$ for the left and right singular vectors of P^π sorted in descending order by singular value. Any stationary point (Φ^*, F^*, Ψ^*) of L_{rec} under the two timescale scenario satisfies* $\operatorname{span}(\Phi^*) = \operatorname{span}(\{u_1, \dots, u_k\})$, $\operatorname{span}({\Psi^*}^\top) = \operatorname{span}(\{v_1, \dots, v_k\})$.

Proof. We first show that under the two timescale scenario, F is stationary and therefore does not change the span of the critical points.

Due to the assumption of the two-timescale scenario, we compute Ψ^* by solving the linear regression problem,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\Psi^*}\|\Phi F \Psi^* - P\|_F^2 &= F^\top \Phi^\top (\Phi F \Psi^* - P) \\
0 &= F^\top \Phi^\top (\Phi F \Psi^* - P) \\
\Leftrightarrow \qquad B^* &= \left(F^\top \Phi^\top \Phi F\right)^{-1} F^\top \Phi^\top P = F^{-1} \left(\Phi^\top \Phi\right)^{-1} \Phi^\top P.
\end{aligned}$$

Plugging this solution back into the original equation,

$$\begin{aligned}
\|\Phi F \Psi^* - P\|_F^2 &= \|\Phi F F^{-1} \left(\Phi^\top \Phi\right)^{-1} \Phi^\top P - P\|_F^2 \\
&= \|\Phi \left(\Phi^\top \Phi\right)^{-1} \Phi^\top P - P\|_F^2,
\end{aligned}$$

we notice that F cancels. Therefore, the optimality conditions for A follow from the Eckart-Young theorem, as presented in Lemma 3 □

Proposition 3. *Assume Assumption 1, Assumption 2, and Assumption 3 hold. Let $\{\Phi^*_{\text{lat/td}}\}$ be the set of critical points of L_{lat} or L_{td} respectively. Then $\mathcal{O}^{-1}\Phi^*_{\text{lat/td}}$ are stationary points for the reparameterized losses $L^{\mathcal{O}}_{\text{lat}}$ and $L^{\mathcal{O}}_{\text{td}}$.*[7] *Furthermore, if $\Phi^*_{\text{lat/td}}$ is an asymptotically stable point of $L_{\text{lat/td}}$ that has a Jacobian with all negative eigenvalues, $\mathcal{O}^{-1}\Phi^*_{\text{lat/td}}$ is an asymptotically stable point of $L^{\mathcal{O}}_{\text{lat/td}}$.*

Proof. We first write out all losses with the observation matrix $\mathcal{O}$. The reward function is assumed to not change under the introduction of $\mathcal{O}$, therefore we do not multiply $\mathcal{O}$ to $x^\top r$.

$$\begin{aligned}
L_{\text{rec}}(\Phi, F, \Psi) =& \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \mathcal{O}\Phi F \Psi - x^\top P^\pi \mathcal{O}\right\|_2^2\right] = \|\mathcal{O}\Phi F \Psi - P^\pi \mathcal{O}\|_2^2, \\
L_{\text{lat}}(\Phi, F) =& \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \mathcal{O}\Phi F - \left[x^\top P^\pi \mathcal{O}\Phi\right]_{\text{sg}}\right\|_2^2\right] = \left\|\mathcal{O}\Phi F - \left[P^\pi \mathcal{O}\Phi\right]_{\text{sg}}\right\|_2^2 \\
L_{\text{td}}(\Phi, F) =& \mathbb{E}_{x\sim\mathcal{D}}\left[\left\|x^\top \mathcal{O}\Phi \hat{V} - \left[x^\top \left(r + \gamma P^\pi \mathcal{O}\Phi \hat{V}\right)\right]_{\text{sg}}\right\|_2^2\right] = \left\|\mathcal{O}\Phi \hat{V} - \left[\left(r + \gamma P^\pi \mathcal{O}\Phi \hat{V}\right)\right]_{\text{sg}}\right\|_2^2.
\end{aligned}$$

[7] Due to space constraints, we present the full equations in the proof.

Note that in the cases of L_{lat} and L_{TD}, all occurrences of $\mathcal{O}$ are multiplied by Φ. Therefore the corresponding gradient flows are reparameterizations as defined in Lemma 6, and the proof follows directly. □

Proposition 4. *Assume Assumption 1, Assumption 2, and Assumption 3 hold. Let* $(u_1, \ldots, u_n)$, $(v_1, \ldots, v_n)$ *be the left and right singular vectors of* $\mathcal{O}^{-1}P^\pi\mathcal{O}$. *Any stationary point* (Φ^*, F^*, Ψ^*) *of* $L^{\mathcal{O}}_{rec}$ *satisfies* $\text{span}(\Phi^*) = \text{span}(\{u_1, \ldots, u_k\})$, $\text{span}({\Psi^*}^\top) = \text{span}(\{v_1, \ldots, v_k\})$.

Proof. As before, note that $L^{\mathcal{O}}_{\text{rec}}$ is of the form $\|\mathcal{O}AXB - P\mathcal{O}\|_F^2$, with $A \in \mathbb{R}^{n\times k}$, $X \in \mathbb{R}^{k\times k}$, and $B \in \mathbb{R}^{k\times n}$.

Due to the assumption of the two-timescale scenario, we compute Ψ^* by solving the linear regression problem,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\psi}\|\mathcal{O}\Phi F\Psi - P\mathcal{O}\|_F^2 &= F^\top\Phi^\top\mathcal{O}^\top(\mathcal{O}\Phi F\Psi - P) \\
0 &= F^\top\Phi^\top\mathcal{O}^\top(\mathcal{O}\Phi F\Psi^* - P) = 0 \\
\Leftrightarrow \qquad \Psi^* &= \left(F^\top\Phi^\top\mathcal{O}^\top\mathcal{O}\Phi F\right)^{-1} F^\top\Phi^\top\mathcal{O}^\top P.
\end{aligned}$$

Substituting into $L^{\mathcal{O}}_{\text{rec}}$, we obtain

$$\begin{aligned}
\|\mathcal{O}\Phi F\Psi - P\mathcal{O}\|_F^2 &= \|\mathcal{O}\Phi F\left(F^\top\Phi^\top\mathcal{O}^\top\mathcal{O}\Phi F\right)^{-1} F^\top\Phi^\top\mathcal{O}^\top P - P\mathcal{O}\|_F^2 \\
&= \|\mathcal{O}\Phi\left(\Phi^\top\mathcal{O}^\top\mathcal{O}\Phi\right)^{-1}\Phi^\top\mathcal{O}^\top P - P\mathcal{O}\|_F^2,
\end{aligned}$$

which again implies that F is stationary.

We note that $\|\mathcal{O}\Phi\Psi - P\mathcal{O}\|_F^2$ is the reduced rank regression problem solved in Lemma 3 which solution is given by the top-k left and right singular vectors of $\mathcal{O}^{-1}P\mathcal{O}$. □

Proofs for Section 5

Proposition 5. *Suppose that* P^π *is real diagonalizable, and that Assumption 1, Assumption 3, and Assumption 4 hold. There exists a non-trivial critical point* Φ^* *of the two-timescale TD loss* L_{td} *such that* $\text{span}(r^\pi) \subseteq \text{span}(\Phi^*)$. *Furthermore,* Φ^* *is a critical point of the two-timescale joint loss* L_{td+lat}. *Therefore combining* L_{TD} *and* L_{Lat} *does not exclude the existence of a stationary point with* 0 *value function approximation error.*

Proof. By Assumption 4 there exists a set of k vectors $\phi_1, \ldots, \phi_k$ such that $r^\pi \in \text{span}(\phi_1, \ldots, \phi_k)$ and $\phi_1, \ldots, \phi_k$ span an invariant subspace of P^π. We can choose this set of vectors to orthonormal, e.g. by applying the Gram Schmidt procedure to the eigenvectors $(w_{i1}, \ldots w_{im})$. By Lemma 7 the matrix $\Phi \in \mathbb{R}^{n\times k}$ whose columns are $\phi_1, \ldots, \phi_k$ is a critical point for L_{lat} and spans an invariant subspace of P^π. In addition, by Lemma 5, Φ forms a complete basis for the value function V^π. By Lemma 13, Φ is also a critical point of L_{TD}, with 0 value function approximation error. As Φ is a critical point for both L_{TD} and L_{Lat}, it is a critical point of $L_{\text{TD}} + L_{\text{Lat}}$.

□

Proposition 6. *Let Assumption 1, Assumption 2, and Assumption 4 hold. If the reward spanning eigenvectors do not lie within the span of the top-k singular vectors,* $\text{span}(w_{i_1}, \ldots, w_{i_m}) \not\subseteq \text{span}(u_1, \ldots, u_k)$, *the critical points of the two-timescale joint loss* L_{td+rec} *are guaranteed to not be minimizers of the value function approximation error.*

Proof. Following from Assumption 4 and Lemma 5, we have that any critical point Φ^* of L_{td} with perfect value function approximation fulfils $\text{span}(r^\pi) \subseteq \text{span}(\Phi^*)$, as without this condition, Φ would not have all necessary basis vectors to represent V^π. Under our low-dimensionality assumption

Parameter	MinAtar	DMC
Initial steps (Random policy)	5000	5000
Env steps per update step	4	1
Batch size	512	512
Exploration ϵ	0.05	0.01
RL learning rate	0.0003	0.0003
Model/decoder learning rate	0.0003	0.0003
Encoder learning rate	0.0001	0.0001
Target network update interval	1000	n/a
Soft update τ	n/a	0.995
Discount factor γ	0.99	0.99
Model forward prediction steps	4	4

Table 1: Hyper-parameters the RL experiments

$r^\pi \in \text{span}(w_{i_1}, \ldots, w_{i_m})$, this condition implies that $\text{span}(w_{i_1}, \ldots, w_{i_m}) \subseteq \text{span}(\Phi^*)$. From Proposition 2 we know that any critical point Φ^* of L_{rec} satisfies $\text{span}(\Phi^*) = \text{span}(\{u_1, \ldots, u_k\})$, where $u_1, \ldots, u_k$ are the top k left singular vectors of P^π. Under the assumption that $\text{span}(w_{i_1}, \ldots, w_{i_m}) \not\subseteq \text{span}(u_1, \ldots, u_k)$ these conditions cannot happen simultaneously, and hence no critical point of the joint loss achieves perfect value function reconstruction. □

Proposition 7 (Suboptimality of top k eigenspaces with distractions)**.** *Assume an MDP with distraction composed of two independent processes $\mathcal{M}$ and $\mathcal{N}$ according to Definition 1. Let $v_1, \ldots, v_n$ and $u_1, \ldots, u_m$ be the eigenvectors of $\mathcal{M}$ and $\mathcal{N}$ respectively, with associated real eigenvalues λ_j and μ_i ordered. Assume $\forall i < k : \mu_i > \lambda_2$ and $r_\mathcal{N} = 0$. Let U_k be an orthonormal basis for the top-k eigenspace of $\mathcal{M} \otimes \mathcal{N}$. Then* $\text{span}(U_k) = \text{span}(\{\mathbf{1} \otimes v_i | \forall i \leq k\})$ *and* $\text{proj}_{U_k}(r_\mathcal{M} \otimes \mathbf{1}) = (\sum_{i=0}^m r_i/n)\mathbf{1}_{n \cdot m}$.

Proof. We first note that $\text{span}(V_k) = \text{span}(\{v_i \otimes u_1 | \forall i \leq k\})$ follows directly from Lemma 1 and Lemma 2. The eigenvector $u_1 = \mathbf{1}$ as M is a stochastic matrix.

As V_k is an orthogonal basis, write the projection operation as

$$\begin{aligned} V_k V_k^\top (r_\mathcal{M} \otimes \mathbf{1}) &= \frac{1}{m} (\mathbf{1} \otimes \text{orth}(V)) \left(\mathbf{1} \otimes \text{orth}(V)^\top\right) (r_\mathcal{M} \otimes \mathbf{1}) \\ &= \frac{1}{m} \left((\mathbf{1} \otimes \mathbf{1}) \otimes (\text{orth}(V)\,\text{orth}(V)^\top)\right) (r_\mathcal{M} \otimes \mathbf{1}) \\ &= \frac{1}{m} (\mathbf{1} \otimes \mathbf{1}) r_\mathcal{M} \otimes (\text{orth}(V)\,\text{orth}(V)^\top) \mathbf{1} \\ &= \sum_{i=1}^m \frac{r_i}{m} \mathbf{1}. \end{aligned}$$

□

E Implementation details for the experiments

For the Minatar experiments, we use a simple Double DQN architecture (Van Hasselt et al., 2016). We find that our implementation performs roughly on par with those reported by Young & Tian (2019), however we changed the network architecture slightly to allow a clear "encoder" and "prediction head" split.

We implement the latent self-prediction loss using a periodically updated copy of the encoder network. This hard update of the encoder target is synchronized with the Q target network update.

Our networks are parameterized as presented in Table 2

Encoder Φ	ConvLayer	channels=16, kernel=$(3,3)$, padding=0, stride=$(1,1)$
	ELU activation	–
	Dense Layer	out_size=100
	ELU activation	–
Latent Model F	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=100
Decoder Ψ	Dense Layer	out_size=$10 \cdot 10 \cdot 16$
	ELU activation	–
	ConvTranspose Layer	kernel=$(3,3)$, padding=1, stride=$(1,1)$
Q head $\hat{V}$	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=action_space

Table 2: Network architecture for the MinAtar experiments.

Encoder Φ	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=256
	ELU activation	–
Latent Model F	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=100
Decoder Ψ	Dense Layer	out_size=$10 \cdot 10 \cdot 16$
	ELU activation	–
	Dense Layer	out_size=obs_dim
Q head $\hat{V}$	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=1
Actor head	Dense Layer	out_size=256
	ELU activation	–
	Dense Layer	out_size=action_dim

Table 3: Network architecture for the DMC experiments.

Relevant hyper-parameters are shown in Table 1.

The random noise matrix is sampled from a Bernoulli distribution with $p(1) = 0.1$. The distortion matrix was created using a random matrix of size $10 \cdot 10 \cdot \text{channels} \times 10 \cdot 10 \cdot \text{channels}$ with entries independently sampled from a Bernoulli distribution with $p(1) = 0.2$. This was chosen to have a similar sparsity in the distraction as in the main observation channels. Additionally, we verified that the matrix was invertible and $\|\mathcal{O}\|_1 < 255$ to ensure that the replay buffer implementation would not overflow.

The DMC results were obtained using the TD3 algorithm (Fujimoto et al., 2018) with constant Gaussian noise with standard deviation of 0.01 added for exploration following Yarats et al. (2021). In addition to the Q function network, TD3 also requires an actor network. We do not propagate any gradients from the actor network into the encoder, following standard practice in actor-critic learning.

Networks for the DMC implementation are shown in Table 3.

In the DMC experiments, we used isotropic Gaussian noise for the distracting noise, and a copy of the *humanoid* environment as the distraction.

Full code is available at https://github.com/adaptive-agents-lab/understading_auxilliary_tasks.

F DMC results

The DMC environments have a very different observation structure and transition dynamics compared to those of the MinAtar games. The observation spaces are dense and often contain topological discontinuities such as those outlined in Subsection A.1.

All experiments are repeated across 10 seeds.

We find slightly different results in these environments compared to MinAtar, especially about the efficacy of the latent self-prediction and observation reconstructions in the stand-alone setting (Figure 8 and Figure 9). Overall, latent self-prediction performs much more strongly in these environments compared to the MinAtar experiments, especially when using it as a stand-alone loss. Curiously, in the only environment where observation prediction shines (quadruped-walk), stand-alone latent self-prediciton and reward prediction seems to outperform all other test settings. This highlights the second important difference between the MinAtar games and the DMC suite: dense rewards. We conjecture that most of the differences between observation prediction and latent self-prediction in the DMC suite comes from dense rewards and different topological continuity, but a precise investigation is needed in future work.

Finally, our assumptions about the impact of noise distortions on the efficacy of different loss functions seems to be much more clearly apparent in DMC than in MinAtar. This suggests that the differing primary observation spaces also change how the learning process interacts with the noise, i.e. because the spectral properties of the underlying environments might differ. Especially in the challenging distraction setting of structured noise, non trivial policies (as measured by a strong performance improvement over a random baseline) can only be observed in 7 out of the 15 environemnts. Of these 7, latent self prediction seems to outperform the other baselines in 3 cases, only falling behind in 1. However, we observed that the pendulum environment has a strong bimodal distribution with some runs completely failing to perform for each experimental scenario, so the number of seeds might be insufficient to disambiguate performance here, as evidenced by the large confidence interval at the 95% level.

Overall, 10 seeds, even though widely used in the community, might not be sufficient to fully represent the performance in the DMC suite, suggesting that more work is needed on the inherent variations of these environments. We do not present aggregate performance over the environments, as the different reward scales makes a comparison prone to be driven by outliers such as the humanoid environments, in which several algorithms fail to learn at all at 10 seeds.

G Limitations

While our paper aims to minimize the theory-practice gap with careful experimentation, we nonetheless need to make several assumptions that are both limitations and also potential for additional analysis in future work. As our aim is to make theory useful and accessible for practitioners, we aim to be very open and clear about our limitations here.

Limitations of the analytical framework: From previous work (Tang et al., 2022; Le Lan et al., 2023) we inherit the limitation of studying deterministic models in potentially stochastic environments. This is a necessary limitation, as considering the stochastic equivalents of e.g. the observation prediction loss would render the model and their gradients non-linear due to the introduction of a softmax or similar constraint. As other works carry the same limitation, we believe that this does not render our work inapplicable, but it does suggest the need for more powerful mathematical tools in future work.

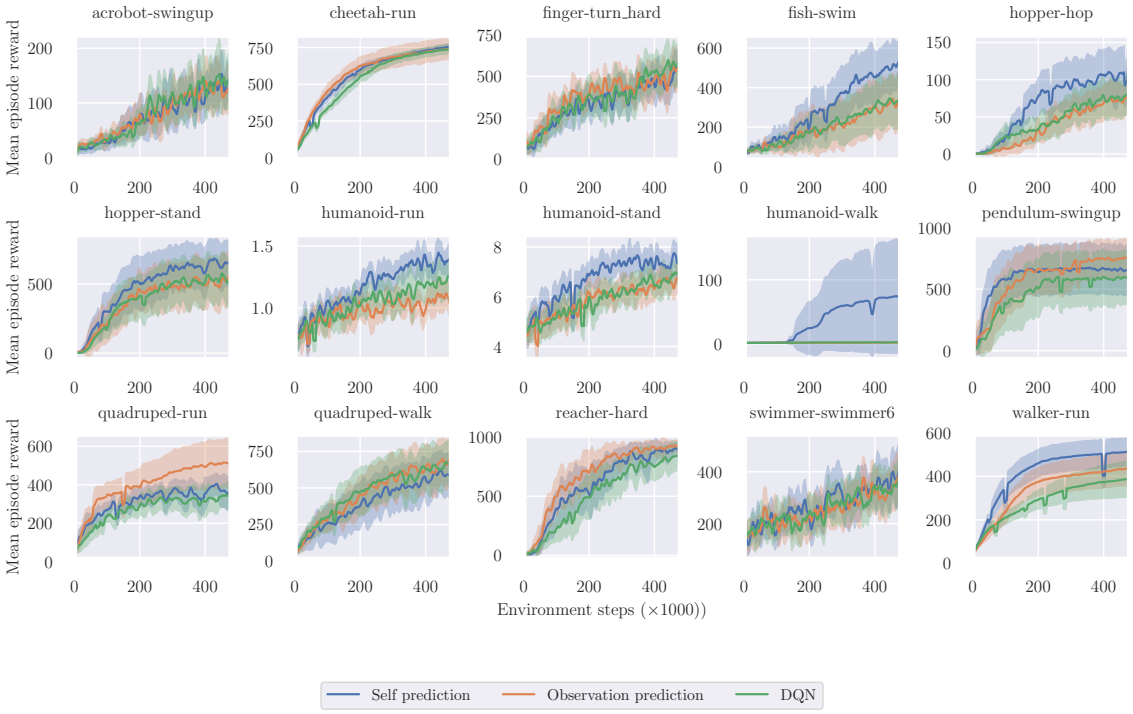

Figure 7: DMC: Auxiliary task scenario

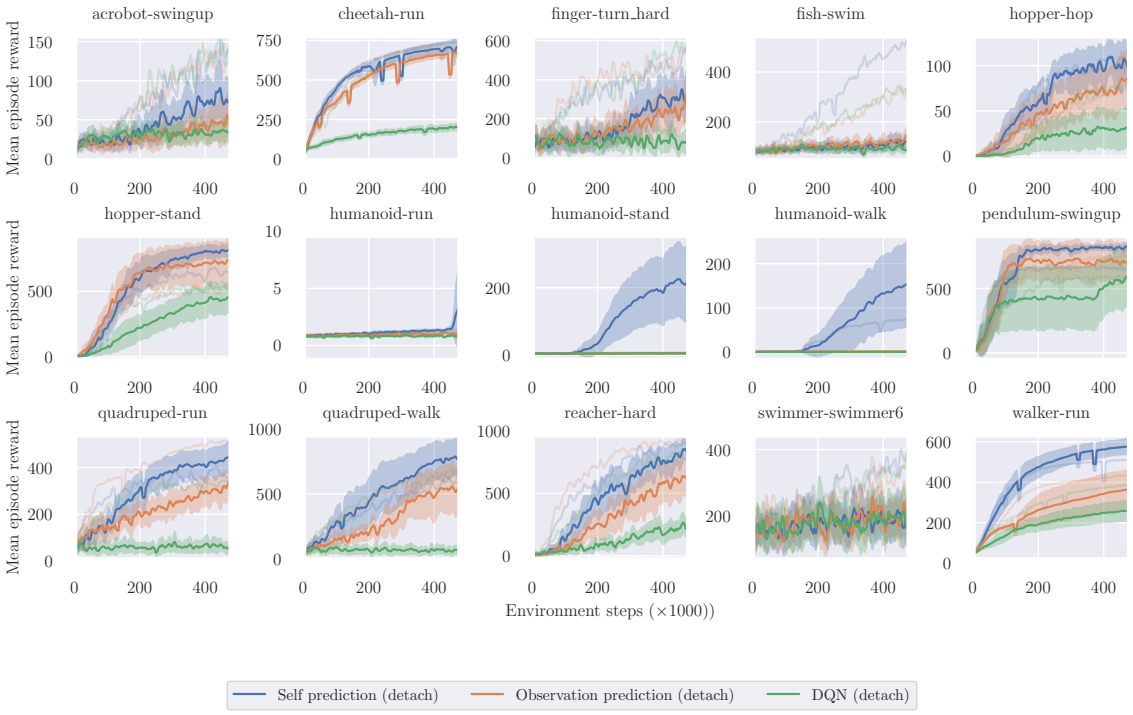

Figure 8: DMC: Stand alone scenario

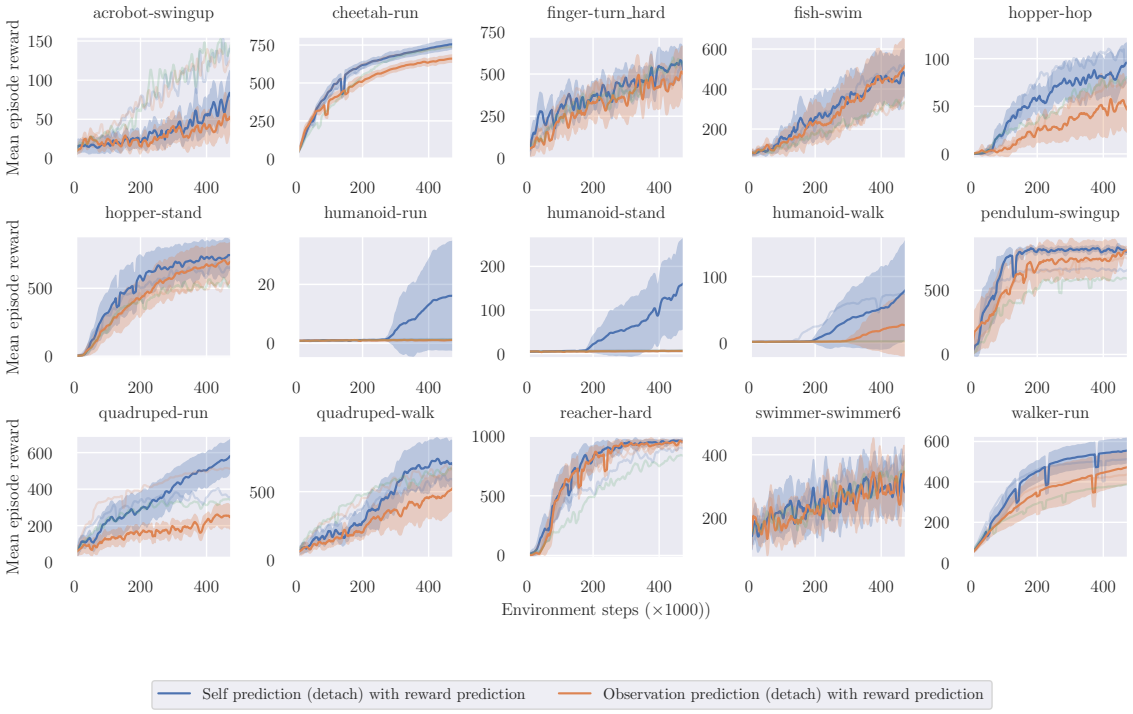

Figure 9: DMC: Auxiliary loss + reward prediction, not TD

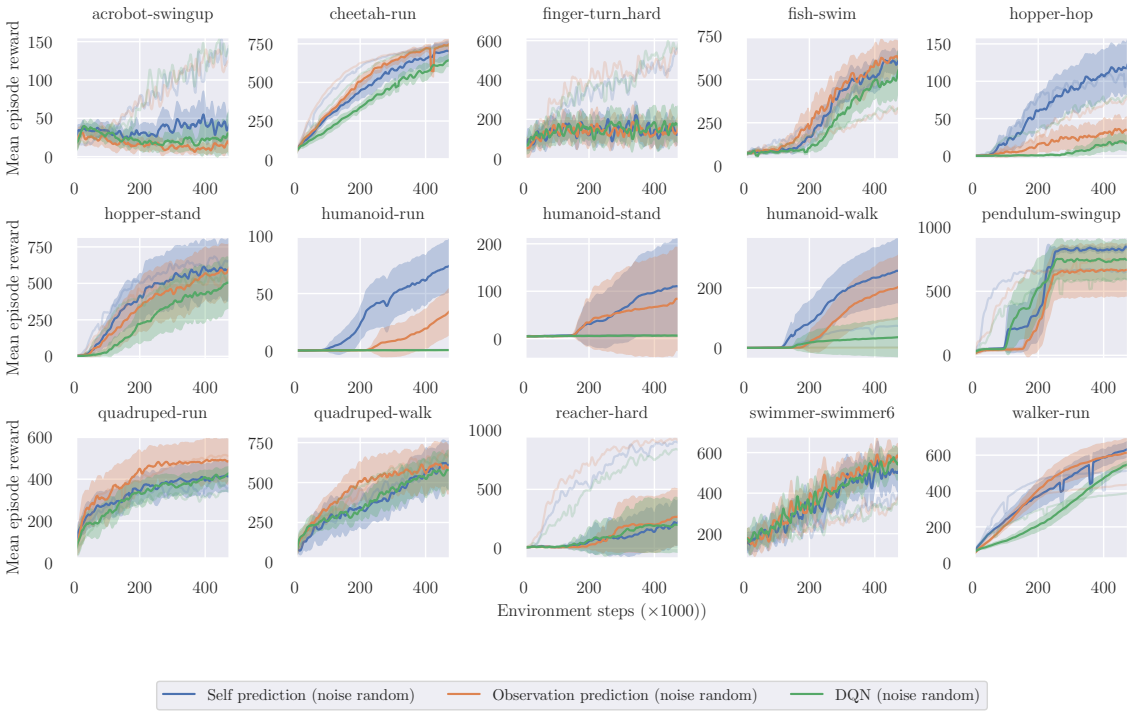

Figure 10: DMC: Random noise distraction

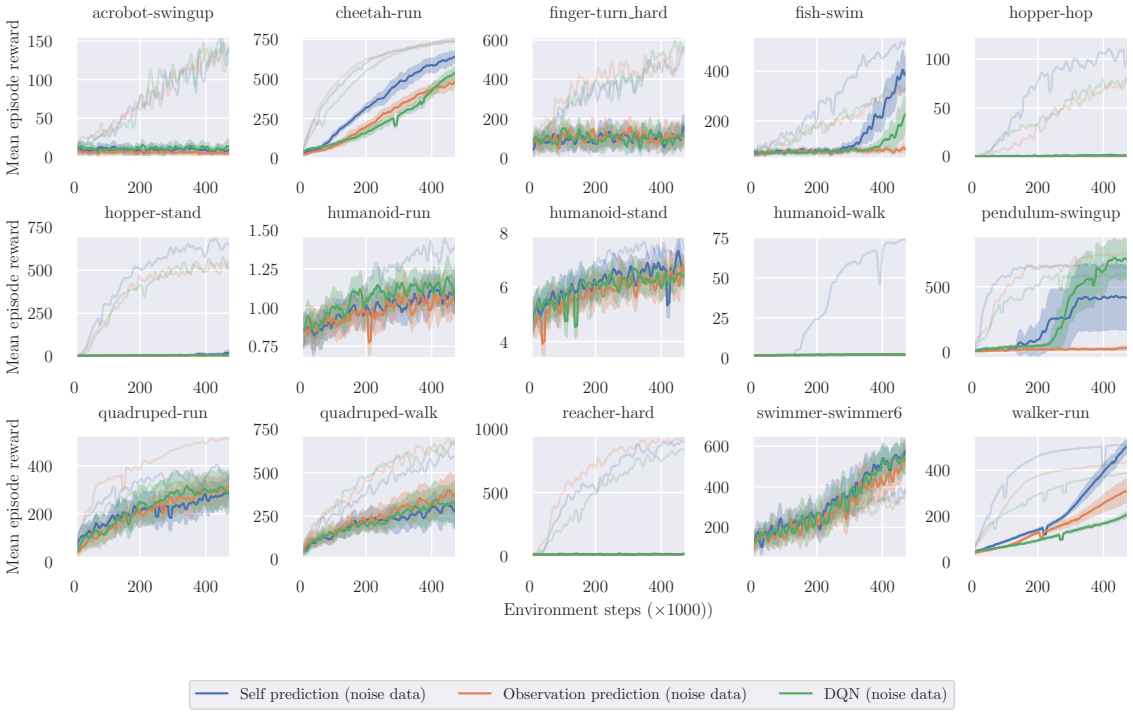

Figure 11: DMC: Structured distraction

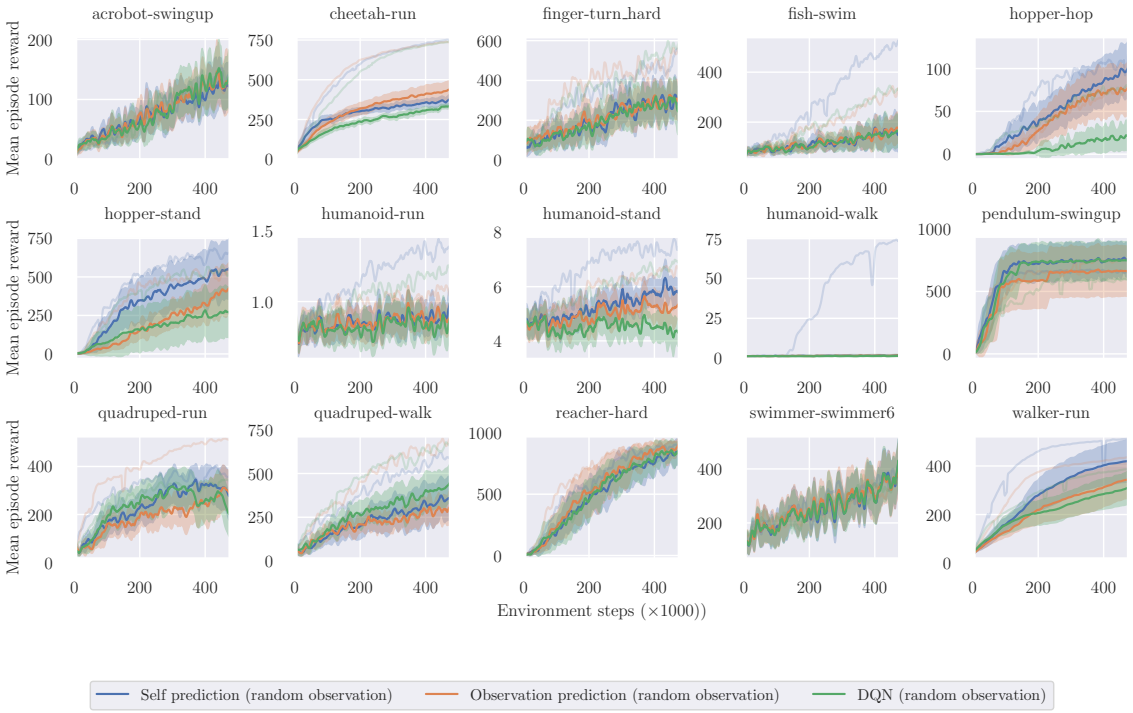

Figure 12: DMC: Observation space distortion

We also conduct all of our theoretical work in the on-policy policy evaluation regime, while our empirical study includes both off-policy policy estimation and policy improvement. Again, this is a limitation inherited from all related work. As our theoretical predictions are still validated, we consider this an acceptable limitation, but studying the impact of off-policy samples and shifting policies is an important step for future work.

Limitations of the formalism: Our notion of observation distortion requires unnaturally large observation spaces. Again, this stems from our adherence to a linear framework. While the observation space of e.g. the MinAtar games is relatively large, depending on the game a 300-1000 dimensional vector, this is still substantially smaller than the total number of states. Studying these kinds of nonlinear image transformations in more detail would require deviating farther from the previous literature.

In this paper, we aim to (re-)introduce the notion of distraction into the learning dynamics literature, and so we use a relatively simplistic notion of distraction. Going beyond the independence assumption in the distraction model (due to the Kronecker formulation) and analysing more complex forms of distractions e.g. processes in which the reward-relevant process causally influences the distracting process but not vice versa is an exciting direction for future work.