How JEPA Avoids Noisy Features: The Implicit Bias of Deep Linear Self Distillation Networks

Etai Littwin **Omid Saremi** **Madhu Advani**
Vimal Thilak Preetum Nakkiran Chen Huang Joshua Susskind

Apple

Abstract

Two competing paradigms exist for self-supervised learning of data representations. *Joint Embedding Predictive Architecture* (JEPA) is a class of architectures in which semantically similar inputs are encoded into representations that are predictive of each other. A recent successful approach that falls under the JEPA framework is self-distillation, where an online encoder is trained to predict the output of the target encoder, sometimes using a lightweight predictor network. This is contrasted with the *Masked AutoEncoder* (MAE) paradigm, where an encoder and decoder are trained to reconstruct missing parts of the input in the data space rather, than its latent representation. A common motivation for using the JEPA approach over MAE is that the JEPA objective prioritizes abstract features over fine-grained pixel information (which can be unpredictable and uninformative). In this work, we seek to understand the mechanism behind this empirical observation by analyzing the training dynamics of deep linear models. We uncover a surprising mechanism: in a simplified linear setting where both approaches learn similar representations, JEPAs are biased to learn high-influence features, i.e., features characterized by having high regression coefficients. Our results point to a distinct implicit bias of predicting in latent space that may shed light on its success in practice.

1 Introduction

Representation learning has arguably been one of the most promising prospects and longest-standing goals of deep learning. In simple terms, representation learning refers to the process of automatically discovering and extracting meaningful features or representations from raw data. The learned representations can then be used to solve various downstream tasks, such as classification, and regression, or as general-purpose representations in robotics and embodied agents. In recent years, the practice of representation learning has witnessed remarkable strides, particularly in the domain of Self-Supervised Learning (SSL) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The SSL paradigm has garnered significant attention due to its ability to exploit vast amounts of unlabeled data, freeing practitioners from the burden of expensive annotation efforts. While many SSL approaches have been proposed in the literature, two paradigms have emerged as particularly successful, driving the bulk of the recent breakthroughs in the field:

Masked AutoEncoders (MAE) The original MAE [8] and its successive variants (e.g. [13, 14]) introduced a training objective that seeks to reconstruct missing pieces of data from its partially masked input. This approach, while simple, has proven efficient and scalable [15]. Formally, the MAE objective uses an encoder function $f_W(X) : \mathbb{R}^d \to \mathbb{R}^d$ and a decoder function $g_V(f) : \mathbb{R}^d \to \mathbb{R}^d$ to predict a target y given input x:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{2}\mathbb{E}_{x,y\sim p(x,y)}\|g_V(f_W(x)) - y\|^2, \tag{1}$$

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

where we define a joint distribution over inputs and targets $p = p(x, y)$ with $x, y \in \mathbb{R}^d$. Here we assume tied dimensions for simplicity.

In general, we can think of x and y as different inputs sharing the same underlying semantic information. One common practice is to let the input x correspond to a partially masked input, and the target y corresponds to the masked-out portions of the input. Note that a critical design choice of the MAE objective is that the loss penalizes the reconstruction error directly in the target's input space. This implies that, as perhaps some parts of the target y cannot be predicted given x, the quality of the encoding f_W crucially depends on how x, y are sampled.

Joint Embedding Predictive Architectures (JEPA) The JEPA family of models does not attempt to predict y directly and instead opts to predict the latent representation of y given by some encoding function. When the encoder function for x and y is shared, a simple JEPA objective is given by:

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{2}\mathbb{E}_{x,y\sim p(x,y)}\|g_V(f_W(x)) - f_W(y)\|^2. \tag{2}$$

Unlike the objective in (1), the objective given by (2) is susceptible to collapse: it can be minimized trivially by employing an encoder and decoder that map all inputs to the same vector. Contrastive methods [16, 6, 17] avoid this issue by including an additional loss term that pushes apart representations for negative pairs (two different samples). The drawback to this approach is that it can require comparing a sample to many negative examples to work effectively. This has led to a rise in the popularity of non-contrastive SSL. With non-contrastive methods, it becomes imperative to implicitly or explicitly bias the representation mapping $f_W(x)$ to avoid its collapse to a trivial solution (e.g., $f_W(x) = 0,\ \forall x$). In this paper, we leverage the widely adopted [11] *StopGrad* operator, preventing the flow of gradients through the $f_W(y)$ branch. Prominent examples of this foundational architecture include BYOL [9], data2vec [10], SimSiam [11] and the more recent I-JEPA [18] and V-JEPA [19] models. Consequently, the JEPA models under consideration fall under non-contrastive self-distillation methods. We will refer to such methods as JEPA for the remainder of this paper.

While both paradigms have found practical use, characterizing the implicit bias of each remains a major research question. Recently, it was shown in [20] that in the vision domain, perceptual features mostly reside in the data subspace which explains a relatively small portion of the observed variance. This suggests that reconstruction losses are sub-optimal for perception tasks since they tend to prioritize learning subspaces that explain the bulk of the input variance. On the other hand, the JEPA approach appears to be more suitable for perception tasks: empirically, JEPA often achieves better downstream classification performance with fewer optimization steps [3, 19].

In this work, we seek to uncover the mechanisms behind these observations by analyzing tractable deep linear neural networks. To that end, we introduce a fully solvable toy model admitting a complete characterization of its training dynamics for both objectives. Our analysis is inspired by previous work on the greedy learning dynamics of gradient descent, and SSL in particular [21, 22, 23]. However, going beyond previous work, here we present a more nuanced analysis showing how the greedy nature of learning is affected by the choice of objective, model, and the properties of the data. Our analysis reveals a distinct implicit bias of JEPAs: the propensity of the JEPA objective towards learning influential features, defined as features with a large regression coefficient, a property for which the MAE objective is mostly agnostic. Moreover, we show that this distinct bias takes hold only when the encoder is deep and diminishes substantially when the encoder is shallow. This allows JEPAs to steer away from the top subspace of the data which explains its observed variance, and focus on the more semantically rich subspace. Our work therefore uncovers a fundamental implicit bias, shedding light on both the advantages and drawbacks of predicting in latent space.

To summarize, our specific contributions are:

1. We analytically solve the learning dynamics of JEPA and MAE, in the tractable theoretical setting of deep linear networks.
2. In this setting, we prove a significant difference between JEPA and MAE: informally, JEPA prioritizes "influential features" (features which are most informative in prediction), whereas MAE only prioritizes highly covarying features (even if they are noisy and thus less informative).
3. We show that our theoretical setting is rich enough to capture two more realistic generative processes, based on static and dynamic linear models.

Our results help clarify the precise advantages of joint-embedding architectures, and complement the recent results of [20] on limitations of reconstruction-based methods.

Preliminaries

In this paper, we consider linear over-parameterized encoders and a shallow linear decoder. This tractable setting allows the non-linear training dynamics [24, 25, 26, 27] of the deep linear encoder network to be solved exactly under a certain set of assumptions. Let $\{W^a\}_{a=1\dots L} \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{d\times d}$ denote the set of weights for a linear encoder $f(x)$ of depth L, and that of a linear decoder $g(x)$ respectively, and consider the linear parameterization

$$\bar{W} = \prod_{a=1}^{L} W^a, \quad f(x) = \bar{W}x, \quad g(f(x)) = Vf(x), \tag{3}$$

where $\prod_{a=1}^{L} W^a$ indicates the conventional right-to-left product of matrices. As for the input data distribution, we will take x, y to be samples from a centered distribution with covariances:

$$\mathbb{E}[xx^\top] = \Sigma^{xx}, \quad \mathbb{E}[yy^\top] = \Sigma^{yy}, \quad \mathbb{E}[xy^\top] = \Sigma^{xy}. \tag{4}$$

For the purpose of deriving learning dynamics, we will later assume that $\Sigma^{xx}, \Sigma^{xy}, \Sigma^{yy} \in \mathbb{R}^{d\times d}$ are diagonal. Let SG($\bullet$) denote the StopGrad operator applied on $\bullet$. Thus, the JEPA and MAE objectives become

$$\mathcal{L}_{\text{jepa}} = \frac{1}{2}\mathbb{E}_{x,y\sim p(x,y)}\|V\bar{W}x - \text{SG}(\bar{W})y\|^2, \quad \mathcal{L}_{\text{mae}} = \frac{1}{2}\mathbb{E}_{x,y\sim p(x,y)}\|V\bar{W}x - y\|^2. \tag{5}$$

We will denote the diagonal elements of the diagonalized covariance and correlations matrix by $\sigma_i^2 = \Sigma_{ii}^{xx}$ and $\lambda_i = \Sigma_{ii}^{yx}$ respectively, and let $\rho_i = \frac{\lambda_i}{\sigma_i^2}$ denote the regression coefficients.

To motivate our theoretical analysis, we next show empirically that depending on the particular values of $\{\lambda_i\}$ and $\{\rho_i\}$, optimizing the objectives in eq. (5) via gradient descent may learn different encoders entirely.

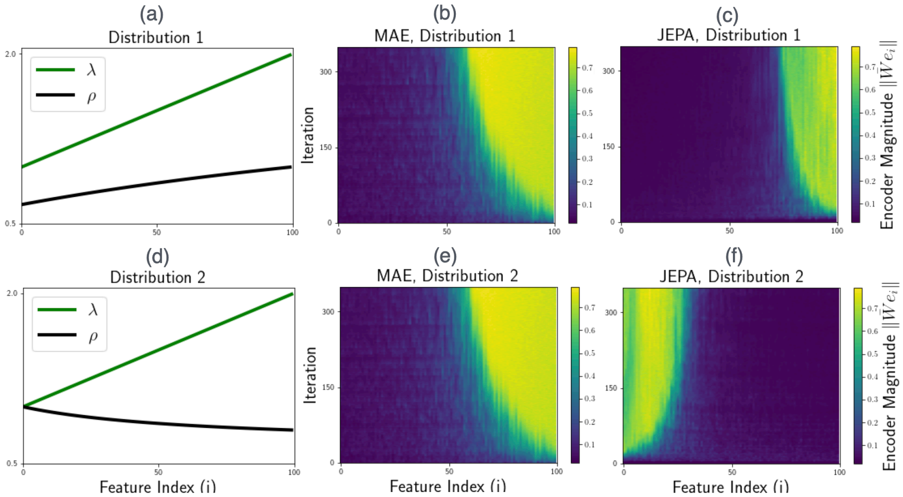

Figure 1: **Deep linear model trained using the MAE and JEPA objectives eq. (5).** Features indices (x-axis) are organized s.t. λ_i is monotonically increasing and ρ_i is (a) - (c) monotonically increasing or (d) - (f) monotonically decreasing. (b),(c): Both objectives learn features in the same order given distribution 1. In (e),(f) the MAE objective maintains the same learning order as in (b) on distribution 2, however, the JEPA objective reverses the learning order, due to sensitivity to ρ_i.

2 The Implicit Bias of JEPA

Before presenting the theoretical results, we describe a simple experiment using the setup in section 1 which will flesh out a surprising key difference between the JEPA and MAE objectives in eq. (5).

First, note that each coordinate x_i of the input data represents an independent feature characterized by the covariance and regression coefficient λ_i, ρ_i, which we refer to as *feature parameters*. Let $e_i \in \mathbb{R}^d$ denote the i-th standard basis vector. We say that an encoder $\bar{W}$ has learned the i-th feature if its projection on e_i, namely $\|\bar{W}e_i\|$ is large. Intuitively, one should expect that if ρ_i is small, $\|\bar{W}e_i\|$ should be small post training as well. Note that this should not be overly surprising since, at least in the case of the MAE objective, the global minimizer is achieved with the regression coefficients $V\bar{W} = \Sigma^{yx}(\Sigma^{xx})^{-1}$. In this investigation, however, we seek an understanding of the dynamics of training, by studying how the MAE and JEPA objectives prioritize different features in training, based on their feature parameters.

To this end, we track how the projections $\|\bar{W}e_i\|$ evolve during training for each component i by training models on Gaussian data with feature parameters λ_i, ρ_i according to the following two distributions (see Figure 1(a) and (d) for an illustration):

1. **Distribution 1**: Set both $\{\lambda_i\}, \{\rho_i\}$ to be monotonically increasing (i.e $\forall_{i>j},\ \lambda_i > \lambda_j, \rho_i > \rho_j$).
2. **Distribution 2**: Set $\{\lambda_i\}$ to be monotonically increasing, and $\{\rho_i\}$ to be monotonically decreasing.

For each distribution, we initialize all the weights with a Gaussian initialization and train the models using both the MAE and JEPA objectives using gradient descent on batches of randomly sampled data. We illustrate the dynamics of training by measuring the magnitude of the encoder components $\|\bar{W}e_i\|$ during training in Figure 1. On the first distribution, we observe clear greedy learning dynamics for both objectives, where components with a higher λ_i are learned earlier on in training. This behavior in greedy settings is in line with the findings in [21] which showed the same behavior in similar settings. However, on the second distribution, we observe a distinction: while the MAE objective retains its order of learning as in distribution 1, the JEPA objective exhibits a complete inversion of the learning order where features with a higher ρ_i, rather than a high λ_i are learned first. This implies that in practical scenarios with a finite training budget, JEPA and MAE objectives can potentially learn entirely different features with little to no overlap, as illustrated in Figure 1. To better understand these phenomena, we turn to a dynamical analysis of a simple model that fully recovers our observations.

3 Dynamical Analysis

In the following theoretical analysis, we train the corresponding models employing objectives eq. (5) by optimizing all weights simultaneously starting at some initial weights configuration $\{W^a(0)\}_{a=1...L}$ and $V(0)$, using the gradient flow equations

$$\forall_a,\ \ \dot{W}^a = -\nabla_{W^a}\mathcal{L},\ \ \dot{V} = -\nabla_V\mathcal{L}, \tag{6}$$

where $\mathcal{L}$ is either JEPA or MAE objective. Generically, the two systems of ODEs are intractable without any simplifying assumptions. A special class of initializations discussed below permits analytical treatment of the problem, allowing us to derive a depth separation result indicative of differences in the inductive bias of the two objectives. In the following, we describe our set of assumptions on the encoder and decoder weights at initialization, and data distribution formally:

Assumption 3.1.

1. Σ^{xx}, Σ^{yx} are diagonal and positive definite.
2. All weight matrices $W^a(0), a = 1...L, V(0)$ are initialized as $W^a = \epsilon^{\frac{1}{L}}U^a, V(0) = \epsilon^{\frac{1}{L}}U^v$ where $\{U^a\}_a, U^v$ are orthogonal matrices, and $0 < \epsilon \ll 1$ is an initialization scale.
3. For JEPA, we set U^v to be diagonal, and for MAE we set U^v such that $U^v(\prod_a U^a)$ is diagonal.

A few notes on assumption 3.1 before stating our results. The assumption on orthogonal weights implies that the weights are "balanced" at initialization, as is typically assumed in prior works on deep linear networks [28, 22]. The extra technical constraint on the uniformity of the initialization eigenvalues is necessary to deal with the non-uniformity of the data eigenvalues and approximately holds in typical initializations with large Gaussian matrices. Likewise, we initialize V differently for JEPA and MAE to achieve alignment of eigenvectors between the model and data at initialization. Although somewhat unrealistic, our special initialization schemes will prove informative to the general case. Applying assumption 3.1 to JEPA and MAE objectives in eq. (5) respectively, and using the gradient flow equations in eq. (6), one arrives at the following theorem:

Theorem 3.2 (ODE Equivalence)**.** *Suppose $\{W^a\}_{a=1\cdots L}$ and V are initialized according to assumption 3.1. Let $\bar{w}_i = \|\bar{W}e_i\|$, where e_is are the standard basis. Furthermore, assume the JEPA objective in eq.* (5) *is optimized using gradient flow according to eq.* (6)*. Then, we have*

$$\textit{JEPA} : \dot{\bar{w}}_i(t) = \bar{w}_i(t)^{3-\frac{1}{L}}\lambda_i - \bar{w}_i(t)^3\lambda_i\rho_i^{-1}. \tag{7}$$

Similarly, the MAE objective eq. (5) *is optimized using gradient flow according to eq.* (6) *yielding:*

$$\textit{MAE} : \dot{\bar{w}}_i(t) = \bar{w}_i(t)^{2-\frac{1}{L}}\lambda_i - \bar{w}_i(t)^3\lambda_i\rho_i^{-1}. \tag{8}$$

Remarkably, the only difference between the JEPA and MAE dynamical equations is in an exponent. For a derivation of these dynamical equations see appendix B.1. Although the assumption 3.1 enables us to characterize the learning dynamics fully, we will show via numerical simulations (see section 4), that the observations made in this paper will not change qualitatively under the more general and realistic initialization. Since ODEs for each component i decouple, for the remainder of this paper, we drop the subscript i and consider a single ODE for $\bar{w}$ parameterized by the encoder and data feature parameters $\{\lambda, \rho\}$. A direct corollary to theorem 3.2 is that the training dynamics of MAE for $L >> 1$ as described in eq. (8) approach those of JEPA in eq. (7) for $L = 1$. Additionally, a depth-dependent difference is apparent in the fixed-point solutions between the two objectives. This is formalized in the following corollary:

Corollary 3.3. *Let $\bar{w}_{\mathit{MAE}}(t, L), \bar{w}_{\mathit{JEPA}}(t, L)$ denote the solutions to eqs.* (7) *and* (8) *for depth L encoders and at time t, given initial condition $\bar{w}(0) = \epsilon$, we have*

$$\bar{w}_{\mathit{MAE}}(\infty, L) = \rho^{\frac{L}{L+1}}, \quad \bar{w}_{\mathit{JEPA}}(\infty, L) = \rho^L. \tag{9}$$

In addition, it holds that the dynamics of a 1-layer JEPA model match an infinite-depth MAE

$$\lim_{L'\to\infty} \bar{w}_{\mathit{MAE}}(t, L') = \bar{w}_{\mathit{JEPA}}(t, 1). \tag{10}$$

When $L > 1$, JEPA will suppress the encoding of noisier directions with a lower regression coefficient and this suppression increases with the network depth. At large depths, the encoder becomes low-rank to the point a single dominating eigenvalue corresponding to the feature with the largest regression coefficient remains. Note that corollary 3.3 also indicates that the JEPA solution for $L > 1$ is not reachable by the MAE objective, and vice versa. While interesting in itself, the difference between the two objectives runs deeper than their fixed-point solutions and is found by analyzing the training dynamics.

In the small initialization regime, the evolution of the weights during training for both objectives exhibits incremental learning dynamics, in which the encoder learns features progressively. We define the *critical time*, denoted as t^*, to be the time it takes for the encoder projection $\bar{w}$ to reach a positive finite-but-close-to-1 fraction p of its final fixed point value. It is a quantity we wish to track since it captures an important data-dependent difference between JEPA and MAE training dynamics, concerning the order in which feature learning proceeds in the encoder:

$$\bar{w}(t^*) = p\bar{w}(\infty). \tag{11}$$

The dependence of this definition of the critical time on a choice for p introduces a degree of arbitrariness in the actual value of t^*. However, as long as p is not too close to 1 or 0, the leading order in the Laurent expansion for $t^* = t^*(\epsilon, \lambda, \rho)$ in ϵ, as we show in section 3.1, is not affected by the specific values of p. The temporal ordering of $t^\star(\lambda, \rho)$s across all features, allows us to observe which feature takes priority in learning according to each objective.

3.1 Critical Time Analysis

To derive the critical time we first solve the dynamical equations in eqs. (7) and (8) in the following. The JEPA dynamics given by eq. (7) can be solved implicitly in a closed form as given by theorem B.7 with a full proof in the supplementary material section.

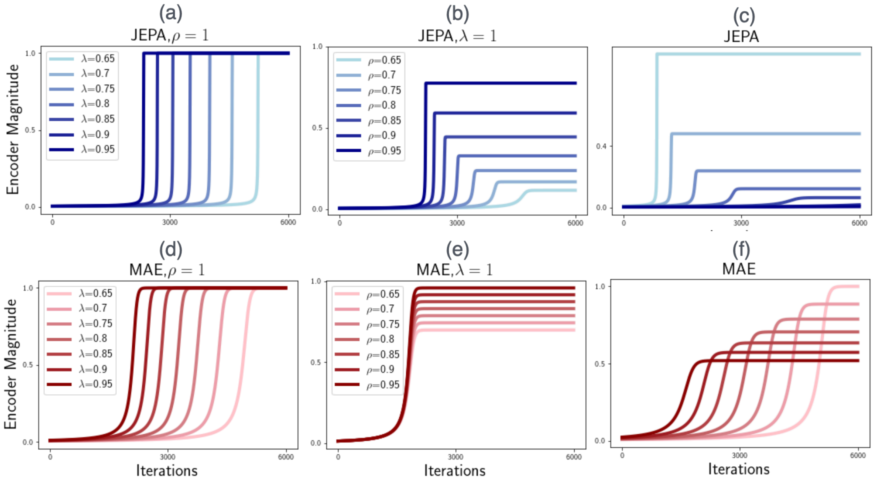

Figure 2: **Simulations of the JEPA and MAE equivalent ODEs (eqs. (7) and (8)).** Each curve represents a numerical simulation of the corresponding ODE, for different values of ρ, λ. (a), (d) darker curves correspond to higher λ and $\rho = 1$. (b), (e) darker curves correspond to higher ρ and $\lambda = 1$. As can be seen, both objectives exhibit greedy learning dynamics with respect to λ, however, only JEPA exhibits greedy dynamics with respect to ρ. (c), (f) darker curves correspond to higher λ but lower ρ. In this case, the order of learning is inverted between the objectives due to the different trends in ρ, λ.

The closed form solution implicitly describes the full trajectory of $\bar{w}(t)$ for any time t. From this solution, it is clear that the encoder projection in the standard basis $\bar{w}$, starting from its initial value ϵ at $t = 0$, will reach its corresponding asymptotic value at $t = \infty$. The critical time t^*, defined in (11), is the time scale for which $\bar{w}$ exits the dynamical region near $t = 0$ and reaches a finite fraction of its asymptotic value. Theorem 3.4 gives the critical time for JEPA with arbitrary depth L encoders:

Theorem 3.4 (JEPA critical time)**.** *The critical time t^* in the small initialization regime $\bar{w}(t = 0) = \epsilon \ll 1$ for JEPA is given by*

$$t^*_{jepa} = \frac{1}{\lambda} \sum_{n=1}^{2L-1} \frac{L}{n\rho^{2L-n-1}\epsilon^{\frac{n}{L}}} + \Theta\Big[\log(\epsilon)\Big], \tag{12}$$

as long as p is not too close, as defined in eq. (81)*, to zero or one.*

In the small initialization regime, the JEPA solution given in theorem B.7 then describes a step-wise learning process of features where each feature is learned in the time scale given by eq. (12). We also derive an analogous theorem B.10 applied to the MAE objective. For technical reasons, we assume $L > 1$ in the following and provide the equivalent theorems for $L = 1$ in Appendix B.4. From the MAE dynamics summarized in , eq. (12) we arrive at the following result for MAE critical time:

Theorem 3.5 (MAE critical time)**.** *The MAE critical time t^* in the small initialization regime of $\bar{w}(t = 0) = \epsilon \ll 1$ and $L > 1$ is given by*

$$t^*_{mae} = \frac{L}{\lambda(L-1)\epsilon^{\frac{L-1}{L}}} + \Theta(1). \tag{13}$$

as long as p is not too close, as defined in eq. (109)*, to zero or one.*

3.2 Comparing Learning Dynamics: JEPA vs. MAE

theorem 3.4 and theorem 3.5 reveal a crucial distinction between the two objectives. To understand how each objective prioritizes features, we note the functional form of the leading orders in ϵ in the critical time $t^\star$. For MAE, we observe that the leading order term depends principally on the inverse of λ. Crucially, we note that the next to leading order term (NLO) is small in comparison for any encoder depth L. This implies step-wise learning dynamics where the step ordering is predominantly set by the feature covariance λ. Conversely, features with identical λ will be learned in the same timescale. In the case of JEPA, we observe that the leading order term is again principally a function of λ, however, the NLO term, which depends inversely on the regression coefficient ρ, is increasingly large with the encoder depth L. That is, features with identical λ but with different ρ will be learned in different timescales, where the feature with the highest regression coefficient is prioritized. Moreover, the priority towards large ρ increases with depth. This implies a meaningful and distinct implicit bias of the JEPA objective towards learning features with a high regression coefficient, increasingly so with depth. This is summarized in the following corollary:

Corollary 3.6. *Let $t^*_{jepa}(\epsilon, \lambda, \rho), t^*_{mae}(\epsilon, \lambda, \rho)$ denote the critical time for the JEPA and MAE objectives given feature parameters ϵ, λ, ρ. WLOG assume $\rho' > \rho$, let $\Delta_\rho = \frac{1}{\rho} - \frac{1}{\rho'}$. Then, it holds that*

$$\frac{t^*_{jepa}(\epsilon, \lambda, \rho)}{t^*_{jepa}(\epsilon, \lambda, \rho')} = 1 + \frac{2L-1}{2L-2}\Delta_\rho \epsilon^{\frac{1}{L}} + \Theta(\epsilon^{\frac{2}{L}}), \quad \frac{t^*_{mae}(\epsilon, \lambda, \rho)}{t^*_{mae}(\epsilon, \lambda, \rho')} = 1 + \Theta\Big[\epsilon^{\frac{L-1}{L}}\Big] \tag{14}$$

We note the strong dependency on depth L and Δ_ρ in the JEPA case through the term $\Delta_\rho \epsilon^{\frac{1}{L}}$ in eq. (14), indicating greedy learning dynamics with respect to ρ, a property absent in the case of MAE. We next turn to empirical simulations to test the validity of our theory, as well as its generalizability in less stringent settings.

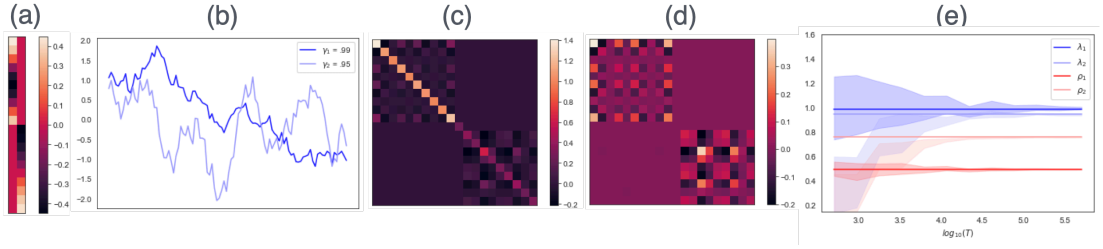

Figure 3: **Temporal model** (15): (a) v_1, v_2, (b) temporal dynamics of flickering u, autocorrelation $\gamma_1 = 0.99$ and $\gamma_2 = 0.95$, (c) $\hat{\Sigma}^{xx}$ empirical covariance with 500k samples, (d) $\hat{\Sigma}^{xy}$ empirical correlation, (e) predictions versus simulations of parameters λ and ρ varying $\log_{10}(T)$. λ decreases from feature 1 to 2 whereas ρ increase because noise added standard deviation decreases from 1 to 0.5. Note (e) show 2 standard deviation with 10 runs.

4 Numerical Experiments

We also conduct experiments directly verifying our theoretical findings. To that end, we numerically simulate the ODEs in eqs. (7) and (8) for different distributions and depths starting from a small initialization. In Figure 2, we fix the encoder depth $L = 5$ while varying the data parameters. As predicted by the theory, we observe a clear greedy learning process for both objectives. However, only the JEPA objective exhibits greedy dynamics with respect to ρ, where components with larger ρ are learned first. Moreover, we observe that the JEPA objective can reverse the feature learning order relative to MAE when ρ and λ have the opposite trends, qualitatively reproducing the observation in section 2. In Figure 4, we run further simulations by varying the depth parameter L. As predicted, we observe that deeper encoders introduce a wider temporal spread between learning features with the same λ but different ρ, in contrast to MAE which learns all features at the same time independent of the depth. Moving beyond gradient flow and the assumptions on initialization, in Figure 5 we train randomly initialized deep linear neural networks using stochastic gradient descent on randomly sampled batches of Gaussian data with varying values for ρ, λ, using both objectives. As is evident, our results carry over to this setup as well, where greedy learning of features can be seen with an objective dependent ordering according to ρ, λ.

4.1 Linear Generative Models

We consider generative models which satisfy our simultaneously-diagonalizable assumption (in the large sample limit). This is a setting in which our theory holds exactly and datasets with this property allow both JEPA and MAE models to learn the same modes, although not necessarily in the same order. We analyze the simplest linear generative models which still lead to non-trivial learning differences between JEPAs and MAEs for two settings: *random masking* and a *temporal model*. The full details of these different settings can be found in Appendix C. In the first case, our views are random masks of the data, this corresponds to a static setting and we derive closed-form expressions for λ and ρ (see C.1). For the remainder of this section, we will focus in on the temporal model due to its simplicity and intuitive interpretation.

The temporal model considers the scenario where our views x, y correspond to our data $z^1, ..., z^T$ at two consecutive times: $x = z^t$ and $y = z^{t+1}$. The goal is to predict the next frame from the previous one (this setting is consistent with comparing VideoMAE and V-JEPA video models). We now define a simple linear model corresponding to combining independently time-varying images $\{v^a\}$ to the model simultaneously with random noise. We can write this mathematically as:

$$z_i^t = \sum_a u^a(t) v_i^a + \xi_i^t, \quad t = 1, ...T. \tag{15}$$

Note that we are allowing the amplitude of the images to vary with time according to a scalar function $u(t)$; we make this choice to study the simplest non-trivial temporal variability. The full details of the setup and derivation for simultaneous diagonalizability can be found in Appendix C.2, where we derive the correlation coefficient for each feature:

$$\lambda_a = \gamma_a \|v^a\|^2, \quad \rho_a = \frac{\gamma_a \|v^a\|^2}{\sigma_a^2 + \|v^a\|^2}. \tag{16}$$

Here σ_a represents the noise amplitude applied to mode a. See Figure 3 for a demonstration of our theory matched to simulations and an example of higher correlation modes with lower correlation coefficient in Figure 3(e). Additionally, a simulation demonstrating an approach to simultaneously diagonalizable empirical covariance and correlation $\hat{\Sigma}^{xx}$ and $\hat{\Sigma}^{xy}$ can be found in Figure 6. Combining (16) with our theory (14) demonstrates how JEPAs can be slower to learn noisy features in a way that MAEs are not. JEPA will also learn noisy features by converging to lower amplitude weights (9), making such features potentially less likely to be used downstream from the embedding.

5 Related Work

Self-supervised learning. One successful SSL paradigm is MAE [8] and its variants [13, 14] which learn representations via input space reconstruction. However, recent works (e.g., [20]) show that such generative paradigms often learn uninformative features for perception, since MAE-like methods tend to learn the data subspace that explains the bulk of observed variance and can include unhelpful information for perception. By contrast, the JEPA paradigm [9, 10, 11, 18, 19] learns to predict the representations of similar image views in the latent space. Such objectives have been found to prioritize semantic features over detailed pixel information, leading to superior performance for perception tasks. To prevent the feature collapse of JEPA, stop-gradients, and other architectural tricks are widely used. More recent works propose techniques like spectrum modulation [29] and weight regularization [30]. Another popular collapse prevention method is based on contrastive loss against negative image pairs [2, 5, 6, 11, 7]. In this work, we focus on MAE and JEPA methods, seeking to understand their implicit bias that leads to different training dynamics. To measure representation quality in JEPAs, [31] devised a metric that effectively counts the number of directions with a large regression coefficient in latent space. Our work can be seen as further motivation for this approach.

Theoretical analysis of SSL. There have been several recent works attempting to understand the success of non-contrastive SSL (JEPA) from various perspectives. [32] studies how the self-distillation in JEPA avoids representation collapse, finding the key role of eigenspace alignment between the predictor and input correlation matrix. The authors of [33] further provide a theoretical bridge between contrastive and non-contrastive objectives towards global and local spectral embedding methods respectively. It is shown that all these methods can be deployed successfully if the pairwise relations during SSL are aligned with the downstream task. While for MAE-based methods, [20] shows they tend to learn uninformative features by pixel reconstruction, unlike the JEPA objective that prioritizes semantic features. In this paper, we seek to understand the mechanism behind this empirical observation by characterizing the implicit bias of both methods using deep linear models.

Learning dynamics in linear networks. Deep linear neural networks have been used in [24, 25, 26, 27, 34, 28, 35, 36] to study the nonlinear dynamics of learning in various settings applying different assumptions, including the saddle point behavior as well as the step-wise behavior where distinct features are learned at different time scales. This is also in line with the observed greedy learning dynamics for modern architectures like Transformers [22, 23]. In the context of SSL, the authors of [21] observed a similar step-wise nature which corresponds to the learning of eigenmodes of the contrastive kernel $\Sigma^{yx} + \Sigma^{xy}$. However, we go beyond this observation by characterizing the distinctive implicit bias of JEPA relative to MAE, i.e., they learn the same features in a step-wise fashion but not necessarily in the same order. Surprisingly, this difference in behavior is only apparent when the encoder network is deep. Our findings about JEPA are related to empirical observations in [37] which show JEPA-based methods focus on "slow features" that vary slowly over time. This is confirmed in our studies where JEPA-based methods will prioritize "slow features" for which the marginal distribution has the smallest variance.

6 Limitations

Our theory has several clear limitations. The theoretical results presented in this paper are restricted to conditions in assumption 3.1, which include a deep linear MLP with some restrictions on the initialization scheme, as well as a simultaneously diagonalizable covariances Σ^{yx}, Σ^{xx}. It is worth discussing the implications of these assumptions, and the generalizability of our results to more typical settings. On the model side, empirical simulations with deep linear models initialized using the default initialization [38] indicate our results qualitatively generalize to popular initialization schemes. However, a potentially stronger assumption we make is simultaneously diagonalizability of Σ^{yx} and Σ^{xx}, which allows JEPA and MAE to learn the same features. We provide generative frameworks that adhere to this property, however we expect data distributions of interest to significantly deviate from this assumption. Moreover, the generalizability of our claims to fully practical scenarios yet remains unexplored, and precise characterization of the implicit biases pertaining to more general data distributions and architectures is a direction for future work.

7 Discussion

We have introduced toy models of two of the most popular self-supervised algorithms, namely JEPA and MAE, where we can completely characterize the learning process. We have uncovered a novel learning principle that separates the two objectives: while MAE focuses on highly co-varying features early in training, the JEPA objective focuses on highly influential features (indicated by high values of the regression coefficient). This implicit bias of the JEPA objective allows it to focus on features that are predictive, yet contain minimal noise, as measured by their variance across the training set. This new understanding of the learning dynamics of JEPA sheds light on recent empirical observations and opens the door to new avenues of research strengthening these results. Certainly, one must first generalize these results to non-simultaneously-diagonalizable data distributions, as well as the more imposing challenge of non-linear models. Still, a less than rigorous leap to practical settings may still provide valuable insights. For example, an implicit bias towards predictive yet low-variance features may explain the tendency of JEPA to learn more semantic features, which are inherently less noisy. Conversely, it may shed light on JEPAs vulnerability to spurious or *slow* features. Additionally, the feature prioritization distinction between the objectives may bear on the efficiency of JEPA in learning semantic features quickly, as these features tend to reside in the low variance subspace of the data, as shown recently [20]. Finally, it is also worth investigating whether the mechanism uncovered here generalizes to other joint embedding architectures not relying on self-distillation.

8 Acknowledgments

We thank Dan Busbridge, Arwen Bradley, Stefano Cosentino, and Shuangfei Zhai for stimulating discussions and useful feedback on the research and writing of this paper.

References

[1] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *ArXiv*, abs/2105.04906, 2021.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[7] Chen Huang, Hanlin Goh, Jiatao Gu, and Josh Susskind. MAST: Masked augmentation subspace training for generalizable self-supervised priors. In *ICLR*, 2023.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

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

[10] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[12] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ArXiv*, abs/1808.06670, 2018.

[13] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[14] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[15] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[18] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[19] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

[20] Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

[21] James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pages 31852–31876. PMLR, 2023.

[22] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *ArXiv*, abs/2012.09839, 2020.

[23] Enric Boix-Adserà, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua M. Susskind. Transformers learn through gradual rank increase. *ArXiv*, abs/2306.07042, 2023.

[24] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[25] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[26] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature communications*, 9(1):4950, 2018.

[27] Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, 2022.

[28] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Neural Information Processing Systems*, 2019.

[29] Xi Weng, Yunhao Ni, Tengwei Song, Jie Luo, Rao Muhammad Anwer, Salman Khan, Fahad Shahbaz Khan, and Lei Huang. Modulate your spectrum in self-supervised learning. *International Conference on Learning Representations (ICLR)*, 2024.

[30] Ali Saheb Pasand, Reza Moravej, Mahdi Biparva, and Ali Ghodsi. WERank: Towards rank degradation prevention for self-supervised learning using weight regularization. *ArXiv*, abs/2402.09586, 2024.

[31] Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Josh Susskind, and Etai Littwin. Lidar: Sensing linear probing performance in joint embedding ssl architectures. *ArXiv*, abs/2312.04000, 2023.

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

[33] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022.

[34] Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[35] Raphael Berthier. Incremental learning in diagonal linear networks. *J. Mach. Learn. Res.*, 24:171:1–171:26, 2022.

[36] Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *ArXiv*, abs/2304.00488, 2023.

[37] Vlad Sobal, V JyothirS, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. *ArXiv*, abs/2211.10831, 2022.

[38] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[41] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The central claims made in the paper are stated in the Abstract and the Introduction. Additionally, we enumerate our contributions in the Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 6 in the main text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate ”Limitations” section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions and main theoretical results are included in the main text. Detailed proofs are available in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup details are included in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The major contributions of the paper are theoretical in nature. The experiments included in the paper are used to support and illustrate analytical results. Consequently, the models used are simple linear models trained with synthetic data. The paper describes how to generate data used in the experiments as well as other setup information needed to run and reproduce experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all details used to construct datasets used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments are used to verify predictions made by the analytical model developed in the paper. The theory developed in this paper suggests strong effects that we illustrate via experiments. These effects are strong and do not require making fine-grained comparisons via statistical significance testing.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is included with experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the work presented in this paper is performed in a manner consistent with NeurIPS Ethics Guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact of our research in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work is theoretical in nature so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We include this information in experimental details.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

Contents

Appendix A Additional Figures

A.1 Experimental Details

For Figure 1 and Figure 5, we train MLPs using stochastic gradient descent on random Gaussian data. We use a constant $d = 100$ width network and fix the encoder depth to be $L = 5$. We sample x, y from a jointly Gaussian distribution with diagonal covariances, with different values of ρ, λ per feature, as illustrated in the first column of each figure. At each iteration, we randomly sample a minibatch of 1000 samples and train for 5000 iterations.

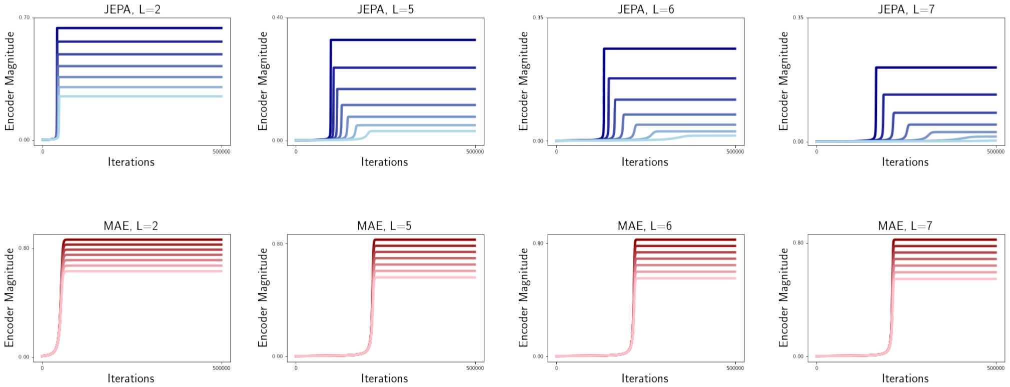

Figure 4: **Simulations of the JEPA and MAE equivalent ODEs for** $L = 2, 5, 6, 7$ **(eqs. (7) and (8)).** The covariance λ is fixed to 1 across all curves, and darker curves correspond to a higher ρ. As evident, only in the case of the JEPA objective, deeper encoders induce a more pronounced incremental learning of features with respect to ρ.

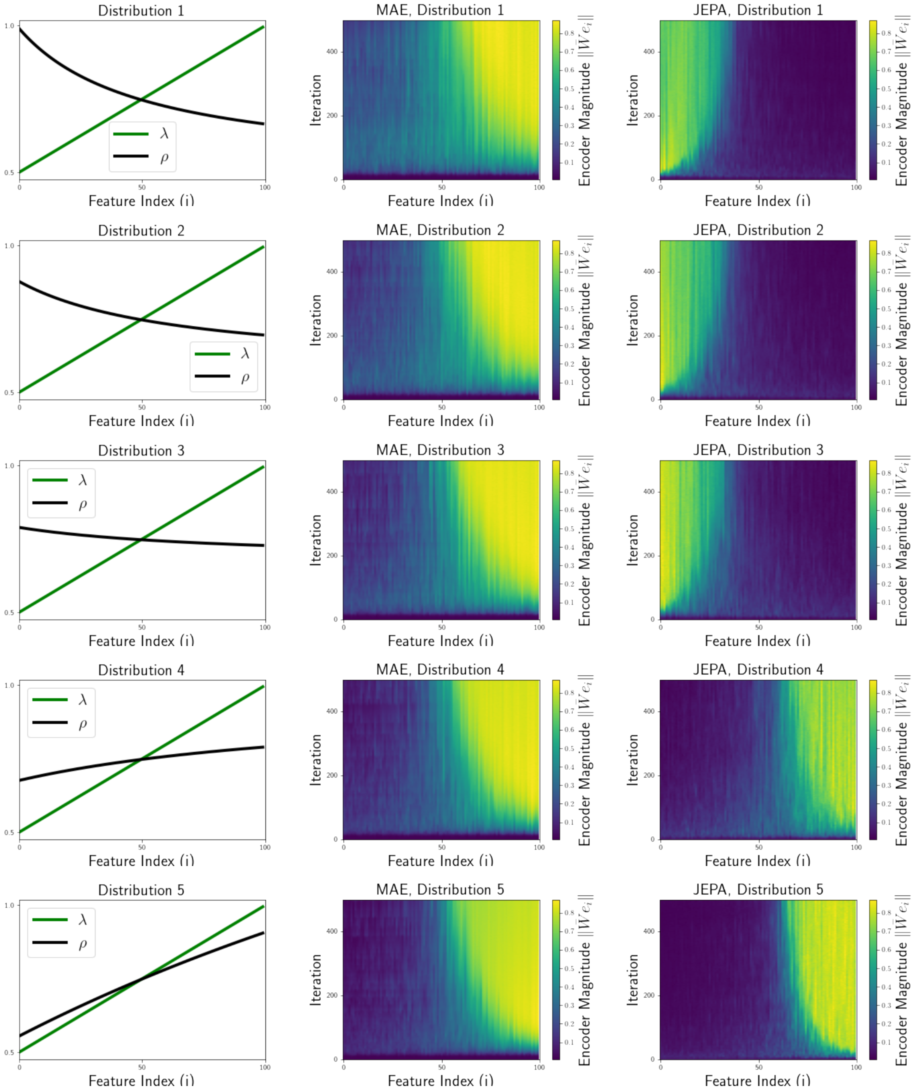

Figure 5: **Deep linear networks trained on Gaussian data.** The left most column represents the values of λ, ρ used to generate the data. All networks were initialized using standard gaussian initialization with default scale, and the encoder depth is fixed to $L = 5$. The order of feature learning is dictated by ρ for the JEPA objective, and λ for the MAE objective.

Appendix B Proofs

In the following sections, we derive dynamical equations for JEPA and MAE and characterize their training dynamics. We also write down closed-form implicit solutions for their dynamics, all presented in the form of proofs for the theorems/corollaries mentioned in the main paper.

B.1 Proofs of theorem 3.2 and corollary 3.3

Theorem 3.2 (ODE Equivalence)**.** *Suppose $\{W^a\}_{a=1\cdots L}$ and V are initialized according to assumption 3.1. Let $\bar{w}_i = \|\bar{W}e_i\|$, where e_is are the standard basis. Furthermore, assume the JEPA objective in eq.* (5) *is optimized using gradient flow according to eq.* (6)*. Then, we have*

$$\textit{JEPA}: \dot{\bar{w}}_i(t) = \bar{w}_i(t)^{3-\frac{1}{L}}\lambda_i - \bar{w}_i(t)^3\lambda_i\rho_i^{-1}. \tag{7}$$

Similarly, the MAE objective eq. (5) *is optimized using gradient flow according to eq.* (6) *yielding:*

$$\textit{MAE}: \dot{\bar{w}}_i(t) = \bar{w}_i(t)^{2-\frac{1}{L}}\lambda_i - \bar{w}_i(t)^3\lambda_i\rho_i^{-1}. \tag{8}$$

Proof. Let us begin with the derivation of the JEPA dynamical equation. The gradient flow equations are given by

$$\forall_a,\ \ \dot{W}^a(t) = -\nabla_{W^a(t)}\mathcal{L}_{jepa},\quad \dot{V}(t) = -\nabla_{V(t)}\mathcal{L}_{jepa}. \tag{17}$$

Evaluating the gradient using the JEPA loss in eq. (5), one obtains

$$\forall_a,\ \ \nabla_{W^a(t)}\mathcal{L}_{jepa} = \mathbb{E}\Big[\Big(\prod_{b=a+1}^{L} W^b(t)\Big)^\top V(t)^\top\Big[V(t)\bar{W}(t)x - \bar{W}(t)y\Big]x^\top\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top\Big], \tag{18}$$

$$\nabla_{V(t)}\mathcal{L}_{jepa} = \mathbb{E}\Big[(V(t)\bar{W}(t)x - \bar{W}(t)y)x^\top\bar{W}(t)^\top\Big]. \tag{19}$$

Substituting for the expected values using $\mathbb{E}[xx^\top] = \Sigma^{xx}$ and $\mathbb{E}[yx^\top] = \Sigma^{yx}$, one gets

$$\begin{aligned}\forall_a,\ \ \nabla_{W^a(t)}\mathcal{L}_{jepa} = &\Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top V(t)\bar{W}(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top \\ &- \Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top\bar{W}(t)\Sigma^{yx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top,\end{aligned} \tag{20}$$

$$\nabla_{V(t)}\mathcal{L}_{jepa} = \mathbb{E}\Big[(V(t)\bar{W}(t)x - \bar{W}(t)y)x^\top\bar{W}(t)^\top\Big] = V(t)\bar{W}(t)\Sigma^{xx}\bar{W}(t)^\top - \bar{W}(t)\Sigma^{yx}\bar{W}(t)^\top. \tag{21}$$

With the gradient flow equations eq. (17) in mind, we get the following system of ODEs for $\{w^a, V\}$ to solve

$$\begin{aligned}\forall_a,\ \ \dot{W}^a(t) = &\Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top\bar{W}(t)\Sigma^{yx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top \\ &- \Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top V(t)\bar{W}(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top,\end{aligned} \tag{22}$$

$$\dot{V}(t) = \bar{W}(t)\Sigma^{yx}\bar{W}(t)^\top - V(t)\bar{W}(t)\Sigma^{xx}\bar{W}(t)^\top. \tag{23}$$

Similarly, using the MAE loss given by eq. (5), the gradient flow equations for MAE lead to the following system of ODEs

$$\begin{aligned}\forall_a,\ \ \dot{W}^a(t) = &\Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top\Sigma^{yx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top \\ &- \Big[\prod_{b=a+1}^{L} W^b(t)\Big]^\top V(t)^\top V(t)\bar{W}(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1}W^b(t)\Big]^\top,\end{aligned} \tag{24}$$

$$\dot{V}(t) = \Sigma^{yx}\bar{W}(t)^\top - V(t)\bar{W}(t)\Sigma^{xx}\bar{W}(t)^\top. \tag{25}$$

For ease of presentation, in this appendix we will use $V(t) = W^{L+1}(t)$, and denote $W(t) = W^{L+1}(t)\bar{W}(t)$. With this notation, the dynamical equations for JEPA can be rewritten for $W = W(t)$ and $\bar{W} = \bar{W}(t)$ as follows

$$\begin{aligned}\forall_a,\ \dot{W}^a(t) = {} & \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \bar{W}(t)\Sigma^{yx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top \\ & - \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top.\end{aligned} \tag{26}$$

For $\dot{V}(t) = \dot{W}^{L+1}(t)$ we obtain

$$\dot{W}^{L+1}(t) = \bar{W}(t)\Sigma^{yx}\bar{W}(t)^\top - W(t)\Sigma^{xx}\bar{W}(t)^\top. \tag{27}$$

From the product rule for derivative, we have

$$\dot{W}(t) = \sum_{a=1}^{L+1}\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\dot{W}^a(t)\Big[\prod_{b=1}^{a-1} W^b(t)\Big]. \tag{28}$$

Substituting the expression for $\dot{W}^a(t)$ in (28) gives

$$\begin{aligned}\dot{W}(t) = {} & \sum_{a=1}^{L+1}\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \bar{W}(t)\Sigma^{yx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top\Big[\prod_{b=1}^{a-1} W^b(t)\Big] \\ & - \sum_{a=1}^{L+1}\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top\Big[\prod_{b=1}^{a-1} W^b(t)\Big].\end{aligned} \tag{29}$$

Similarly, the dynamical equations for MAE can be re-expressed for $W = W(t)$

$$\begin{aligned}\forall_a,\ \dot{W}^a(t) = {} & \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \Sigma^{yx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top \\ & - \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top.\end{aligned} \tag{30}$$

For $\dot{V}(t) = \dot{W}^{L+1}(t)$ we get

$$\dot{W}^{L+1}(t) = \Sigma^{yx}\bar{W}(t)^\top - W(t)\Sigma^{xx}\bar{W}(t)^\top. \tag{31}$$

For MAE, plugging the expression for $\dot{W}^a(t)$ into (28) results in

$$\begin{aligned}\dot{W}(t) = {} & \sum_{a=1}^{L+1}\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \Sigma^{yx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top\Big[\prod_{b=1}^{a-1} W^b(t)\Big] \\ & - \sum_{a=1}^{L+1}\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top\Big[\prod_{b=1}^{a-1} W^b(t)\Big].\end{aligned} \tag{32}$$

In the following, we study the relevant consequences of these dynamical equations and the assumption 3.1. Let us begin by noting that the above assumptions on initialization imply that the weights are "balanced" at initialization, i.e., $W^{a+1\top}(0)W^{a+1}(0) = W^a(0)W^{a\top}(0)$, for $a = 1...L$. In the following lemma, we show during training, the weights will remain balanced:

Lemma B.1. *$W^{a+1\top}(t)W^{a+1}(t) = W^a(t)W^{a\top}(t)$ for $a = 1...L$, if $W^a(t)$ satisfies the ODE sets* (26) *and* (27) *or* (30) *and* (31).

Proof. Multiplying both sides of the MAE dynamic equation (30) from the right by $W^{a\top}(t)$ for $a = 1 \cdots L-1$ gives (after absorbing $W^{a\top}$ in the relevant products)

$$\begin{aligned}\dot{W}^a(t)W^{a\top}(t) = &\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \Sigma^{yx}\Big[\prod_{b=1}^{a} W^b(t)\Big]^\top \\ &- \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a} W^b(t)\Big]^\top.\end{aligned} \tag{33}$$

Rewriting (30) for $a+1 \leq L$ layer index and multiplying by $W^{a+1\top}$ from the left results in

$$\begin{aligned}W^{a+1\top}(t)\dot{W}^{a+1}(t) = &\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top \Sigma^{yx}\Big[\prod_{b=1}^{a} W^b(t)\Big]^\top \\ &- \Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top W(t)\Sigma^{xx}\Big[\prod_{b=1}^{a} W^b(t)\Big]^\top \\ &= \dot{W}^a(t)W^{a\top}(t).\end{aligned} \tag{34}$$

Transposing (34), one obtains

$$\dot{W}^{a+1\top}(t)W^{a+1}(t) = W^a(t)\dot{W}^{a\top}(t). \tag{35}$$

Using (27) and (31), it can also be shown that

$$\begin{aligned}V^\top(t)\dot{V}(t) &= \dot{W}^L(t)W^{L\top}(t), \\ \dot{V}^\top(t)V(t) &= W^L(t)\dot{W}^{L\top}(t).\end{aligned} \tag{36}$$

Therefore, for $a = 1 \cdots L$

$$W^{a+1\top}(t)\dot{W}^{a+1}(t) + \dot{W}^{a+1\top}(t)W^{a+1}(t) = \dot{W}^a(t)W^{a\top}(t) + W^a(t)\dot{W}^{a\top}(t), \tag{37}$$

which can be written as

$$\frac{d}{dt}[W^{a+1\top}(t)W^{a+1}(t) - W^a(t)W^{a\top}(t)] = 0, \tag{38}$$

at all times. Consequently, $W^{a+1\top}(t)W^{a+1}(t) - W^a(t)W^{a\top}(t) = \mathcal{C}$ where $\mathcal{C}$ is a matrix with constant entries. On the other hand, according to assumption 3.1, W^as $a = 1 \cdots L+1$ are scaled orthogonal matrices at initialization, therefore $\mathcal{C} = \mathbf{0}$. This implies

$$W^{a+1\top}(t)W^{a+1}(t) = W^a(t)W^{a\top}(t), \tag{39}$$

at all times for $a = 1 \cdots L$, which proves the claim for MAE. The proof for the JEPA case is similar. This concludes the proof of the lemma. □

Using the above lemma one can write the following alternative form for (29) and (32)

Lemma B.2. *The dynamical equations for JEPA and MAE share the same general form as follows*

$$\begin{aligned}\dot{W} = &\sum_{a=1}^{L+1}[W(t)W(t)^\top]^{1-\frac{a}{L+1}}\Xi_{jepa/mae}[W(t)^\top W(t)]^{\frac{a-1}{L+1}} \\ &- \sum_{a=1}^{L+1}[W(t)W(t)^\top]^{1-\frac{a}{L+1}}W(t)\Sigma^{xx}[W(t)^\top W(t)]^{\frac{a-1}{L+1}},\end{aligned} \tag{40}$$

where $\Xi_{jepa} = \bar{W}(t)\Sigma^{yx}$ *and* $\Xi_{mae} = \Sigma^{yx}$.

Proof. Let $W^a(t) = Q_a(t)\omega_a(t)P_a(t)$ denote its SVD decomposition, where Q^a, P^a are the left and right eigenbasis of the corresponding SVD decomposition. Using lemma B.1, we know

$$W^{a+1\top}(t)W^{a+1}(t) = W^a(t)W^{a\top}(t), \tag{41}$$

for $a = 1 \cdots L$. Substituting the corresponding SVD decomposition for W^as implies

$$P_{a+1}^\top(t)\omega_{a+1}^2(t)P_{a+1}(t) = Q_a(t)\omega_a^2(t)Q_a^\top(t), \tag{42}$$

which leads to

$$P_{a+1}^\top(t) = Q_a(t), \forall_{a=1...L}, \tag{43}$$

and the conclusion that all ωs are equal

$$\omega_a(t) = \omega(t), \forall_{a=1..L+1}. \tag{44}$$

It is easy to see that this implies

$$\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]\Big[\prod_{b=a+1}^{L+1} W^b(t)\Big]^\top = Q_{L+1}(t)\omega^{2(L-a+1)}Q_{L+1}^\top(t), \tag{45}$$

$$\Big[\prod_{b=1}^{a-1} W^b(t)\Big]^\top\Big[\prod_{b=1}^{a-1} W^b(t)\Big] = P_1(t)\omega^{2(a-1)}P_1^\top(t). \tag{46}$$

On the other hand

$$W(t)W^\top(t) = Q_{L+1}(t)\omega^{2(L+1)}(t)Q_{L+1}^\top(t), \tag{47}$$

$$W^\top(t)W(t) = P_1^\top(t)\omega^{2(L+1)}(t)P_1(t). \tag{48}$$

Using the above relations, the equations (29) and (32) result in the statement of the lemma. This concludes the proof. □

Now we are in a position to state and prove the following theorem:

Theorem B.3. *For JEPA, suppose $\bar{\omega}(t) = \omega(t)^L$ with a diagonal $\omega(t)$. For any time t, the following ansatz*

$$\bar{W}(t) = \mathcal{U}\bar{\omega}(t), \tag{49}$$

$$W(t) = \mathcal{U}\bar{\omega}(t)^{1+\frac{1}{L}}, \tag{50}$$

where $\mathcal{U}$ is a constant orthogonal matrix, solves dynamical equations for JEPA, if $\bar{\omega}(t)$ obeys the following differential equation

$$\dot{\bar{\omega}}(t) = L\Big[\bar{\omega}(t)^{3-\frac{1}{L}}\Sigma^{yx} - \bar{\omega}^3(t)\Sigma^{xx}\Big], \tag{51}$$

with $\bar{\omega}(0) = \epsilon\mathbb{1}$ at initialization.

Proof. First note that the ansatz satisfies conditions 2 and 3 of the assumption 3.1 at initialization. Plugging the ansatz into (40) give

$$\mathcal{U}(1+\frac{1}{L})\bar{\omega}^{\frac{1}{L}}(t)\dot{\bar{\omega}}(t) = (L+1)\mathcal{U}\Big[\bar{\omega}^3(t)\Sigma^{yx} - \bar{\omega}^{3+\frac{1}{L}}(t)\Sigma^{xx}\Big], \tag{52}$$

which is the claimed ODE after rearrangement of the terms and noting $\mathcal{U}$ is non-singular. Also, note that the equation for $V(t)$ is satisfied. In fact, from the ansatz one obtains

$$V(t) = W^{L+1}(t) = \mathcal{U}\bar{\omega}^{\frac{1}{L}}(t)\mathcal{U}^\top. \tag{53}$$

Substituting this in (27), we get

$$\mathcal{U}\frac{d\bar{\omega}^{\frac{1}{L}}(t)}{dt}\mathcal{U}^\top = \mathcal{U}\bar{\omega}^2(t)\Sigma^{yx}\mathcal{U}^\top - \mathcal{U}\bar{\omega}^{2+\frac{1}{L}}(t)\mathcal{U}^\top. \tag{54}$$

Rearranging the terms and expanding the left-hand side gives

$$\mathcal{U}\dot{\bar{\omega}}(t)\mathcal{U}^\top = L\mathcal{U}\Big[\bar{\omega}^{3-\frac{1}{L}}(t)\Sigma^{yx} - \bar{\omega}^3(t)\Sigma^{xx}\Big]\mathcal{U}^\top, \tag{55}$$

which is satisfied if the claimed ODE in the theorem is satisfied. This concludes the proof. □

Theorem B.4. *For MAE, suppose $\bar{\omega}(t) = \omega^L(t)$ with a diagonal $\omega(t)$. The following ansatz*

$$\begin{aligned} \bar{W}(t) &= \mathcal{U}\omega^L(t), \\ W^{L+1}(t) &= \omega(t)\mathcal{U}^\top, \end{aligned} \tag{56}$$

where $\mathcal{U}$ is a constant orthogonal matrix, solves the dynamical equations (40) *and* (31)*, if*

$$\dot{\bar{\omega}}(t) = L[\bar{\omega}^{2-\frac{1}{L}}(t)\Sigma^{yx} - \bar{\omega}^3(t)\Sigma^{xx}], \tag{57}$$

with $\bar{\omega}(0) = \epsilon\mathbb{1}$ at initialization.

Proof. Let us start by noticing that the ansatz is consistent with conditions 2 and 3 of the assumption 3.1 at initialization. From the ansatz

$$W(t) = W^{L+1}\bar{W}(t) = \omega^{1+L}(t). \tag{58}$$

Plugging this expression into the equation (40) gives

$$(L+1)\omega^L(t)\dot{\omega}(t) = \sum_{a=1}^{L+1} \omega^{2(L+1-a)+2(a-1)}\Sigma^{yx} - \sum_{a=1}^{L+1} \omega^{2(L+1-a)+L+1+2(a-1)}\Sigma^{xx}, \tag{59}$$

where the fact that Σ^{yx} and Σ^{xx} are diagonal was used. Simplifying the above expression, one obtains

$$\dot{\omega}(t) = \omega^L(t)\Sigma^{yx} - \omega^{2L+1}(t)\Sigma^{xx}. \tag{60}$$

One can rewrite this ODE in terms of $\bar{\omega} = \bar{\omega}(t)$. Multiply both sides of (60) $L\omega^{L-1}$(t) gives

$$\dot{\bar{\omega}}(t) = L[\bar{\omega}^{2-\frac{1}{L}}(t)\Sigma^{yx} - \bar{\omega}^3(t)\Sigma^{xx}]. \tag{61}$$

Now let us turn to (31). Using the ansatz eq. (56), we can write (31) as

$$\dot{\omega}(t) = \Sigma^{yx}\omega^L(t) - \omega^{L+1}(t)\Sigma^{xx}\omega^L(t) \tag{62}$$

which is the same as (60) given Σ^{yx}, Σ^{xx} and ω are diagonal. This concludes the proof. □

Keeping in mind $\bar{\omega}(0)$ is proportional to identity at initialization and Σ^{xx} and Σ^{xx} are assumed diagonal, $\bar{\omega} = \bar{\omega}(t)$ remains diagonal at all times during its evolution according to (51) and eq. (57). This also means the assumption $\omega = \omega(t)$ diagonal used during derivation of eq. (51) and eq. (57) is a self-consistent one. Let us denote $\bar{w}_i(t) = \bar{\omega}_{ii}(t)$. Observe that

$$\bar{w}_i(t) = \bar{\omega}_{ii}(t) = \|\bar{W}(t)e_i\|. \tag{63}$$

Using the above, one arrives at the following ODE for MAE

$$\dot{\bar{w}}_i(t) = \bar{w}_i(t)^{2-\frac{1}{L}}(\Sigma^{yx})_{ii} - \bar{w}_i(t)^3(\Sigma^{xx})_{ii}, \tag{64}$$

where the constant L was absorbed into the definition of the gradient flow time. At this point, we can drop index i to avoid clutter

$$\dot{\bar{w}}(t) = \bar{w}(t)^{2-\frac{1}{L}}\lambda - \bar{w}(t)^3\lambda\rho^{-1}. \tag{65}$$

Similarly, for JEPA one arrives at

$$\dot{\bar{w}}(t) = \bar{w}(t)^{3-\frac{1}{L}}\lambda - \bar{w}(t)^3\lambda\rho^{-1}, \tag{66}$$

where the definition of the regression coefficient ρ was used. This completes the proof. □

We may now proceed to prove corollary 3.3.

Corollary B.5. *Let $\bar{w}_{\mathit{MAE}}(t, L), \bar{w}_{\mathit{JEPA}}(t, L)$ denote the solutions to eqs.* (7) *and* (8) *for depth L encoders and at time t, given initial condition $\bar{w}(0) = \epsilon$, we have*

$$\bar{w}_{\mathit{MAE}}(\infty, L) = \rho^{\frac{L}{L+1}}, \quad \bar{w}_{\mathit{JEPA}}(\infty, L) = \rho^L. \tag{9}$$

In addition, it holds that the dynamics of a 1-layer JEPA model match an infinite-depth MAE

$$\lim_{L' \to \infty} \bar{w}_{\mathit{MAE}}(t, L') = \bar{w}_{\mathit{JEPA}}(t, 1). \tag{10}$$

Proof. The fixed-point solutions for a finite L can be easily derived by setting $\dot{\bar{w}} = 0$ for both the JEPA and MAE equivalent ODEs and solving for $\bar{w}$. To show that the limit of the MAE fixed-point at $L = \infty$ equals the JEPA fixed-point at $L = 1$, we note that the right-hand side of eq. (8) is a smooth function of $\zeta = \frac{1}{L}$ for $L > 0$. Hence, we have that $\bar{w}_{\text{MAE}}(t, L = \infty)$ is the solution to

$$\dot{\bar{w}}(t, \infty) = \lim_{\zeta \to 0} \bar{w}(t)^{2-\zeta}\lambda - \bar{w}(t)^3\rho^{-1}\lambda = \bar{w}(t)^2\lambda - \bar{w}(t)^3\rho^{-1}\lambda, \tag{67}$$

with the initial condition $\bar{w}(0)$, which is identical to the JEPA dynamics in eq. (7). □

We will now proceed to discuss the results on JEPA dynamics next:

B.2 JEPA dynamics: proofs of theorem B.7 and theorem 3.4

In this section, we solve the JEPA training dynamics given by the ODE in (66) for general depth L. Let us start by stating and proving the following lemma:

Lemma B.6. *Define*

$$\mathcal{I}_L(\psi) = \int \frac{d\psi}{\psi^{2L} - \psi^{2L+1}}, \tag{68}$$

then

$$\mathcal{I}_L = -\sum_{n=1}^{2L-1} \frac{1}{n\psi^n} + \log(\psi) - \log(1-\psi) + \mathcal{C}, \tag{69}$$

where $\mathcal{C}$ is a constant of integration.

Proof. We use induction to prove (69). For $L = 1$, the integral is elementary given by

$$\mathcal{I}_1 = -\frac{1}{\psi} + \log(\psi) - \log(1-\psi) + \mathcal{C}, \tag{70}$$

hence, the statement of the lemma holds. Next, we show if the statement of the lemma holds for an arbitrary $L \neq 1$, it will also hold for $L + 1$. Note that

$$\frac{1}{\psi^{2L+2}(1-\psi)} - \frac{1}{\psi^{2L}(1-\psi)} = \frac{1}{\psi^{2L+2}} + \frac{1}{\psi^{2L+1}}. \tag{71}$$

Integrating both sides with respect to ψ gives the following recursion relation

$$\mathcal{I}_{L+1} - \mathcal{I}_L = -\frac{1}{(2L+1)\psi^{2L+1}} - \frac{1}{2L\psi^{2L}} + \mathcal{C}. \tag{72}$$

Assuming the statement of the lemma holds for some $L \neq 1$, one will have

$$\begin{aligned} \mathcal{I}_{L+1} &= \mathcal{I}_L - \frac{1}{2L\psi^{2L}} - \frac{1}{(2L+1)\psi^{2L+1}} + \mathcal{C}, &(73)\\ &= -\sum_{n=1}^{2L-1} \frac{1}{n\psi^n} + \log(\psi) - \log(1-\psi) - \frac{1}{2L\psi^{2L}} - \frac{1}{(2L+1)\psi^{2L+1}} + \mathcal{C}, &(74)\\ &= -\sum_{n=1}^{2(L+1)-1} \frac{1}{n\psi^n} + \log(\psi) - \log(1-\psi) + \mathcal{C}. &(75) \end{aligned}$$

Hence, the claimed relation (69) holds for $L + 1$ as well. An alternative proof strategy would have been based on integrating both sides of the following identity

$$\frac{1}{\psi^{2L}(1-\psi)} = \frac{1}{1-\psi} + \frac{\sum_{n=1}^{2L-1} \psi^n}{\psi^{2L}}. \tag{76}$$

This concludes the proof. □

We are now in a position to state and prove the main theorem:

Theorem B.7 (JEPA dynamics)**.** *Suppose $\psi = \frac{1}{\rho}\bar{w}^{\frac{1}{L}}$. The JEPA dynamics given by eq.* (7) *admits the following implicit solution in ψ*

$$\psi(t) = \frac{\exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}{1 + \exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}. \tag{77}$$

where $\psi(t=0) = \frac{1}{\rho}\epsilon^L$, and $\mathcal{C}$ is a constant of integration.

Theorem 3.4 (JEPA critical time)**.** *The critical time t^* in the small initialization regime $\bar{w}(t=0) = \epsilon \ll 1$ for JEPA is given by*

$$t^*_{jepa} = \frac{1}{\lambda}\sum_{n=1}^{2L-1}\frac{L}{n\rho^{2L-n-1}\epsilon^{\frac{n}{L}}} + \Theta\Big[\log(\epsilon)\Big], \tag{12}$$

as long as p is not too close, as defined in eq. (81)*, to zero or one.*

Proof. A change for variable of the form $\bar{w} = (\rho\psi)^L$ in the JEPA dynamical equation given by (66) leads to

$$\mathcal{I}_L = \int \frac{d\psi}{\psi^{2L} - \psi^{2L+1}} = \frac{\sigma^2\rho^{2L}}{L}t + \mathcal{C}. \tag{78}$$

To fix $\mathcal{C}$, we need to impose the initial condition at $t = 0$. Using lemma B.6 and noting $\psi_{|t=0} = \frac{1}{\rho}\epsilon^{\frac{1}{L}}$

$$\mathcal{C} = -\sum_{n=1}^{2L-1}\frac{\rho^n}{n\epsilon^{\frac{n}{L}}} + \frac{1}{L}\log(\epsilon) - \log(\rho) - \log(1 - \frac{1}{\rho}\epsilon^{\frac{1}{L}}). \tag{79}$$

To compute the critical time, take $\psi^* = \psi^*(t^*)$ to be the ψ value at which we measure the critical time following the definition and arguments in (11). Putting (78) and (79) together leads to

$$\frac{\sigma^2\rho^{2L}}{L}t^* = -\sum_{n=1}^{2L-1}\frac{1}{n(\psi^*)^n} + \log(\psi^*) - \log(1-\psi^*) + \sum_{n=1}^{2L-1}\frac{\rho^n}{n\epsilon^{\frac{n}{L}}} - \frac{1}{L}\log(\epsilon) + \log(\rho) + \log(1 - \frac{1}{\rho}\epsilon^{\frac{1}{L}}). \tag{80}$$

As long as ψ^* is not too close to 0 or 1 in an ϵ-dependent way, meaning $\psi^* \sim \Theta(\epsilon^\beta)$ where $\beta < \frac{1}{L}$, or $1 - \psi^* \sim \Theta\Big[\exp(-\epsilon^{-\beta})\Big]$ where $\beta < \frac{2L-1}{L}$, the first three terms on the right-hand side in the above equation are sub-leading compared to the rest of the terms. Given that $\psi^* = p^{\frac{1}{L}}$, these conditions for ψ^* translate to constraints on the scaling of p, defined in eq. (11). That is to say

$$p^{\frac{1}{L}} \sim \Theta(\epsilon^\beta) \text{ with } \beta < \frac{1}{L}, \quad 1 - p^{\frac{1}{L}} \sim \Theta\Big[\exp(-\epsilon^{-\beta})\Big] \text{ with } \beta < \frac{2L-1}{L}. \tag{81}$$

This implies the following Laurent expansion for the critical time holds

$$t^* = \sum_{n=1}^{2L-1}\frac{L\rho^{n-2L}}{n\sigma^2\epsilon^{\frac{n}{L}}} - \frac{1}{\sigma^2\rho^{2L}}\log(\epsilon) + \Theta(1). \tag{82}$$

Using $\rho = \frac{\lambda}{\sigma^2}$, the above equation can be rewritten as

$$t^*_{\text{jepa}} = \frac{1}{\lambda}\sum_{n=1}^{2L-1}\frac{L}{n\rho^{2L-n-1}\epsilon^{\frac{n}{L}}} + \Theta\Big[\log(\epsilon)\Big]. \tag{83}$$

This concludes the proof. □

Embedded in the steps of the theorem's proof is an implicit closed-form solution to the JEPA dynamical equation in (66), which for completeness we state as a theorem below:

Theorem B.7 (JEPA dynamics)**.** *Suppose $\psi = \frac{1}{\rho}\bar{w}^{\frac{1}{L}}$. The JEPA dynamics given by eq.* (7) *admits the following implicit solution in ψ*

$$\psi(t) = \frac{\exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}{1 + \exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}. \tag{77}$$

where $\psi(t=0) = \frac{1}{\rho}\epsilon^L$, and $\mathcal{C}$ is a constant of integration.

Proof. Substituting $\bar{w} = (\rho\psi)^L$ in (66), one obtains

$$\int \frac{d\psi}{\psi^{2L} - \psi^{2L+1}} = \frac{\sigma^2}{L}\rho^{2L}t + \mathcal{C}. \tag{84}$$

From lemma B.6, we have

$$-\log(1-\psi) + \log\psi + \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi^{n-2L} = \frac{\sigma^2\rho^{2L}}{L}t + \mathcal{C}, \tag{85}$$

where

$$\psi_{|t=0} = \frac{1}{\rho}\epsilon^L, \tag{86}$$

$$\mathcal{C} = -\log(1-\psi_{|t=0}) + \log(\psi_{|t=0}) + \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi_{|t=0}^{n-2L}. \tag{87}$$

which, using the definition of the regression coefficient, can be rewritten as

$$\psi(t) = \frac{\exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}{1 + \exp\left[\frac{\lambda}{L}\rho^{2L-1}t - \sum_{n=1}^{2L-1}\frac{1}{n-2L}\psi(t)^{n-2L} + \mathcal{C}\right]}, \tag{88}$$

This completes the proof. □

Now we move on to studying the MAE dynamics. It turns out, it makes sense to discuss MAE dynamics for $L = 1$ and $L > 1$ cases, separately:

B.3 MAE dynamics: proofs of theorem B.10 and theorem 3.5 for $L > 1$

Now we turn to analyzing the MAE dynamics for an arbitrary encoder depth L. Let us start with the following lemma:

Lemma B.8. *Suppose $\Omega = [\delta, u]$, where δ is a constant and u, δ belong to $(0, 1)$, then the following holds*

$$\mathcal{I}(u) = \int_\Omega \frac{d\psi}{\psi^2 - \psi^\alpha} = \frac{1}{(\alpha-2)u}\sum_{n=0}^{\infty}\frac{u^{n(\alpha-2)}}{n - \frac{1}{\alpha-2}} + \mathcal{C}, \tag{89}$$

where $\alpha > 2$ and $\mathcal{C}$ is an integration constant.

Proof. The integrand admits the following Laurent series expansion convergent within the integration domain Ω

$$\frac{1}{\psi^2 - \psi^\alpha} = \frac{1}{\psi^2} + \sum_{n=1}^{\infty}\psi^{n(\alpha-2)-2}, \tag{90}$$

therefore, the series expansion can be integrated term by term

$$\mathcal{I}(u) = -\frac{1}{u} + \sum_{n=1}^{\infty}\frac{u^{n(\alpha-2)-1}}{n(\alpha-2)-1} + C_\delta \tag{91}$$

$$= \sum_{n=0}^{\infty}\frac{u^{n(\alpha-2)-1}}{n(\alpha-2)-1} + C_\delta \tag{92}$$

$$= \frac{1}{(\alpha-2)u}\sum_{n=0}^{\infty}\frac{u^{n(\alpha-2)}}{n - \frac{1}{\alpha-2}} + C_\delta, \tag{93}$$

where C_δ is some delta-dependent constant. In passing, we also recognize that the last expression above can be written in terms of the *Lerch transcendent function* $\Phi = \Phi(z, s, a)$ for the special value of $s = 1$

$$\mathcal{I}(u) = \frac{1}{(\alpha-2)u}\Phi\left(u^{\alpha-2}, 1, -\frac{1}{\alpha-2}\right), \tag{94}$$

where $\Phi = \Phi(z, s, a)$ can be represented as a series over the complex plane

$$\Phi = \Phi(z, s, a) = \sum_{n=0}^{\infty} \frac{z^n}{(n+a)^s}. \tag{95}$$

It is convergent for $|z| < 1$ or on the unit circle $|z| = 1$ when $\Re(s) > 1$. This concludes the proof of the lemma. □

Next, we proceed to state the main theorem of this section:

Theorem 3.5 (MAE critical time)**.** *The MAE critical time t^* in the small initialization regime of $\bar{w}(t=0) = \epsilon \ll 1$ and $L > 1$ is given by*

$$t^*_{\mathit{mae}} = \frac{L}{\lambda(L-1)\epsilon^{\frac{L-1}{L}}} + \Theta(1). \tag{13}$$

as long as p is not too close, as defined in eq. (109)*, to zero or one.*

Proof. Let us make the following change of the dependent variable in (64)

$$\bar{w}(t) = \rho^{\frac{L}{L+1}} u(t)^{\frac{L}{L-1}}. \tag{96}$$

Then it can be integrated to give the following

$$\mathcal{I} = \int \frac{du}{u^2 - u^{\frac{-3L+1}{1-L}}} = -\frac{\rho^{-\frac{1-L}{L+1}}\lambda(1-L)}{L}t + \mathcal{C}, \tag{97}$$

where $\mathcal{C}$ is an integration constant. The integral $\mathcal{I}$ can be performed using lemma (B.8)

$$\mathcal{I} = \frac{L-1}{(L+1)u}\Phi\left(u^{\frac{L+1}{L-1}}, 1, \frac{1-L}{1+L}\right), \tag{98}$$

where Φ is the *Lerch transcendent function*. An initial condition needs to be imposed to fix the integration constant. Recall that

$$u_{|t=0} = \rho^{\frac{1-L}{L+1}}\epsilon^{\frac{L-1}{L}}. \tag{99}$$

Note that for $L > 1$

$$u_{|t=0} \to 0, \tag{100}$$

under the small initialization assumption where $\epsilon \to 0$. From (97)

$$\mathcal{C} = \mathcal{I}(u_{|t=0}). \tag{101}$$

Together with (100), we deduce

$$\mathcal{C} = \text{Leading Order}_{u\to 0}\Big[\mathcal{I}(u)\Big]_{|u=u_{|t=0}} + \Theta(\epsilon^r), \tag{102}$$

for some exponent r. Given

$$\frac{1}{u}\Phi(u^{\frac{L+1}{L-1}}, 1, \frac{1-L}{1+L}) = \frac{1+L}{1-L}\frac{1}{u} + \sum_{n=1}^{\infty}\frac{u^{\frac{L+1}{L-1}n-1}}{n+\frac{1-L}{1+L}}, \tag{103}$$

This allows us to write

$$\frac{\Phi}{u} = \frac{1+L}{1-L}\frac{1}{u} + \Theta(u^{\frac{2}{L-1}}). \tag{104}$$

This fixes $\mathcal{C}$ to be

$$\mathcal{C} = -\frac{1}{\rho^{\frac{1-L}{1+L}}\epsilon^{\frac{L-1}{L}}} + \Theta(\epsilon^{\frac{2}{L}}). \tag{105}$$

To compute the critical time, we start with (97) and (105)

$$\frac{\rho^{-\frac{1-L}{L+1}}\lambda(1-L)}{L}t^* = -\frac{1}{\rho^{\frac{1-L}{1+L}}\epsilon^{\frac{L-1}{L}}} - \frac{L-1}{(L+1)u^*}\Phi(u_*^{\frac{L+1}{L-1}}, 1, \frac{1-L}{1+L}). \tag{106}$$

To further proceed, we make the following observation, stated as a lemma:

Lemma B.9. *The second term on the right-hand side of* (106) *is always sub-leading compared to the first unless u_* is chosen so that $u_* \sim \Theta(\epsilon^{\beta})$ or $1 - u_* \sim \Theta\Big[\exp(-\epsilon^{-\beta})\Big]$, where $\beta > \frac{L-1}{L}$.*

Proof. Singular points of $\mathcal{I} = \mathcal{I}(u)$ are at $u = 0$ or $u = 1$. The leading term in the Taylor expansion near $u = 0$ is

$$\mathcal{I}(u) = -\frac{1}{u} + \Theta(u^{\frac{2}{L-1}}). \tag{107}$$

If $u_* \sim \Theta(\epsilon^{\beta})$ with $\beta < \frac{L-1}{L}$, then $\mathcal{I}(u_*)$ would be sub-leading compared to the first term on the right-hand side of (106). In addition, $\mathcal{I}(u)$ contains a logarithmically singular point at $u = 1$. Taylor-expanding around $u = 1^-$ gives

$$\mathcal{I} = -\frac{L-1}{L+1}\log(1-u) + \Theta(1). \tag{108}$$

If $1 - u_* \sim \Theta\Big[\exp(-\epsilon^{-\beta})\Big]$ then $\mathcal{I}(u_*) \sim \epsilon^{-\beta}$ which is sub-leading compared to the first term of the right-hand side of (106) for $\beta < \frac{L-1}{L}$. This completes the proof. □

Therefore, for any choice of u^*, unless it is too close to either of the two singular points at $u = 0$ and $u = 1$ (as quantified in the above lemma), the $\mathcal{I}$ term in (106) is sub-leading. Given $u^* = p^{\frac{L-1}{L}}$ where p is the fraction in the definition of the critical time, defined in eq. (11), the conditions on u^* become constraints on p not scaling too close to zero or one with small ϵ

$$p^{\frac{L-1}{L}} \sim \Theta(\epsilon^{\beta}) \text{ or } 1 - p^{\frac{L-1}{L}} \sim \Theta\Big[\exp(-\epsilon^{-\beta})\Big], \text{with } \beta < \frac{L-1}{L}. \tag{109}$$

This means in the limit $\epsilon \to 0$ the leading term in the Laurent expansion in ϵ of the critical time is given by

$$t^*_{mae} = \frac{L}{\lambda(L-1)\epsilon^{\frac{L-1}{L}}}. \tag{110}$$

The sub-leading $\Theta(1)$ contribution to the critical that depends on the choice of u^*

$$t^*_{mae} = \frac{L}{\lambda(L-1)\epsilon^{\frac{L-1}{L}}} + \frac{\gamma}{(L-1)\lambda} + \Theta(\epsilon^{\frac{2}{L}}), \tag{111}$$

where $\gamma = L\rho^{\frac{1-L}{1+L}}\mathcal{I}(u^*)$, which is u^*-dependent. This concludes the proof. □

Let us state one of the intermediate results (closed-form solution to MAE dynamics) in proving the above theorem:

Theorem B.10 (MAE dynamics)**.** *Suppose $L > 1$ and let $\psi = \rho^{\frac{1-L}{L+1}}\bar{w}^{-\frac{L-1}{L}}$. The MAE dynamics given by eq.* (8) *admits the following implicit solution in ψ*

$$t = -\frac{\rho^{\frac{1-L}{L+1}}\mathcal{C}L}{\lambda(L-1)} + \frac{\rho^{\frac{1-L}{L+1}}L}{\lambda(L+1)\psi(t)}\Phi\big(\psi(t)^{\frac{L+1}{L-1}}, 1, \frac{1-L}{1+L}\big) \tag{112}$$

where $\Phi := \Phi(z, a, q)$ is the Lerch transcendent function defined by

$$\Phi(z, s, a) = \sum_{n=0}^{\infty}\frac{z^n}{(a+n)^a}, \quad |z| < 1, \tag{113}$$

and $\mathcal{C}$ is an integration constant.

From eq. (112) we can extract the critical time:

Proof. To prove this theorem, note that (97) gives

$$-\frac{\rho^{-\frac{1-L}{L+1}}\lambda(1-L)}{L}t = -\mathcal{C} + \frac{L-1}{u(L+1)}\Phi\left(u^{\frac{L+1}{L-1}}, 1, \frac{1-L}{1+L}\right), \tag{114}$$

which reduces to the statement of the theorem after rearranging terms and constants, where $\mathcal{C}$ is given by (105). This proves the theorem. □

B.4 MAE: shallow encoder $L = 1$

For a single-layer encoder we have (64)

$$\frac{1}{\sigma^2}\int_{\bar{w}} \frac{d\bar{w}(t)}{\rho\bar{w}(t) - \bar{w}^3(t)} = t + \mathcal{C}, \tag{115}$$

which has a closed-form solution

$$\ln[\bar{w}^2(t)] - \ln\left(|\bar{w}^2(t) - \rho|\right) = 2\sigma^2\rho t + \mathcal{C}. \tag{116}$$

Solving for $\bar{w}(t)$ and imposing the initial condition leads to

$$\bar{w}^2(t) = (\rho - \bar{w}^2(t))\exp\left[2\lambda t + \mathcal{C}\right], \tag{117}$$

$$\bar{w}(t) = \frac{\sqrt{\rho}\exp\left[\lambda t + \mathcal{C}/2\right]}{\sqrt{1 + \exp\left[2\lambda t + \mathcal{C}\right]}}, \tag{118}$$

$$\mathcal{C} = \log\left(\frac{\epsilon^2}{\rho - \epsilon^2}\right). \tag{119}$$

Qualitatively, the dynamics of $\bar{w}(t)$ for $\epsilon \ll 1$ can be described as follows:

1. $\bar{w}(t)$ remains indistinguishable from zero until a critical time $t^* = t^\star(\epsilon)$ is reached.
2. At $t > t^\star(\epsilon)$, $\bar{w}(t)$ grows rapidly until it converges to its asymptotic value given by $\bar{w}(\infty) = \sqrt{\rho}$.

We can derive the critical time t^* defined in (11). The solution (116) can be written

$$2\lambda t^* + \log\left(\frac{\epsilon^2}{\rho - \epsilon^2}\right) = \log(\bar{w}_*^2) - \log\left(|\bar{w}_*^2 - \rho|\right) \sim \Theta(1). \tag{120}$$

The leading contribution to the critical time in the limit $\epsilon \to 0$ is

$$t^\star(\epsilon, \lambda) \approx \frac{|\log(\epsilon)|}{\lambda}. \tag{121}$$

In summary, for $L = 1$ and small initialization, learning of projections of the encoder in the standard basis occurs in a step-wise manner over the time scale t^* which is *controlled by λ only.*

B.5 Proof of corollary 3.6

Corollary B.11. *Let $t^*_{jepa}(\epsilon, \lambda, \rho), t^*_{mae}(\epsilon, \lambda, \rho)$ denote the critical time for the JEPA and MAE objectives given feature parameters ϵ, λ, ρ. WLOG assume $\rho' > \rho$, let $\Delta_\rho = \frac{1}{\rho} - \frac{1}{\rho'}$. Then, it holds that*

$$\frac{t^*_{jepa}(\epsilon, \lambda, \rho)}{t^*_{jepa}(\epsilon, \lambda, \rho')} = 1 + \frac{2L-1}{2L-2}\Delta_\rho \epsilon^{\frac{1}{L}} + \Theta(\epsilon^{\frac{2}{L}}), \quad \frac{t^*_{mae}(\epsilon, \lambda, \rho)}{t^*_{mae}(\epsilon, \lambda, \rho')} = 1 + \Theta\left[\epsilon^{\frac{L-1}{L}}\right] \tag{14}$$

Proof. The statement of the corollary is easily seen to hold if one forms JEPA/MAE critical time ratios using the results in theorem 3.4 and theorem 3.5) for some fixed λ but two values of the regression coefficient, denoted as ρ and ρ'. Taylor-expanding the resulting expression around $\epsilon = 0$ up to sufficient order proves the corollary. □

Appendix C Linear Generative Models

C.1 Random masking with isotropic noise

The framework we will use is to assume a data sample $z \in \mathbb{R}^d$ is generated as:

$$z_i = \sum_k s_k B_{ki} + \eta_i. \tag{122}$$

where η_i is noise and s_k are model coefficients for important factors (rows of B). We will also find it useful to define $\mathbb{E}\left[s^2\right] = \frac{1}{d}\sum_{k=1}^{d} \mathbb{E}\left[s_k^2\right]$ and will assume that the variance of these coefficients and noise is isotropic: $\mathbb{E}\left[\eta_i^2\right] = \mathbb{E}\left[\eta^2\right]$.

We then sample x and y from masked versions of z as:

$$x_i = z_i m_i \quad y_i = z_i(1-m_i) \quad m_i = \begin{cases} 1 & \text{with probability } f \\ 0 & \text{with probability } 1-f \end{cases} \tag{123}$$

Under assumptions that $B_{ki} \sim \mathcal{N}(0, \frac{1}{d})$ and in the high dimensional limit of infinite d, infinite sample size and some assumptions on the distribution of model coefficients (for instance that they are γ-sparse with finite γ), we can show:

Theorem C.1. *Σ^{xx} and Σ^{xy} are simultaneously diagonalizable in the basis B and have diagonal elements:*

$$\sigma_i^2 = f^2\mathbb{E}\left[s_i^2\right] + f\mathbb{E}\left[\eta^2\right] + (f-f^2)\mathbb{E}\left[s^2\right], \tag{124}$$

$$\lambda_i = f(1-f)(\mathbb{E}\left[s_i^2\right] - \mathbb{E}\left[s^2\right]). \tag{125}$$

From this, we can see that

$$\rho_i = \frac{\lambda_i}{\sigma_i^2} = \frac{(1-f)(\mathbb{E}\left[s_i^2\right] - \mathbb{E}\left[s^2\right])}{\mathbb{E}\left[s^2\right] + \mathbb{E}\left[\eta^2\right] + f(\mathbb{E}\left[s_i^2\right] - \mathbb{E}\left[s^2\right])} \tag{126}$$

We can see from this that the order in which components are learned does not depend on the masking ratio, but it does impact the final solution learned and the precise timing in which components are learned. This derivation can also be applied to show that all the parameters in a sparse basis will eventually be learned, but the weighting will depend on the sparsity and noise level of the data and which architecture and loss function we choose.

In order to derive these results, we compute the covariances and correlation terms of this model since our results in the main paper are computed in terms of these:

$$\mathbb{E}\left[x_i x_j\right] = \mathbb{E}\left[m_i m_j\right]\mathbb{E}\left[z_i z_j\right] = \left((f-f^2)\delta_{ij} + f^2\right)\mathbb{E}\left[z_i z_j\right]. \tag{127}$$

It follows that:

$$\mathbb{E}\left[xx^T\right] = f^2\mathbb{E}\left[zz^T\right] + (f-f^2)\text{Diag}(\mathbb{E}\left[zz^T\right]). \tag{128}$$

Diag($\bullet$) here is an operator removing all non-diagonal elements in the matrix $\bullet$. Similarly we can derive that $\mathbb{E}\left[m_i(1-m_j)\right] = f(1-f)(1-\delta_{ij})$ and $\mathbb{E}\left[(1-m_i)(1-m_j)\right] = f(1-f)\delta_{ij} + (1-f)^2$ to show that:

$$\mathbb{E}\left[xy^T\right] = f(1-f)\mathbb{E}\left[zz^T\right] - f(1-f)\text{Diag}(\mathbb{E}\left[zz^T\right]), \tag{129}$$

$$\mathbb{E}\left[yy^T\right] = (1-f)^2\mathbb{E}\left[zz^T\right] + (f-f^2)\text{Diag}(\mathbb{E}\left[zz^T\right]). \tag{130}$$

Note that all of these matrices have the same form: a weighted sum of the full data covariance and the diagonal of this covariance. It is helpful therefore to compute this covariance directly:

$$\mathbb{E}\left[zz^T\right]_{\epsilon,s,m} = B^T\Sigma^s B + \sigma_\epsilon^2 I = B^T\left(\Sigma^s + \sigma_\epsilon^2 I\right)B \tag{131}$$

where $\Sigma^s \in \mathbb{R}^{d\times d}$ is a diagonal with components $\Sigma_{jk}^s = \mathbb{E}\left[s_k^2\right]\delta_{jk}$. Unfortunately, the diagonal of the data covariance matrix above is not generally going to be exactly simultaneously diagonalizable with the data covariance unless it is proportional to an identity matrix. We can however show that this will occur in a certain limit. To see this consider the diagonal elements of the full data covariance:

$$\mathbb{E}\left[z_i^2\right] = b_i^T\left(\Sigma^s + \sigma_\epsilon^2\right)b_i = \sum_k B_{ki}^2(s_k^2 + \sigma_\epsilon^2) \tag{132}$$

In the limit where $\sum_k B_{ki}^2 s_k^2$ becomes a self-averaged quantity meaning that it does not fluctuate as we draw new instances of B and s from the dataset, we will have that these diagonal elements are also the same for all i and it is therefore proportional to an identity matrix.

To keep our derivation simple we will assume that the coefficients are γ-sparse, meaning $[\gamma \cdot d]$ are drawn from a distribution P_s and the rest are zero ($[\bullet]$ rounds to the nearest integer). It follows that:

$$\lim_{\gamma d \to \infty} \mathbb{E}\left[z_i^2\right] = \mathbb{E}\left[s^2\right] + \sigma_\epsilon^2 \tag{133}$$

Thus, we have simultaneous diagonalizability.

C.2 Temporal model

$$z_i^t = \sum_a u^a(t) v_i^a + \xi_i^t \tag{134}$$

This corresponds to combining a set of images $\{\mathbf{v}^a\}$ but flickering them with a time-varying intensity $u(t)$. We will assume the following properties in the large T limit and then construct appropriate functions:

$$\mathbb{E}\left[(u^a(t))^2\right] = 1, \quad \mathbb{E}\left[u^a(t)u^a(t+1)\right] = \gamma_a, \quad \mathbb{E}\left[u^a(t)\right] = 0. \tag{135}$$

$$\mathbb{E}\left[u^a(t)u^b(t)\right] = 0, \quad \mathbb{E}\left[u^a(t)u^b(t+1)\right] = 0. \tag{136}$$

This will hold for instance if we generate

$$u^a(1) = \eta_0^a, \quad u^a(t+1) = \gamma_a u^a(t) + \sqrt{1-\gamma_a^2}\eta_t^a, \quad \eta_t^a \sim \mathcal{N}(0,1). \tag{137}$$

We must ensure T is large enough to allow for good mixing of values for simulations. Note that the mixing time will depend on γ_a.

We will also assume that the noise is uncorrelated in time so that: $\mathbb{E}\left[\xi_i^t \, \xi_i^{t+1}\right] = 0$. It follows that in the large T limit:

$$C_{ij}^{xy} = \sum_a \gamma_a v_i^a v_j^a \tag{138}$$

For the purposes of this experiment, we assume the images take up different parts of the input space so that we can easily tune the level of noise that shows up in each to explore the trade-offs between temporal correlation and noise. In this case, all vectors v^a are orthogonal by definition so we immediately know that the above expression for C^{xy} is diagonal. We then assume the noise is $\xi_i \sim \mathcal{N}(0, \sigma_{a(i)}^2)$ where $a(i)$ denotes the image that pixel i belongs to.

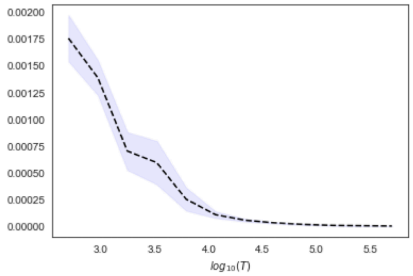

Figure 6: **Simultaneous diagonalizability demonstration**. We are in the same setting as Fig 3. Simultaneous diagonalizability is demonstrated by measured by mean of squared error in off-diagonal elements when attempting to diagonalize $\hat{\Sigma}^{xy}$ using the eigenbasis of $\hat{\Sigma}^{xx}$. The plot show 2 standard deviation with 10 runs.

We then have that

$$C_{ij}^{xx} = \sum_{a=1}^{M} v_i^a v_j^a + \sum_a \sigma_a^2 \delta_{ij} \delta_{a(i)a(j)} \tag{139}$$

Here δ_{ij} refers to the Kroneker delta function which is zero when $i \neq j$ and one otherwise. It follows that v^a are all rescaled eigenvectors and these matrices are simultaneously diagonalizable with the corresponding parameters:

$$\lambda_a = \gamma_a \|v^a\|^2, \quad a = 1, ..., M, \quad \lambda_a = 0 \text{ otherwise.} \tag{140}$$

$$\rho_a = \frac{\gamma_a \|v^a\|^2}{\sigma_a^2 + \|v^a\|^2}, \quad a = 1, ..., M, \quad \rho_a = 0 \text{ otherwise.} \tag{141}$$

Appendix D Imagenet Experiments

Our results in the paper predict that JEPA will tend to focus on the lower subspace of data variance where most of the perceptual features reside as claimed in [20]. To test this prediction in a realistic setting, we use the empirical setup described in [20] to study the differences in the feature learning dynamics of I-JEPA [3] and MAE [8]. We first describe the empirical setup and then make our observations.

Setup : We train a vision transformer (ViT) [39] encoder using I-JEPA [3] and MAE [8] on ImageNet-1K [40] data resized to 64×64 pixels. We use MAE-style masking to ensure parity with the masking function and copy all other hyperparameters described in I-JEPA [3] and MAE [8]. We build additional datasets by removing certain principal components of the full data subspace as described in [20] — **bottom** refers to the dataset that preservers bottom 25% of explained variance while **top** denotes the dataset with top 75% of explained variance. We use the original (**full**) images as well the filtered images described above and extract representations using encoders trained with I-JEPA and MAE. We then compare the representation between **full-bottom** and **full-top** via canonical correlation analysis (CCA) [41] for several checkpoints that are gathered during training.

Observe that I-JEPA's features obtained from the full and bottom datasets show higher similarity compared to MAE throughout training while the trend is reversed for full and top images. This implies that JEPA learns features from the bottom portion of the PCA space **faster** than MAE. Balestrieo and Lecun [20] suggest that the bottom portion of the spectrum contain features useful for discrimination tasks which is what JEPA focuses on during optimization.

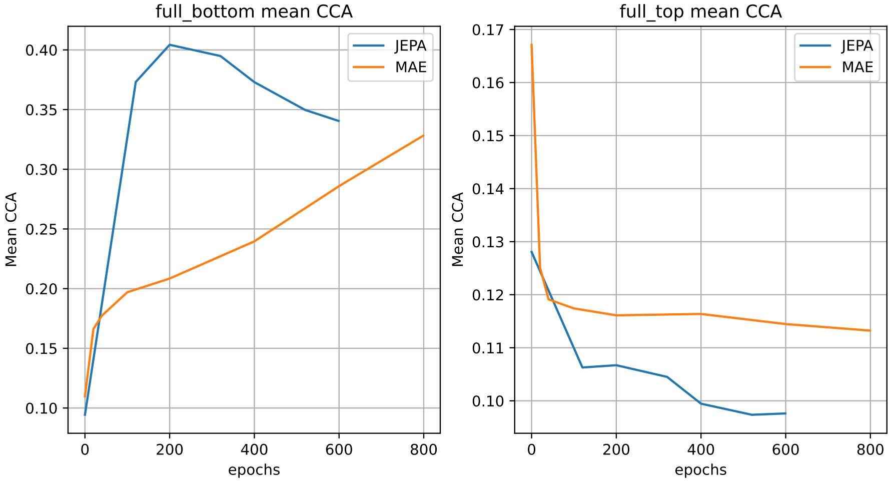

Figure 7: Feature similarity dynamics for I-JEPA vs MAE on Imagenet dataset.

Appendix E Broader Impact

This work is a theoretical contribution to shed light on techniques in self-supervised learning to try to understand their implicit biases. The analyses reveal differences in the kinds of features that get learned, which can potentially be helpful to practitioners in selecting which technique to use when. These benefits may be important in deriving better models for a variety of downstream tasks that people care about such as classification, planning, and decision-making.