PEIRA: Learning Predictive Encoders through Inter-View Regressor Alignment

Michael Arbel*
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK
michael.arbel@inria.fr

Basile Terver*
Ecole Normale Supérieure / PSL
Inria Paris
basile.terver@ens.psl.edu

Jean Ponce
Ecole Normale Supérieure / PSL
New York University
jean.ponce@ens.fr

Abstract

Non-contrastive self-supervised learning (SSL) is an effective framework for predictive representation learning, but popular (and in practice effective) methods such as SimSiam, BYOL, I-JEPA or DINO, which rely on a form of self-distillation to train a teacher-student network, remain poorly understood as they typically do not minimize a well-defined objective. We analyze the dynamics of a variant of the Joint Embedding Predictive Architecture (JEPA) using a regularized linear regressor to predict the learned representations of two views of the data from one another, and fully characterize its stability: non-collapsed stable equilibria align with leading nonlinear canonical correlation subspaces, while collapsed equilibria may also be stable attractors. Motivated by this result, we introduce PEIRA, a non-contrastive SSL method with an explicit objective defined through the trace of the optimal linear regressor. We show that its only stable equilibria are nontrivial global minimizers and recover the same canonical correlation subspaces, with regularization selecting the effective dimension. Experiments on ImageNet-1K and CIFAR-10 show PEIRA is competitive with VICReg and LeJEPA baselines, and qualitative empirical results support the theory. Code: https://github.com/MichaelArbel/peira-loss (loss implementation), https://github.com/Basile-Terv/PEIRA (experiment reproduction).

1 Introduction

Non-contrastive self-supervised learning has emerged as a successful approach to representation learning, achieving strong empirical performance across vision and video understanding [Grill et al., 2020, Caron et al., 2021, Chen and He, 2021, Bardes et al., 2022, 2024, Assran et al., 2025]. In this setting, one learns encoders of different views of the same data so the corresponding features predict each other using a learned inter-view regressor, thereby avoiding the need to mine negative examples as in *contrastive* methods [Chen et al., 2020, He et al., 2020]. These approaches are, however, susceptible to *representational collapse*, in which the learned encoders and predictor respectively degenerate to constants and the identity [LeCun, 2022]. A large body of work has studied and sought to prevent collapse. Among non-contrastive methods, one can distinguish *regularization-based*

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

approaches, such as Barlow Twins [Zbontar et al., 2021], VICReg [Bardes et al., 2022], and LeJEPA [Balestriero and LeCun, 2025], from *self-distillation* methods, such as SimSiam [Chen and He, 2021], BYOL [Grill et al., 2020], DINO [Caron et al., 2021, Oquab et al., 2023, Siméoni et al., 2025], I-JEPA [Assran et al., 2023] and V-JEPA [Bardes et al., 2024]. The latter use mechanisms such as stop-gradient and exponential moving averages, leading to dynamics that differ from standard gradient descent and may avoid collapse [Tian et al., 2021]. Regularization-based methods, like VICReg or Barlow Twins, have been connected to the recovery of canonical correlation subspaces [Balestriero and LeCun, 2022, Chapman et al., 2024], providing a useful interpretation of the representations they learn. By contrast, *self-distillation* methods remain poorly understood despite their empirical success: beyond restricted analyses [Tian et al., 2021, Tang et al., 2023, Littwin et al., 2024, Ponce et al., 2026], the learned representations and their collapse behavior are still not fully characterized.

In this work, we close this gap for a self-distillation variant of JEPA dynamics that learns nonlinear encoders coupled with an optimal regularized linear predictor. We establish, to our knowledge, the first rigorous connection between these dynamics and nonlinear canonical correlation analysis (CCA) [Michaeli et al., 2015]. The dynamics converge to representations spanning a nonlinear CCA subspace, with effective dimension selected implicitly by regularization. We characterize stable equilibria, identifying those that encode predictive structure and those that collapse. This extends prior analyses of non-contrastive self-distillation, which are restricted to linear encoders and assumptions such as isotropic covariance or orthogonal initialization [Tian et al., 2021, Tang et al., 2023, Littwin et al., 2024]. The characterization motivates a new unconstrained objective, *Predictive Encoders by Inter-View Regressor Alignment* (**PEIRA**), for JEPA-style learning, with the same statistical target that rules out collapse as a stable solution, both theoretically and empirically. PEIRA aligns encoders by maximizing the trace of the best regularized linear predictor between views while controlling feature scale. We analyze its dynamics and equilibria, showing that it recovers the same nonlinear CCA subspaces as the considered self-distillation dynamics, again with regularization-induced spectral selection, while excluding collapse through the objective rather than training heuristics. Finally, we give a scalable stochastic implementation for parametric encoders inspired by stochastic compositional optimization [Wang et al., 2017], which avoids differentiating through the optimal predictor and achieves competitive performance on SSL image representation learning. Specifically, the main contributions of this paper are threefold:

1. We present a full theoretical characterization of the dynamics of JEPA-style self-distillation learning in terms of canonical correlation analysis (CCA) when the predictor is linear and the encoders are arbitrary square-integrable functions. This includes identifying all equilibria of the corresponding dynamics **(Proposition 4.1)** and characterizing the stable ones, which include the collapsed zero solution **(Proposition 4.2)**.

2. We introduce PEIRA, a novel approach to self-supervised learning with a well-defined unconstrained objective. We characterize the equilibria of its dynamics and show that PEIRA recovers the same CCA subspaces as self-distillation **(Proposition 4.3)**. Unlike self-distillation, however, all stable equilibria of PEIRA are nontrivial global minimizers of the objective **(Proposition 4.4)**.

3. We propose a scalable stochastic training procedure for PEIRA **(Algorithm 1)**, for parametric encoders (in practice neural networks), and demonstrate in Section 6 that it delivers competitive performance on ImageNet-1K with respect to baselines like VICReg and LeJEPA on ImageNet-1K and CIFAR-10 benchmarks. Experiments on these benchmarks also confirm the theoretical claims about spectral alignment, show strong correlation between the PEIRA objective and downstream performance, and robustness of the method to variations in hyperparameter values.

2 Related work

Canonical correlation analysis and self-supervised learning. CCA is the classical framework for extracting maximally correlated components across paired views under a whitening constraint [Hardoon et al., 2004, Andrew et al., 2013, Fukumizu et al., 2007]. Recently, several SSL methods have incorporated CCA-inspired structure to prevent collapse. Whitening-based losses [Ermolov et al., 2021] and their analysis [Huang et al., 2024] do so through explicit decorrelation, while penalized relaxations [Zhang et al., 2021, Chang et al., 2018] improve practicality but lose exact CCA recovery guarantees. Chapman et al. [2024] proposed an unconstrained objective for CCA with provable recovery in the linear setting, and proposed to use it for SSL. However, its quartic (4th

order polynomial) dependence on the representation can make optimization prone to instabilities [Shi et al., 2023, De Sa et al., 2015]. A few works also established connections between some existing SSL methods and CCA. Balestriero and LeCun [2022] show that VICReg and Barlow Twins recover spectral embedding methods, while Chapman et al. [2024] further rigorously establish that VICReg recovers CCA subspaces, in the linear setting, and show that it can still collapse. Similarly, Lyu et al. [2021] show that maximizing correlation under decorrelation constraints recovers shared components, and discuss similarities with methods such as BYOL and Barlow Twins. However, these works stop short of providing a formal CCA recovery theory for self-distillation methods.

Self-distillation methods for non-contrastive SSL. Theory for these methods remains confined to highly simplified regimes, typically linear models with isotropic data or orthogonal initialization [Tian et al., 2021, Tang et al., 2023, Littwin et al., 2024]. Tang et al. [2023] have further shown that, provided the training procedure converges, the corresponding equilibrium spans a linear eigenspace of the cross-covariance matrix. Related analyses with shallow neural networks still rely on simplified data models [Wen and Li, 2022]. Extending these results is difficult: even in the linear case, relaxing isotropy leads to hard-to-characterize equilibria and unclear convergence, since the dynamics are not the gradient flow of any global objective [Ponce et al., 2026].

3 From Self-Distillation to PEIRA

Following the JEPA approach to self-supervised representation learning, we are interested in learning encoders of different views of the data such that the encoded views predict each other well. In general, this problem can be formulated as follows: Given a distribution $\mathbb{P}$ of data pairs (x, y) (the views) in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, the goal is to learn square-integrable and in general *nonlinear* encoders $U : \mathbb{R}^{d_1} \to \mathbb{R}^k$ and $V : \mathbb{R}^{d_2} \to \mathbb{R}^k$ for views x and y of the data using a predictive objective. We consider a regularized *symmetric*[1] objective which involves a *linear* prediction operator $P : \mathbb{R}^k \to \mathbb{R}^k$ represented as a $k \times k$ matrix

$$\mathcal{P}(P, U, V) = \mathcal{P}_{U \triangleright V}(P, U, V) + \mathcal{P}_{V \triangleright U}(P, V, U) + \frac{\lambda}{2} \|P\|_F^2 , \tag{1}$$

where $\|P\|_F$ is the Frobenius norm of P, whereas $\mathcal{P}_{U \triangleright V}$ and $\mathcal{P}_{V \triangleright U}$ are directed objectives encouraging P to linearly predict one encoded view from the other. Specifically, defining $\|U\|_2^2 = \mathbb{E}[\|U(x)\|_2^2]$ and $\|V\|_2^2 = \mathbb{E}[\|V(y)\|_2^2]$, we define these objectives as follows

$$\begin{aligned} \mathcal{P}_{U \triangleright V}(P, U, V) &= \frac{1}{2} \mathbb{E}\left[\|PU(x) - V(y)\|^2 \right] + \frac{\lambda}{2} \|U\|_2^2, \\ \mathcal{P}_{V \triangleright U}(P, V, U) &= \frac{1}{2} \mathbb{E}\left[\|PV(y) - U(x)\|^2 \right] + \frac{\lambda}{2} \|V\|_2^2. \end{aligned} \tag{2}$$

where expectations are taken w.r.t. the data distribution $\mathbb{P}$. Here, the symbol $U \triangleright V$ means that V is treated as a *target/teacher* view for prediction, whereas U is a *source/student* view. Throughout, *self-distillation* refers to updates in which the teacher view is detached by a stop-gradient or tracked by an EMA; *collapse* means that the encoded features span a subspace of lower dimension than achievable, either totally ($U = V = 0$) or partially; and the *effective rank* of a feature covariance is the exponential of the entropy of its normalized spectrum [Garrido et al., 2023a]. Ideally, we would like to learn P, U and V by minimizing $\mathcal{P}$ w.r.t. these variables. Following [Tang et al., 2023], we consider the fast-predictor regime where the regressor is optimized much faster than the encoders. This regime will allow us to provide a precise characterization of equilibria of the dynamics under consideration, by exploiting a closed form expression of the regressor. Namely, for each pair of encoders (U, V), the regressor is set to its optimal value,

$$P_{U,V} = \arg \min_{P \in \mathbb{R}^{k \times k}} \mathcal{P}(P, U, V), \tag{3}$$

which can be computed in closed form as the solution of a ridge regression problem. Thus, minimizing $\mathcal{P}$ w.r.t. P, U and V is equivalent to minimizing w.r.t. U and V the *residual* objective function

$$\mathcal{P}_{\text{res}}(U, V) = \mathcal{P}_{U \triangleright V}(P_{U,V}, U, V) + \mathcal{P}_{V \triangleright U}(P_{U,V}, V, U) + \frac{\lambda}{2} \|P_{U,V}\|_F^2 . \tag{4}$$

[1] Symmetrized objectives were also considered in Tang et al. [2023], Grill et al. [2020], and naturally accommodate views in different spaces. There is no loss of generality due to symmetrization and all our results can be adapted to asymmetric objectives, provided the views are interchangeable.

Optimizing either $\mathcal{P}$ or $\mathcal{P}_{\text{res}}$ is known to be prone to trivial solutions, e.g., U and V and P are all equal to 0, leading to representation collapse. To address this, self-distillation methods such as SimSiam, BYOL, V-JEPA or DINO update the encoders using a so-called *semi-gradient* [Sutton and Barto, 2018, Chapter 9.3], that is, a proxy for the actual gradient obtained by applying either a stop-gradient operation or exponential moving averages to the *teacher* view. In this way, the *student* view is trained to predict the *teacher* view, rather than both views being updated through the full gradient of the objective. Although this heuristic works well empirically, the learned representations remain poorly understood. In Section 4.1, we show that the self-distillation dynamics derived from the objective $\mathcal{P}_{\text{res}}$ recovers nonlinear CCA subspaces, but is still prone to (partial) collapse.

We propose instead to minimize with respect to U and V a slightly different objective which also depends on the optimal regressor $P_{U,V}$ but which *is guaranteed to* prevent collapse. Let us first note that $P_{U,V}$ can be written in closed form as a matrix-valued *signal-to-noise ratio* (SNR):

$$P_{U,V} = \Sigma_{U,V}(N_{U,V} + \lambda I)^{-1}, \text{ where } \begin{cases} \Sigma_{U,V} = \mathbb{E}[V(y)U(x)^\top + U(x)V(y)^\top], \\ N_{U,V} = \mathbb{E}[U(x)U(x)^\top + V(y)V(y)^\top], \end{cases} \tag{5}$$

where the matrix $\Sigma_{U,V}$ measures shared cross-view signal, while $N_{U,V}$ measures within-view variability. The trace of the predictor appears as a natural scalar summary of the SNR as it sums the contributions of different eigen-modes of $P_{U,V}$ and maximizing it should favor representations that share most information between views. Thus, instead of $\mathcal{P}_{\text{res}}$, we propose to minimize the **PEIRA (for Predictive Encoders through Inter-view Regressor Alignment)** objective:

$$\mathcal{E}_{\text{PEIRA}}(U, V) = -\frac{1}{2}Tr\big(P_{U,V}\big) + \frac{\lambda}{2}\left(\|U\|_2^2 + \|V\|_2^2\right). \tag{6}$$

When both encoders are linear, minimizing the negative trace of $P_{U,V}$ is closely related to ratio-trace optimization for dimensionality reduction [Wang et al., 2007, Jia et al., 2009] discussed in Appendix A. The quadratic regularization on U and V controls feature scale, preventing the negative trace term from being decreased merely by enlarging their norms. The trace of the regressor can be written as the Frobenius scalar product between the symmetric matrices $\Sigma_{U,V}$ and $(N_{U,V} + \lambda I)^{-1}$. Therefore, minimizing $\mathcal{E}_{\text{PEIRA}}$ should also encourage an *alignment* between the eigenbases of the signal and noise matrices $\Sigma_{U,V}$ and $N_{U,V}$. Such an alignment has been shown to arise implicitly for some self-distillation methods with linear encoders [Tian et al., 2021].

4 Canonical structure and stability of the learning dynamics

We characterize the equilibria of the *self-distillation dynamics* arising from the objective $\mathcal{P}_{\text{res}}$ in Equation (4) and the critical points of PEIRA introduced in Equation (6) in terms of nonlinear CCA.

Nonlinear CCA identifies pairs of functions of the two views that are maximally correlated after normalization by whitening constraint (see Appendix A for a formal definition). These directions capture the shared signal between the views, whereas directions with small canonical correlation correspond to signal components that are only weakly correlated across views. To make the characterizations in this section independent of any particular model architecture, we do not impose parametric restrictions on the encoders. Instead, we work directly in the natural infinite dimensional function spaces of square-integrable encoders, $\mathcal{U} = L^2(\mathbb{P}_X; \mathbb{R}^k)$, $\mathcal{V} = L^2(\mathbb{P}_Y; \mathbb{R}^k)$, where $\mathbb{P}_X$ and $\mathbb{P}_Y$ denote the marginals of the joint distribution $\mathbb{P}$. Parametric models, such as deep networks, are introduced only later as function approximators in these spaces (see the *capacity hypothesis* in [Huh et al., 2024]). We use the following spectral form of nonlinear CCA throughout the analysis.

Definition 4.1 (Nonlinear Canonical Correlations)**.** *The distribution $\mathbb{P}$ is said to admit a nonlinear canonical correlation decomposition, in the sense of Michaeli et al. [2015], if there exist orthonormal families $(\psi_i)_{1\le i\le R}$ and $(\xi_i)_{1\le i\le R}$ of $L^2(\mathbb{P}_X)$ and $L^2(\mathbb{P}_Y)$, and a non-increasing sequence $(c_i)_{1\le i\le R} \subset (0, 1]$, where R may be infinite, such that $\mathbb{E}[\psi_i(x)\xi_i(y)] = c_i$ and $\mathbb{E}[\psi_i(x)\xi_j(y)] = 0$ for $i \neq j$. Moreover, the families exhaust all cross-correlated directions: any square-integrable function of x orthogonal to all the ψ_i's is uncorrelated with every square-integrable function of y, and conversely for functions of y orthogonal to all the ξ_i's. The functions ψ_i and ξ_i are called the nonlinear canonical directions, and the scalars c_i are the canonical correlations, which quantify the strength of the shared signal along the i-th direction.*

4.1 Characterizing the self-distillation equilibria and their stability

We study the continuous-time self-distillation dynamics arising from the residual objective $\mathcal{P}_{\text{res}}$ in Equation (4), as it allows us to use tools from dynamical systems theory. Specifically, we consider the learning dynamics arising from applying a stop-gradient[2] operation to the *teacher* views in $\mathcal{P}_{\text{res}}$ before differentiating it, as done in SimSiam [Chen and He, 2021]. Such a procedure is mathematically equivalent to the following continuous-time dynamics defined in the non-parametric space $\mathcal{U} \times \mathcal{V}$:

$$\dot{U}_t = -\partial_U \mathcal{P}_{U \triangleright V}(P_{U_t,V_t}, U_t, V_t), \qquad \dot{V}_t = -\partial_V \mathcal{P}_{V \triangleright U}(P_{U_t,V_t}, U_t, V_t), \tag{7}$$

where $\partial_U \mathcal{P}_{U \triangleright V}$ (resp. $\partial_V \mathcal{P}_{V \triangleright U}$) denotes the partial derivatives of $\mathcal{P}_{U \triangleright V}$ (resp. $\mathcal{P}_{V \triangleright U}$) w.r.t. U (resp. V). The resulting flow extends the one studied by Tang et al. [2023] for linear encoders to nonparametric ones, better capturing the rich representation classes used in practice.

The vector field defined in Equation (7) is generally not the gradient of any objective [Ponce et al., 2026], making convergence harder to characterize without additional structure. Our first result establishes the existence of a Lyapunov function for the flow, i.e., a non-increasing function along such a flow, that confers a gradient-like structure to the system [Chill et al., 2009]. This, in turn, allows us to leverage the abstract Lojasiewicz–Simon framework for establishing convergence to equilibria [Chill, 2003, Bolte et al., 2014]. We identify such a function as a polynomial of the encoders expressed in terms of the signal and noise matrices $\Sigma_{U,V}$ and $N_{U,V}$ defined in Equation (5):

$$\mathcal{L}(U,V) = \frac{1}{3} Tr\big((N_{U,V} + \lambda I)^3\big) - \frac{1}{2} Tr\big(\Sigma_{U,V}^2\big)\,. \tag{8}$$

We further show that these equilibria admit an explicit characterization in terms of CCA directions. The following proposition formalizes this result under a standard regularity assumption on the data for non-linear CCA [Michaeli et al., 2015], stated in Appendix A.1 and ensuring the data distribution admits a non-linear canonical correlation decomposition.

Proposition 4.1 (Identification of the equilibria)**.** *Under (A) of Appendix A.1 and assuming that $c_1, \dots, c_R$ are all distinct, for any initial condition $(U_0, V_0) \in \mathcal{U} \times \mathcal{V}$, the dynamics in Equation (7) admit the function $\mathcal{L}$ in Equation (8) as a Lyapunov function, and necessarily converge to an equilibrium point of the form:*

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{\sigma=\pm} \sum_{i \in D^\sigma} f_{\epsilon_i^\sigma}(c_i, \lambda) \psi_i(x) \mathbf{q}_\mathbf{i}^\sigma, \quad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{\sigma=\pm} \sum_{i \in D^\sigma} \rho_\sigma f_{\epsilon_i^\sigma}(c_i, \lambda) \xi_i(y) \mathbf{q}_\mathbf{i}^\sigma \tag{9}$$

where $D^\pm$ are any subsets of $\{1, \dots, R\}$ with $|D^+| + |D^-| \leq k$, $(\mathbf{q}_i^\sigma)_{\sigma \in \{+,-\}, i \in D^\sigma}$ are any orthonormal vectors in $\mathbb{R}^k$, ϵ_i^σ take any values in $\in \{-1, 1\}$, $\rho_\pm = \pm 1$, and f_ϵ is the spectral filter:

$$f_\epsilon(c, \lambda) = \frac{1}{2}\left(|c| + \epsilon\sqrt{c^2 - 4\lambda}\right) \mathbb{1}_{|c|^2 - 4\lambda \geq 0}. \tag{10}$$

The filter f_ϵ thresholds canonical directions according to the regularization level λ: only directions with sufficiently large canonical correlation can appear at an equilibrium. Thus, the equilibria span arbitrary subspaces generated by at most k canonical directions. They may include positively or negatively correlated components, corresponding to the two sums in Equation (9); these signs affect the eigenvalues of the signal matrix $\Sigma_{U^\star,V^\star}$, but not the underlying canonical subspace. Since the index sets D^+ and D^- are arbitrary, Proposition 4.1 alone does not imply recovery of the leading canonical directions. To determine which subspaces the dynamics select, the next result characterizes their stable equilibria, where an equilibrium $w^\star$ of a dynamical system $\dot{w}_t = -F(w_t)$ is *stable* if the Jacobian $\nabla F(w^\star)$ has no eigenvalue with strictly negative real part, and *unstable* otherwise.

Proposition 4.2 (Stability)**.** *Under the assumptions of Proposition 4.1, the stable equilibria are:*

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{\sigma=\pm} \sum_{i=1}^{r^\sigma} f_1(c_i, \lambda) \psi_i(x) \mathbf{q}_\mathbf{i}^\sigma, \quad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{\sigma=\pm} \sum_{i=1}^{r^\sigma} \rho_\sigma f_1(c_i, \lambda) \xi_i(y) \mathbf{q}_\mathbf{i}^\sigma \tag{11}$$

where r^+ and r^- are any integers in $[0, R]$ and satisfying $0 \leq r^+ + r^- \leq k$, with the convention that a sum of $r^\pm = 0$ elements equals 0, $(\mathbf{q}_i^\sigma)_{\sigma \in \{+,-\}, i \in \{1,\dots,r^\sigma\}}$ are any orthonormal vectors in $\mathbb{R}^k$, and $\rho_\pm = \pm 1$. In particular, the totally collapsed equilibrium $U^\star = 0, V^\star = 0$ is stable.

[2]While we focus on stop-gradient rather than exponential-moving-average (EMA) dynamics, our results also characterize the equilibria of the corresponding EMA dynamics as recovering CCA directions, since the two dynamics admit the same equilibria [Ponce et al., 2026]. We leave a convergence and stability analysis of EMA dynamics to future work, and conjecture that the proof techniques developed here may extend to that setting.

Proposition 4.2 shows that collapse is not ruled out by the self-distillation dynamics: the trivial equilibrium is stable, and so are certain low-rank equilibria. Non-collapsed solutions are stable equilibria spanning the largest possible number of independent canonical directions. Specifically, let r^* be the largest index with $f_1(c_{r^*}, \lambda) > 0$ and $r^* \leq \min(k, R)$. Then non-collapsed solutions are equilibria of the form in Equation (9) with $r^+ = r^*$ or $r^- = r^*$. Stable equilibria with $r^+, r^- < r^*$ are partially collapsed, as they do not span a maximal CCA subspace. In Section 4.2, we show that PEIRA yields equilibria with a similar form but without stable collapse.

Sketch of proofs. The general proof strategy and key results are provided in Appendix B with the full proof provided in Appendix C.2.1. We show that the gradient of the Lyapunov function is well aligned with the vector field in Equation (7), making it a *gradient-like* system [Chill et al., 2009]. This property allows us to leverage the framework of Bolte et al. [2014] based on the Lojasiewicz–Simon gradient inequality [Lojasiewicz, 1982, Feehan and Maridakis, 2020] to establish convergence in Proposition B.3. The characterization and stability of equilibria then follow from the spectral decomposition of the Jacobian of the vector field at equilibrium, derived in Proposition B.7, together with the classification of equilibria established in Proposition B.9.

4.2 Characterizing the PEIRA equilibria and their stability

We now study the gradient flow of the PEIRA objective and characterize its critical points and their stability. The first result shows that the flow converges to critical points with the same canonical-subspace structure as the self-distillation equilibria.

Proposition 4.3 (Identification of equilibria)**.** *Under the same assumptions as Proposition 4.1, the gradient flow dynamics* $(\dot{U}_t, \dot{V}_t) = -\nabla_{U,V}\mathcal{E}_{\textit{PEIRA}}(U_t, V_t)$ *with any initial condition* $(U_0, V_0) \in \mathcal{U} \times \mathcal{V}$ *converges to a critical point* $(U^\star, V^\star)$ *of* $\mathcal{E}_{\textit{PEIRA}}$ *of the form:*

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{i \in D} g(c_i, \lambda)\psi_i(x)\mathbf{q_i}, \qquad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{i \in D} g(c_i, \lambda)\xi_i(y)\mathbf{q_i}, \tag{12}$$

where D *is any subset of* $\{1, \dots, R\}$ *with* $|D| \leq k$, $(\mathbf{q_i})_{i \in D}$ *are any orthonormal vectors in* $\mathbb{R}^k$*, and* g *is the spectral filter given by* $g(c, \lambda) = \max(\sqrt{c} - \lambda, 0)^{\frac{1}{2}}$.

The critical points in Proposition 4.3 share the canonical form of the self-distillation equilibria, but use a different spectral filter and contain only positively correlated components. Here, the filter acts as a soft threshold on the canonical correlations. Direct verification shows that $\Sigma_{U^\star,V^\star}$ and $N_{U^\star,V^\star}$ have the same eigenvectors, satisfying the alignment condition discussed in Section 3. Since D is arbitrary, this characterization alone does not ensure recovery of the leading canonical directions. The next proposition shows that the stable equilibria are exactly the ones spanning the largest maximally correlated subspaces, and that they are global minimizers of $\mathcal{E}_{\text{PEIRA}}$.

Proposition 4.4 (Stability)**.** *Under the same assumptions as Proposition 4.1, let* $0 < \lambda < \sqrt{c_1}$ *and* $1 \leq r_{\max} \leq \min(k, R)$ *be the largest integer such that* $g(c_{r_{\max}}, \lambda) > 0$*, with* $(c_i)_{1 \leq i \leq R}$ *from Definition 4.1. Then the only stable equilibria are global minimizers of* $\mathcal{E}_{\textit{PEIRA}}$ *and take the form:*

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{i=1}^{r_{\max}} g(c_i, \lambda)\psi_i(x)\mathbf{q_i}, \qquad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{i=1}^{r_{\max}} g(c_i, \lambda)\xi_i(y)\mathbf{q_i}, \tag{13}$$

where $(\mathbf{q_i})_{i=1}^{r_{\max}}$ *are any orthonormal vectors in* $\mathbb{R}^k$*. Any other critical point of* $\mathcal{E}_{\textit{PEIRA}}$ *is a strict saddle point, and thus unstable. Moreover, the optimal value of the objective is:*

$$\min_{U,V \in \mathcal{U} \times \mathcal{V}} \mathcal{E}_{\textit{PEIRA}}(U, V) = -\frac{1}{2} \sum_{i=1}^{r_{\max}} (\sqrt{c_i} - \lambda)^2. \tag{14}$$

Proposition 4.4 implies that all critical points of $\mathcal{E}_{\text{PEIRA}}$ are either global minimizers or strict saddles, ruling out spurious local minima. Thus, the gradient flow generically[3] converges to a global minimizer spanning the top-$r_{\max}$ canonical subspace. Consequently, stable collapse does not occur whenever $r_{\max} \geq 2$: any critical point spanning a lower-dimensional canonical subspace is unstable. In particular, in contrast to the self-distillation dynamics of Section 4.1, the trivial collapsed representation

[3]Dynamics do not converge *generically* towards unstable equilibria or strict saddle points, that is, this only occurs for particular initializations [Lee et al., 2016, Truong, 2020].

$U = 0, V = 0$ is unstable for PEIRA as it does not reach the optimal value in Equation (14). The effective dimension $r_{\max}$ is selected by λ and is capped by k. Thus, when the target dimension is unknown, one can choose k conservatively large and let λ determine the number of active canonical directions. The condition $\lambda < \sqrt{c_1}$ is necessary: for $\lambda \geq \sqrt{c_1}$, the condition $\sqrt{c_i} > \lambda$ filters out all directions, so that $r_{\max} = 0$ and the trivial solution is the unique stable equilibrium.

Sketch of proofs. Full proofs are provided in Appendix C.1.1. The convergence part follows from the general convergence result for *gradient-like* systems provided as in the case of Proposition 4.1. The stability analysis then relies on the eigen-decomposition of the Hessian at critical points, derived in Proposition B.7, together with the resulting classification of all critical points in Proposition B.8.

5 Optimizing PEIRA's objective

5.1 An auxiliary objective for gradient computation

Computing the gradient of $\mathcal{E}_{\text{PEIRA}}$ requires differentiating through the dependence of the optimal regressor $P_{U,V}$ on (U, V) in Equation (5). In practice, closed-form mini-batch estimates of the regressor, differentiated through by backpropagation, can lead to high-variance gradient estimates and unstable training [Xu et al., 2020, Galashov et al., 2025].

Instead, we derive an auxiliary objective that has the same gradient as $\mathcal{E}_{\text{PEIRA}}$ but which needs only estimating the regressor $P_{U,V}$ without differentiating it. This enables averaging techniques, such as exponential moving averages (EMA), for estimating $P_{U,V}$, which naturally reduce stochastic fluctuations and can improve training stability [Morales-Brotons et al., 2024].

Proposition 5.1. *The objective $\mathcal{E}_{PEIRA}$ is continuously differentiable. Its gradient is L-Lipschitz with $L = \frac{4}{\lambda}$ for any $0 < \lambda < 1$. Moreover, for any elements U and V in $\mathcal{U}$ and $\mathcal{V}$, the gradient satisfies $\nabla \mathcal{E}_{PEIRA}(U, V) = \partial_{(U,V)} \mathcal{L}_{\text{aux}}(U, V; P^*, Q^*)$, where $P^* = P_{U,V}$, $Q^* = (N_{U,V} + \lambda I)^{-1}$, with $P_{U,V}$ and $N_{U,V}$ defined in Equation (5), whereas $\mathcal{L}_{\text{aux}}$ is the following auxiliary objective:*

$$\begin{aligned} \mathcal{L}_{\text{aux}}(U, V; P, Q) :=& \frac{1}{2}\mathbb{E}\left[U(x)^\top Q\left(PU(x) - V(y)\right) + V(y)^\top Q\left(PV(y) - U(x)\right)\right] \\ &+ \frac{\lambda}{2}\mathbb{E}\left[\|U(x)\|^2 + \|V(y)\|^2\right]. \end{aligned} \tag{15}$$

Proposition 5.1 ensures the gradient is $\frac{4}{\lambda}$-Lipschitz whenever $0 < \lambda < 1$, thus supporting gradient-based optimization and informs the learning-rate choice [Nesterov, 2003]. The gradient identity follows by computing the expression of the gradient of $\mathcal{E}_{\text{PEIRA}}$ and noticing that it matches the partial derivative w.r.t. (U, V) of the function $\mathcal{L}_{\text{aux}}$ of 4 variables U, V, P and Q, when the last two variables are set to P^* and Q^*. Next, we describe a stochastic algorithm leveraging such an identity.

5.2 A stochastic compositional algorithm for PEIRA

Optimization over $\mathcal{U}$ and $\mathcal{V}$ is infeasible and statistically ill-posed with finite data. In practice, one restricts to parametric encoders $U_\theta : \mathcal{X} \to \mathbb{R}^k$ and $V_\psi : \mathcal{Y} \to \mathbb{R}^k$, typically neural networks. While the guarantees in Propositions 4.3 and 4.4 do not directly apply, empirical scaling laws suggest that increased model size, data, and compute improve objective optimization [Hestness et al., 2017, Kaplan et al., 2020]. We therefore use this parametrized setting as a scalable approximation to the population problem studied above, and discuss the resulting gap in Section 7.

Using the gradient expression in Proposition 5.1, we derive Algorithm 1 for learning encoders from paired data $\mathcal{D} = \{(x, y)\}$. The algorithm exploits that $P^\star$ and $Q^\star$ are nonlinear functions of the population uncentered cross-covariance and auto-covariance of the features to maintain EMA estimates of these quantities that reduce gradient variance, in analogy with stochastic compositional gradient methods for objectives involving nonlinear functions of expected values [Wang et al., 2017]. Figure G.1 illustrates the resulting gradient estimator.

Table 1: Linear-probe top-1 accuracy (%) after pretraining a ResNet-50 for 100 epochs on ImageNet-1K and a ResNet-18 for 1000 epochs on CIFAR-10. VICReg, SIGReg, and PEIRA numbers are mean ± one-sigma sample standard deviation over 3 seeds, run with the same augmentations, optimizer and probe; SIGReg uses a narrower projector (hidden width 2048 vs. 8192). BYOL and SimSiam ImageNet-1K numbers are quoted from Table 4 of Bardes et al. [2022] (100-epoch ResNet-50 reimplementation under each method's original recipe). The SimCLR ImageNet-1K number is quoted from Table S3 of Garrido et al. [2023a] (100-epoch ResNet-50, 8192-8192-2048 projector). Transfer: mean linear-probe accuracy over the 8 LeJEPA datasets (VICReg, SIGReg: seed 1; PEIRA: mean ± sample standard deviation over 3 seeds; Appendix G.10). k: PEIRA projector output dimension. Best controlled result per column in bold. "–": not run or no comparable published result.

Method	Approach	ImageNet-1K ↑	CIFAR-10 ↑	Transfer ↑
BYOL [Grill et al., 2020]	self-distillation (EMA)	69.3	–	–
SimSiam [Chen and He, 2021]	self-distillation (stop-gradient)	67.9	–	–
SimCLR [Chen et al., 2020]	contrastive (negatives)	68.5	–	–
VICReg [Bardes et al., 2022]	optimization	68.81 ± 0.09	90.92 ± 0.14	**72.07**
SIGReg [Balestriero and LeCun, 2025]	optimization	66.26 ± 0.28	$\mathbf{92.34 \pm 0.21}$	68.27
PEIRA (k=256, ours)	optimization	66.45 ± 0.11	91.36 ± 0.12	70.89 ± 0.18
PEIRA (k=1024, ours)	optimization	$\mathbf{69.17 \pm 0.06}$	91.68 ± 0.18	71.09 ± 0.24

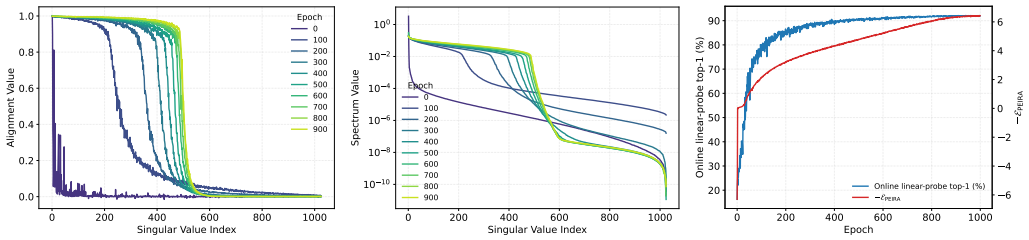


Figure 1: **Empirical dynamics of PEIRA on CIFAR-10.** PEIRA on a single training run. *Left:* alignment $\alpha_i = e_i^\top N_{U,V} e_i / (\|e_i\| \, \|N_{U,V} e_i\|) \in [0,1]$ of the top eigenvectors e_i of the signal $\Sigma_{U,V}$ with respect to the noise $N_{U,V}$ during training. $\alpha_i = 1$ implies e_i is also an eigenvector of $N_{U,V}$. Both matrices gradually align their eigenvectors during training. *Middle:* spectrum of $\Sigma_{U,V}$ during training develops additional significant modes. *Right:* validation top-1 and the negative PEIRA objective $-\mathcal{E}_{\text{PEIRA}}$ during training are strongly correlated.

At each iteration of Algorithm 1, running estimates $\hat{P}^\star$ and $\hat{Q}^\star$ of $P^\star$ and $Q^\star$ are updated from the current features. Specifically, Σ and N maintain EMA estimates of $\Sigma_{U,V}$ and $N_{U,V}$ from mini-batch features $\Phi_X = (U_\theta(x_i))_{i=1}^B$ and $\Phi_Y = (V_\psi(y_i))_{i=1}^B$, stacked as matrices in $\mathbb{R}^{B \times k}$, yielding closed-form approximations of $P^\star$ and $Q^\star$. These quantities define a mini-batch estimate $\widehat{\mathcal{L}}_{\text{aux}}$, through which the encoder parameters are updated by backpropagation, with $\hat{P}^\star$ and $\hat{Q}^\star$ held fixed (lines 7–8). This is efficient for moderate k, since the linear estimators are computed analytically from running feature statistics.

Algorithm 1: SC-PEIRA

Require: Initial encoder parameters θ, ψ, optimizer
Require: Initial $k \times k$ matrices Σ, N, dataset $\mathcal{D}$
Require: Regularization λ, EMA rate $\eta \in (0,1]$ (weight of the current batch)

1: **while** not converged **do**
2: Sample mini-batches $\mathcal{B}$ from $\mathcal{D}$
3: $\Phi_X, \Phi_Y \leftarrow (U_\theta(x))_{(x,y)\in\mathcal{B}}, (V_\psi(y))_{(x,y)\in\mathcal{B}}$
4: $\Sigma \leftarrow (1-\eta)\Sigma + \frac{\eta}{B}(\Phi_X^\top \Phi_Y + \Phi_Y^\top \Phi_X)$
5: $N \leftarrow (1-\eta)N + \frac{\eta}{B}(\Phi_X^\top \Phi_X + \Phi_Y^\top \Phi_Y)$
6: $\hat{P}^\star, \hat{Q}^\star \leftarrow \Sigma(N+\lambda I)^{-1}, (N+\lambda I)^{-1}$
7: $g_\theta, g_\psi \leftarrow \nabla_{\theta,\psi} \widehat{\mathcal{L}}_{\text{aux}}(U_\theta, V_\psi; \hat{P}^\star, \hat{Q}^\star)$
8: Update θ, ψ using gradients g_θ, g_ψ
9: **end while**

Remark 5.1. ***Shared encoders.*** *When the two views are symmetric, a single shared encoder can be used for both, as is common in practice [Chen and He, 2021]. The algorithm can then be adapted by setting $V_\psi = U_\theta$ and performing a single gradient update with $\nabla_\theta \widehat{\mathcal{L}}_{\text{aux}}(U_\theta, U_\theta; \hat{P}^\star, \hat{Q}^\star)$.*

Remark 5.2. ***EMA-rate schedule.*** *The rate η weights the current mini-batch, so $1-\eta$ is the EMA momentum. In practice, η is annealed from η_{init} to a smaller value $\eta_{\min}$ over training, favoring fresh statistics early and lower estimator variance near convergence.*

Remark 5.3. ***Joint updates.*** *Equations (3) and (6) yield a bilevel problem, enabling joint updates of $P^\star$, $Q^\star$, and the encoders [Dagréou et al., 2022, Arbel and Mairal, 2022], similarly to BYOL [Grill et al., 2020] and I-JEPA [Assran et al., 2023]. This would avoid explicit matrix inversion, but add hyperparameters and potentially degrade estimates of $P^\star$ and $Q^\star$.*

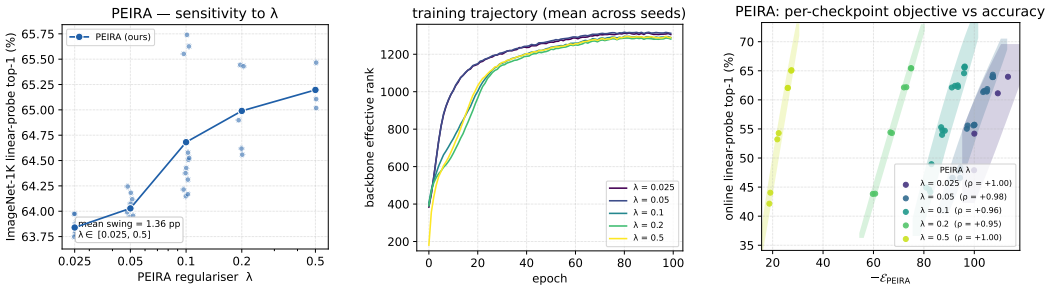


Figure 2: **Study of PEIRA regularization on ImageNet-1K.** ResNet-50, 100 epochs, PEIRA ablation setting (Section G.1); 1–3 random seeds per cell. *Left:* sensitivity of the final online linear-probe top-1 accuracy to the regularizer λ; each point represents a hyperparameter configuration, with $\eta_{\min}$, optimizer weight decay, and gradient clipping varied across runs. *Middle:* backbone entropy-based effective rank [Garrido et al., 2023a] versus epoch, colored by λ; smaller λ accumulates rank faster in the early phase. *Right:* online linear-probe top-1 versus the negative objective $-\mathcal{E}_{\text{PEIRA}}$ (mirroring the rightmost CIFAR-10 panel of Figure 1), evaluated for checkpoints at epochs 25, 50, 75, 100 across runs spanning various λ values; each dot is one (cell, checkpoint) pair, pooling over $\eta_{\min}$, optimizer weight decay, and gradient clipping. The per-λ Spearman correlation is reported as (ρ) in the legend.

6 Experiments

Setup. We evaluate PEIRA through linear-probe classification on ImageNet-1K and CIFAR-10, and transfer learning on eight downstream classification datasets using ImageNet-1K-pretrained representations. We further investigate feature alignment, dimensional collapse, and sensitivity to training settings. We compare against VICReg [Bardes et al., 2022] and SIGReg/LeJEPA [Balestriero and LeCun, 2025], standardizing augmentations, pretraining budgets, and linear-probe evaluation while retaining method-specific projectors and optimization hyperparameters. On ImageNet-1K, VICReg follows the recipe of Bardes et al. [2022], Garrido et al. [2023b], while SIGReg is configured based on the experiments in Section G.4. On CIFAR-10, the baseline configurations selected at 300 epochs by Terver et al. [2026a] are evaluated for 1000 epochs. For PEIRA, the EMA schedule is fixed for all results in Table 1, and only the regularization λ is dataset dependent ($\lambda = 0.7$ on CIFAR-10, $\lambda = 0.1$ on ImageNet-1K). Full experimental details are provided in Appendix G.1.

Representation quality. Table 1 summarizes results on ImageNet-1K, CIFAR-10 and the eight transfer tasks. On ImageNet-1K, PEIRA with $k = 1024$ achieves the strongest performance under our controlled evaluation, outperforming both VICReg and SIGReg. Its accuracy is also comparable to the published BYOL result and exceeds those of SimSiam and SimCLR, although these comparisons are indicative, as the reported results were obtained outside our controlled training protocol. On CIFAR-10, PEIRA outperforms VICReg and achieves performance close to SIGReg. Transfer results further demonstrate the generalization ability of PEIRA: with $k = 1024$, it outperforms SIGReg on average and approaches VICReg's mean performance (Table 1). Detailed results in Table G.9 of Appendix G.10 show that PEIRA surpasses SIGReg on seven of eight datasets and achieves the best performance on three. Interestingly, the negative PEIRA objective closely tracks linear-probe accuracy on CIFAR-10 (Figure 1, right) and remains strongly correlated with it on ImageNet-1K (Figure 2, right), suggesting its potential as a label-free indicator of representation quality. Additional experiments show that PEIRA maintains competitive performance with longer pretraining (Appendix G.7), and achieves training times comparable to the baselines (Appendix G.9).

Spectral alignment and resistance to collapse. Figure 1 (left and middle) shows that the leading eigendirections of the signal matrix $\Sigma_{U,V}$ progressively align with those of the noise matrix $N_{U,V}$, while additional significant spectral modes emerge, consistent with our population analysis. To assess resistance to collapse, we compare PEIRA with its self-distillation counterpart across initialization scales on CIFAR-10 (Appendix G.12). PEIRA maintains high accuracy and substantial effective rank across scales $\{0.01, 0.1, 1\}$, whereas self-distillation exhibits greater sensitivity to initialization and pronounced dimensional collapse at smaller scales. This is consistent with Proposition 4.2: at zero initialization scale, the encoders produce the zero representation, which is a stable attractor of the self-distillation dynamics. Further experiments on synthetic data with known CCA structure confirm the predicted spectral selection and recovery of canonical directions (Appendix G.11).

Sensitivity to hyperparameters. Table 1 shows that projector dimension has a greater impact on ImageNet-1K than on CIFAR-10: reducing k from 1024 to 256 decreases accuracy by 2.72 and 0.32 percentage points, respectively. A broader study confirms improvements up to $k = 1024$ on ImageNet-1K (Appendix G.6), suggesting that higher-dimensional representations are particularly beneficial on more challenging datasets. Figure 2 (left) further demonstrates robustness to regularization, with mean accuracy varying by only 1.36 percentage points over $\lambda \in [0.025, 0.5]$. Smaller λ accelerates early growth of backbone effective rank, while final ranks differ by only approximately 2% (Section G.3), consistent with the spectral role of λ predicted by Proposition 4.4. Finally, additional ablations show limited sensitivity to the EMA schedule and batch sizes (Appendix G.8).

7 Conclusions

We have analyzed the JEPA dynamics in the setting where the predictor is linear although the encoders are arbitrary nonlinear square-integrable functions, identified the corresponding equilibria and their stability, and proposed an alternative objective that provably leads to nontrivial global optimizers. We have also demonstrated through preliminary experiments with standard benchmarks the promise of this approach. We now briefly discuss its limitations and future work.

Limitations. Our analysis is carried out at the population level and assumes access to the optimal linear predictor $P_{U,V}$. This predictor is a deliberate design choice, but in practice it is only estimated: Algorithm 1 replaces population moments with finite-sample EMA estimates, continuous-time flows with discrete stochastic updates, and nonparametric encoders with neural networks. Each approximation is principled in isolation: discrete noisy recursions track limiting continuous-time dynamics under suitable conditions [Benaïm, 1999]; empirical covariance operators concentrate around their population counterparts [Koltchinskii and Lounici, 2017]; averaging reduces estimation variance subject to a bias-variance trade-off [Polyak and Juditsky, 1992]; and expressive networks approximate functions in L^2 [Hornik, 1991], with convergence towards population optima proved in overparameterized and mean-field regimes [Chizat and Bach, 2018, Mei et al., 2018]. A rigorous end-to-end guarantee would require jointly controlling discretization, covariance estimation, EMA bias and variance, matrix-inverse perturbations, approximation, and nonconvex stochastic optimization, e.g., through the theory of singularly perturbed systems [Habets, 1974, Arbel and Mairal, 2022], and is left for future work. Empirically, PEIRA is robust to the quality of the predictor estimate across EMA schedules and batch sizes (Appendix G.8). Our proofs further assume distinct canonical correlations, a simplifying assumption that can be relaxed at the cost of a more complicated characterization (Appendix B.6). Finally, although the experiments have demonstrated competitive performance on benchmarks like ImageNet-1K, more thorough experiments are needed to establish PEIRA as a viable alternative to BYOL, SimSiam, VICReg, V-JEPA 2 and SIGReg for example, notably through experiments on more complex tasks such as video classification [Goyal et al., 2017, Kay et al., 2017].

Future work. Most (but not all) of our discussion and proofs are valid for encoders that belong to Hilbert spaces of infinite dimension, and predictors that are continuous linear operators over these. We plan to investigate next how much of the theory remains valid in this general setting and whether using reproducing kernel Hilbert spaces might lead to a practical computational solution. As shown by recent work on so-called world models [Sobal et al., 2025, Terver et al., 2026b], learning action models in the form of control encoders and conditioning predictors on these models has led to new approaches to robotic planning. We also plan to investigate how PEIRA could be used in this setting, the simplest but probably unsatisfactory instance being to concatenate the state and control features.

Acknowledgments and Disclosure of Funding

This work was supported in part by the French government under management of Agence Nationale de la Recherche as part of the "France 2030" program, PR[AI]RIE-PSAI project, reference ANR-23-IACL-0008. J.P. was supported in part by the Louis Vuitton/ENS chair in artificial intelligence, the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project), and a Global Distinguished Professorship at the Courant Institute of Mathematical Sciences and the Center for Data Science at New York University. M.A. was supported by the ANR project BONSAI (grant ANR-23-CE23-0012-01) and was granted access to the HPC resources of IDRIS under the allocation [AD011016476R1] made by GENCI.

References

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations (ICLR)*, Nov. 2022. URL https://hal.archives-ouvertes.fr/hal-03455458. working paper or preprint.

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. 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.

M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

R. Balestriero and Y. LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35: 26671–26685, 2022.

R. Balestriero and Y. LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

I. Barbălat. Systèmes d'équations différentielles d'oscillations non linéaires. *Revue de Mathématiques Pures et Appliquées*, 4:267–270, 1959.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer, 1999. doi: 10.1007/BFb0096509.

R. Bhatia. *Matrix analysis*. Springer Science & Business Media, 2013.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

L. Bossard, M. Guillaumin, and L. Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, volume 8694 of *Lecture Notes in Computer Science*, pages 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

X. Chang, T. Xiang, and T. M. Hospedales. Scalable and effective deep cca via soft decorrelation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1488–1497, 2018.

J. Chapman, L. Wells, and A. L. Aguila. Unconstrained stochastic CCA: Unifying multiview and self-supervised learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PHLVmV88Zy.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

X. Chen and K. He. Exploring simple siamese representation learning. In *Proc. CVPR*, 2021.

R. Chill. On the łojasiewicz–simon gradient inequality. *Journal of Functional Analysis*, 201(2): 572–601, 2003.

R. Chill, A. Haraux, and M. A. Jendoubi. Applications of the lojasiewicz-simon gradient inequality to gradient-like evolution equations. *Anal. Appl.(Singap.)*, 7(4):351–372, 2009.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, pages 3036–3046, 2018.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

J. B. Conway. *A course in functional analysis*. Springer, 2019.

M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

C. De Sa, C. Re, and K. Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International conference on machine learning*, pages 2332–2341. PMLR, 2015.

A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pages 3015–3024. PMLR, 2021.

P. M. Feehan and M. Maridakis. Łojasiewicz–simon gradient inequalities for analytic and morse–bott functions on banach spaces. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2020(765):35–67, 2020.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *J. Mach. Learn. Res.*, 8:361–383, May 2007. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248659.1248673.

A. Galashov, N. Da Costa, L. Xu, P. Hennig, and A. Gretton. Closed-form last layer optimization. *arXiv preprint arXiv:2510.04606*, 2025.

Q. Garrido, R. Balestriero, L. Najman, and Y. LeCun. Rankme: assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=kDEL91Dufpa.

I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Spectral Theory of Compact Self Adjoint Operators*, pages 171–191. Birkhäuser Basel, Basel, 2003. ISBN 978-3-0348-7980-4. doi: 10.1007/978-3-0348-7980-4_4. URL https://doi.org/10.1007/978-3-0348-7980-4_4.

R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017. URL https://arxiv.org/abs/1706.04261.

R. D. Grigorieff. A note on von neumann's trace inequalitv. *Mathematische Nachrichten*, 151(1): 327–328, 1991.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. 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.

P. Habets. Stabilité asymptotique pour des problèmes de perturbations singulières. In *Stability Problems*, pages 2–18. Springer, 1974.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. doi: 10.1162/0899766042321814.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 2020.

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

L. Huang, Y. Ni, X. Weng, R. M. Anwer, S. Khan, M.-H. Yang, and F. S. Khan. Understanding whitening loss in self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):9479–9492, 2024.

M. Huh, B. Cheung, T. Wang, and P. Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Y. Jia, F. Nie, and C. Zhang. Trace ratio problem revisited. *IEEE Transactions on Neural Networks*, 20(4):729–735, 2009.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. doi: 10.3150/15-BEJ730.

J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

H. O. Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29 (3):719–736, 1958.

Y. LeCun. A path towards autonomous machine intelligence. *OpenReview*, Jun 2022.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257. PMLR, 2016.

E. Littwin, O. Saremi, M. Advani, V. Thilak, P. Nakkiran, C. Huang, and J. Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024.

S. Lojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1982.

Q. Lyu, X. Fu, W. Wang, and S. Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. *arXiv preprint arXiv:2106.07115*, 2021.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

T. Michaeli, W. Wang, and K. Livescu. Nonparametric canonical correlation analysis. *arXiv preprint arXiv:1511.04839*, 2015.

D. Morales-Brotons, T. Vogels, and H. Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=2M9CUnYnBA.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

J. Ponce, B. Terver, M. Hebert, and M. Arbel. Dual perspectives on non-contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2026.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575 [cs]*, Sept. 2014. URL http://arxiv.org/abs/1409.0575. arXiv: 1409.0575.

B. Shi, W. Su, and M. I. Jordan. On learning rates and schrödinger operators. *Journal of Machine Learning Research*, 24(379):1–53, 2023.

O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

V. Sobal, W. Zhang, K. Cho, R. Balestriero, T. G. Rudner, and Y. LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025.

L. Sun, S. Ji, and J. Ye. A least squares formulation for a class of generalized eigenvalue problems in machine learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 977–984, 2009.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2 edition, 2018. ISBN 9780262039246.

Y. Tang, Z. D. Guo, P. H. Richemond, B. A. Pires, Y. Chandak, R. Munos, M. Rowland, M. G. Azar, C. Le Lan, C. Lyle, et al. Understanding self-predictive learning for reinforcement learning. In *International Conference on Machine Learning*, pages 33632–33656. PMLR, 2023.

B. Terver, R. Balestriero, M. Dervishi, D. Fan, Q. Garrido, T. Nagarajan, K. Sinha, W. Zhang, M. Rabbat, Y. LeCun, and A. Bar. A lightweight library for energy-based joint-embedding predictive architectures. In *ICLR 2026 Workshop on World Models*, 2026a.

B. Terver, T.-Y. Yang, J. Ponce, A. Bardes, and Y. LeCun. What drives success in physical planning with joint-embedding predictive world models?, 2026b. URL https://arxiv.org/abs/2512.24497.

G. Teschl. Ordinary differential equations and dynamical systems. *Graduate Studies in Mathematics*, 140:08854–8019, 2000.

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

T. T. Truong. Some convergent results for backtracking gradient descent method on banach spaces, 2020.

H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1–2): 419–449, 2017.

Z. Wen and Y. Li. The mechanism of prediction head in non-contrastive self-supervised learning. In *NeurIPS*, 2022.

L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

H. Zhang, Q. Wu, J. Yan, D. Wipf, and P. S. Yu. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems*, 34:76–89, 2021.

Appendix

Contents

1 Introduction **1**

2 Related work **2**

3 From Self-Distillation to PEIRA **3**

4 Canonical structure and stability of the learning dynamics **4**

4.1 Characterizing the self-distillation equilibria and their stability 5

4.2 Characterizing the PEIRA equilibria and their stability 6

5 Optimizing PEIRA's objective **7**

5.1 An auxiliary objective for gradient computation 7

5.2 A stochastic compositional algorithm for PEIRA 7

6 Experiments **9**

7 Conclusions **10**

Appendix **16**

A Preliminaries and Background on Canonical Correlation Analysis **18**

A.1 Nonlinear canonical correlation analysis 18

A.2 Characterizing nonlinear CCA via generalized eigenvalue decomposition (GEVD) . 18

A.3 Top-k subspaces and ratio-trace optimization. 19

B General Proof Strategy and Main Results **19**

B.1 General Setting and Notations 19

B.2 Reduction to Hilbert–Schmidt coordinate representations 20

B.3 Convergence of Dynamical Systems to Critical Points 21

B.4 Characterization of Critical Points of F^κ 23

B.5 Stability Analysis of Critical Points of F^κ 23

B.6 Degenerate canonical spectra and shared encoders 25

C Proofs of the main results **26**

C.1 Gradient flow of the PEIRA objective 26

C.1.1 Convergence to critical points and characterization of critical points 26

C.1.2 Minimizers span thresholded canonical correlation subspaces 26

C.1.3 The gradient of PEIRA's objective matches that of the auxiliary loss $\mathcal{L}_{\mathrm{aux}}$. 28

C.2 Self-distillation dynamics 29

C.2.1 Convergence of the self-distillation dynamics and characterization of equilibria 29

C.2.2 Stability of equilibria of the self-distillation dynamics 29

D Proofs of the reduction to coordinate representations **30**

D.1 Proof of Proposition B.1: Expressions of the gradient of PEIRA's objective 30

D.2 Proof of Proposition B.2: Expressions of the self-distillation vector field 32

E Proof of convergence to critical points **33**

E.1 Proof of convergence to critical points for general gradient-like dynamical systems 33

E.2 Proof of the auxiliary results of Proposition B.4 and Lemmas B.1 to B.4 36

F Proof of the characterization of equilibria and their stability **40**

F.1 Proof of Proposition B.6 on characterization of equilibria 40

F.2 Proof of the diagonalization of the Jacobian of the vector field F^{κ} at critical points 41

F.3 Proof of Propositions B.8 and B.9 on stability of equilibria 43

G Experimental details and additional results **47**

G.1 Experimental details 47

G.2 Per-epoch correlation of $\mathcal{E}_{\text{PEIRA}}$ with downstream accuracy 49

G.3 Effective rank vs. λ on ImageNet-1K 50

G.4 Comparison with SIGReg on ImageNet-1K 51

G.5 Comparison with VICReg on ImageNet-1K 52

G.6 Projector output width 52

G.7 Longer pretraining 52

G.8 EMA schedule and batch size 53

G.9 Wall-clock cost and conditioning of the $k \times k$ solve 54

G.10 Transfer to the LeJEPA-8 suite 54

G.11 Synthetic two-view Gaussian data 55

G.12 CIFAR-10: robustness to initialization scale 56

A Preliminaries and Background on Canonical Correlation Analysis

A.1 Nonlinear canonical correlation analysis

We start by briefly recalling population nonlinear CCA [Lancaster, 1958, Michaeli et al., 2015], which seeks representations $(U, V) \in \mathcal{U} \times \mathcal{V}$ maximizing inter-view correlation under orthonormality constraints. Specifically, these are solutions to the following constrained optimization problem:

$$\max_{(U,V)\in\mathcal{U}\times\mathcal{V}} \mathbb{E}\left[U(x)^\top V(y)\right], \qquad \text{s.t.} \qquad \mathbb{E}\left[U(x)U(x)^\top\right] = I, \quad \mathbb{E}\left[V(y)V(y)^\top\right] = I, \tag{16}$$

Such representations can be thought of as *predictive* in the sense that $V(y)$ is most predictable from $U(x)$ and vice versa. We need the following mild regularity assumption on the data distribution to ensure nonlinear CCA admits a spectral characterization [Michaeli et al., 2015]:

(A) The density ratio $f(x, y)$, i.e., Radon-Nikodym derivative, of $\mathbb{P}$ with respect to $\mathbb{P}_X \otimes \mathbb{P}_Y$ exists and is integrable under $\mathbb{P}$, i.e., $\mathbb{E}[f(x, y)] < \infty$.

This is a mild assumption (see Remark A.1) ruling out degenerate cases such as a deterministic dependence between x and y. Under Assumption **(A)**, CCA admits a spectral characterization in terms of the canonical directions $(\psi_i, \xi_i)_{1\le i\le R}$ introduced in Definition 4.1. Specifically, for $k \le R$, a solution to the CCA problem is obtained by stacking the first k canonical directions:

$$U(x) = \sum_{i=1}^{k} f_i \psi_i(x) \qquad V(y) = \sum_{i=1}^{k} f_i \xi_i(y), \tag{17}$$

where $(f_i)_{1\le i\le k}$ are the columns of the $k \times k$ identity matrix. Accordingly, the top-k canonical subspaces are the spans of these directions.

Remark A.1. *Assumption (A) holds, for example, on compact domains when x and y are obtained by independently perturbing latent paired samples (x_0, y_0) through kernels whose densities are uniformly bounded above and below. Indeed, if $0 < m_X \le k_X(x \mid x_0) \le M_X$ and $0 < m_Y \le k_Y(y \mid y_0) \le M_Y$, then $p(x, y) \le M_X M_Y$, $p(x) \ge m_X$, and $p(y) \ge m_Y$, so $f(x, y) = \frac{p(x,y)}{p(x)p(y)} \le \frac{M_X M_Y}{m_X m_Y}$. For example, one may take $k_X(x \mid x_0) = (1 - \varepsilon_X)q_X(x \mid x_0) + \varepsilon_X / \operatorname{vol}(\mathcal{X})$, where q_X is any uniformly bounded local perturbation density on $\mathcal{X}$ and $\varepsilon_X \in (0, 1)$, and similarly for k_Y.*

A.2 Characterizing nonlinear CCA via generalized eigenvalue decomposition (GEVD)

A GEVD problem aims to find eigenpairs (s, w) such that $Aw = s\,Bw$, where A and B are bounded symmetric operators in some Hilbert space $\mathcal{H}$, with B positive definite. GEVD problems naturally arise in a range of machine learning problems [Sun et al., 2009] and, in particular, provide a tool for performing CCA. Specifically, a solution of the CCA problem is characterized from a particular GEVD problem by taking its top-k eigenvectors and applying an orthonormalization procedure, such as Gram-Schmidt, to enforce the constraints in Equation (16). Specifically, B is the identity operator defined on the product space $\mathcal{H} = L_2(\mathbb{P}_x) \times L_2(\mathbb{P}_y)$ of real-valued square integrable functions w.r.t. $\mathbb{P}_X$ and $\mathbb{P}_Y$, whereas A is the uncentered cross-covariance operator defined on the same space as:

$$\langle w, Aw' \rangle_{\mathcal{H}} = \mathbb{E}\left[u(x)v'(y) + u'(x)v(y)\right], \qquad \forall w = (u, v), w' = (u', v') \in \mathcal{H}. \tag{18}$$

By Cauchy–Schwarz, A is bounded with operator norm at most 1:

$$|\langle w, Aw' \rangle_{\mathcal{H}}| \le \|w\|_{\mathcal{H}} \|w'\|_{\mathcal{H}}, \qquad \forall\, w, w' \in \mathcal{H}. \tag{19}$$

To ensure the problem[4] has a well-behaved leading spectrum, we need to assume that the operator A is compact[5], which holds as soon as Assumption **(A)** holds [Michaeli et al., 2015]. In this case, the canonical directions $(\psi_i, \xi_i)_{1\le i\le R}$ and canonical correlations $(c_i)_{1\le i\le R}$ correspond to eigenvectors and positive eigenvalues of the GEVD defined by the operators A and B. Recovering representations U, V that span the top-k subspace then yields a CCA solution after orthogonalization.

[4] Since the operator B is the identity, such problem reduces to an EVD problem of the operator A, albeit, possibly, in infinite dimensions.

[5] A compact operator is a linear operator that maps bounded sequences to relatively compact sequences; equivalently, every image sequence has a convergent subsequence.

Eigen structure of the cross-covariance operator A. The following proposition provides an eigen decomposition for operator A under Assumption (A), which allows expressing the solution of non-linear CCA and will be crucial in our analysis.

Proposition A.1. *Under Assumption (A), the operator A is compact and admits a spectral decomposition. Specifically, letting $I = \mathbb{N}$ if $\mathcal{H}$ has infinite dimensions or $I = \{1, \dots, \dim(\mathcal{H})\}$ otherwise, there exists an orthonormal basis $(\zeta_i)_{i\in I}$ of $\mathcal{H}$ and eigenvalues $(s_i)_{i\in I}$ satisfying $A\zeta_i = s_i\zeta_i$. Let I_+ be the index set of positive eigenvalues, and σ be an ordering of the indices in I_+ such that $s_{\sigma(1)} \geq \cdots \geq s_{\sigma(R)} > 0$, with R possibly infinite (in which case $s_{\sigma_j} \to 0$). Set $c_j = s_{\sigma_j}$ and $(\psi_j, \xi_j) = \sqrt{2}\zeta_{\sigma_j}$. Then if c_j are all distinct, $(\psi_j)_{j=1}^R$ and $(\xi_j)_{j=1}^R$ are orthonormal families of $L_2(\mathbb{P}_X)$ and $L_2(\mathbb{P}_Y)$ satisfying $\mathbb{E}[\psi_j(x)\xi_{j'}(y)] = c_j\delta_{j=j'}$ and that exhaust all cross-correlated directions (see Definition 4.1). Moreover, any index $i \in I$ corresponding to negative eigenvalue is of the form $|s_i| = c_j$ for some $j \in \{1, \dots, R\}$, and its eigenvector ζ_i satisfies $\sqrt{2}\zeta_i = (\psi_j, -\xi_j)$. The converse is also true: for any index $j \in \{1, \dots, R\}$ there exists an index $i \in I$ with negative eigenvalue s_i satisfying $|s_i| = c_j$.*

A.3 Top-k subspaces and ratio-trace optimization.

As illustrated through CCA, one often only needs to recover a *top-k subspace representation*, i.e. an element $W = (U, V) \in \mathcal{U} \times \mathcal{V}$ whose k components span the same subspace as the k largest generalized eigenvalues of a given GEVD problem. For a GEVD problem of finite dimension D and when $A \succeq 0$, this subspace was characterized as the span of rows of any maximizer of the ratio-trace objective [Wang et al., 2007, Jia et al., 2009]:

$$\max_{W\in\mathbb{R}^{k\times D}} Tr\Big((WAW^\top)(WBW^\top)^{-1}\Big). \tag{20}$$

While Equation (20) appears to offer an appealing route to direct GEVP via iterative optimization, it is *basis-invariant*, i.e. for any invertible $Q \in \mathbb{R}^{k\times k}$, W and QW attain the same value. Consequently, the above objective is flat along re-parameterizations within the optimal subspace and is typically ill-conditioned for iterative optimization.

The regularization we introduce in the linear predictor considered by PEIRA results in a modified *ratio-trace* objective that breaks this invariance and becomes smooth. The resulting criterion selects a particular subset of optimal subspaces, generally without enforcing B-orthonormality.

B General Proof Strategy and Main Results

B.1 General Setting and Notations

Hilbert spaces of square-integrable functions. Recall the product space $\mathcal{H} = L_2(\mathbb{P}_x) \times L_2(\mathbb{P}_y)$ of real-valued square integrable functions w.r.t. $\mathbb{P}_X$ and $\mathbb{P}_Y$, where $\mathbb{P}_X$ and $\mathbb{P}_Y$ denote the marginals of the joint data distribution $\mathbb{P}$ of pairs (x, y). Let $\mathcal{W}$ be the product space $\mathcal{U} \times \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are the function spaces of square-integrable encoders, $\mathcal{U} = L^2(\mathbb{P}_X; \mathbb{R}^k)$, $\mathcal{V} = L^2(\mathbb{P}_Y; \mathbb{R}^k)$. We will view elements $W = (U, V) \in \mathcal{W}$ as bounded operators from $\mathcal{H}$ to $\mathbb{R}^k$ acting on any element $w = (u, v) \in \mathcal{H}$ as $Ww = \mathbb{E}[U(x)u(x) + V(y)v(y)] \in \mathbb{R}^k$ and denote by $W^\top$ the adjoint of W. For an element $W = (U, V) \in \mathcal{W}$, we denote by $WAW^\top$ and $WW^\top$ the $k \times k$ matrices given by:

$$WAW^\top = \mathbb{E}\left[U(x)V(y)^\top + V(y)U(x)^\top\right], \quad WW^\top = \mathbb{E}\left[U(x)U(x)^\top + V(y)V(y)^\top\right]. \tag{21}$$

Hilbert space of square-integrable sequences. Let I be the index set defined in Proposition A.1 which is equal to $\mathbb{N}$ if $\mathcal{H}$ has infinite dimensions, or to $\{1, \dots, \dim(\mathcal{H})\}$ otherwise. Denote by $\ell^2(I)$ the space of square integrable sequences indexed by I and let $(e_i)_{i\in I}$ be its canonical orthonormal basis. Note that if I is finite, then $\ell^2(I)$ can be identified with $\mathbb{R}^d$. Furthermore, fix $k \in \mathbb{N}$ and consider the space of Hilbert–Schmidt operators $\mathbb{W} = \mathrm{HS}(\ell^2(I), \mathbb{R}^k)$, i.e. the set of bounded linear operators that map elements in $\ell^2(I)$ to $\mathbb{R}^k$ and have a finite Hilbert-Schmidt norm: $\|\cdot\|_{\mathbb{W}}$:

$$\|\mathbf{w}\|_{\mathbb{W}}^2 = \sum_{i\in I} \|\mathbf{w}e_i\|_{\mathbb{R}^k}^2, \qquad \langle \mathbf{w}, \mathbf{w}'\rangle_{\mathbb{W}} := \sum_{i\in I} \langle \mathbf{w}e_i, \mathbf{w}'e_i\rangle_{\mathbb{R}^k}, \qquad \forall \mathbf{w}, \mathbf{w}' \in \mathbb{W}. \tag{22}$$

Any element $\mathbf{w} \in \mathbb{W}$ can therefore be formally identified with a $k \times |I|$ matrix $\mathbf{w} = (w_{pi})_{1 \le p \le k,\ i \in I}$ with $\sum_{p,i} |w_{pi}|^2 < \infty$. We denote by $\mathbf{w}^\top$ its adjoint. Let $(b_1, \dots, b_k)$ be the standard orthonormal basis of $\mathbb{R}^k$ and define the entry-wise rank-one operators which form an orthonormal basis of $\mathbb{W}$.

$$E_{pi} := b_p \otimes e_i^* \in \mathbb{W}, \qquad E_{pi} e_i = b_p, \ \ E_{pi} e_j = 0 \ (j \neq i). \tag{23}$$

Coordinate representation in the eigenbasis of the operator A. Denote by $\mathbf{w} = (\mathbf{w}_i)_{i \in I}$ the components of an element $W \in \mathcal{W}$ in the eigenbasis $(\zeta_i)_{i \in I}$ of the operator A defined in Equation (18). Each component $\mathbf{w}_i$ is a vector in $\mathbb{R}^k$ so that $\mathbf{w}$ can be viewed as an element of $\mathbb{W}$. Then the following holds for any $W \in \mathcal{W}$ with component $\mathbf{w} \in \mathbb{W}$ in the eigenbasis

$$WAW^\top = \mathbf{w} S \mathbf{w}^\top, \qquad WW^\top = \mathbf{w}\mathbf{w}^\top, \tag{24}$$

where S is a bounded compact operator on $\ell^2(I)$ that is diagonal in its canonical basis and whose diagonal elements contain the eigenvalues of A.

Orthogonal decomposition induced by the compact operator S. Let S be the operator considered in Equation (24). Since S is a diagonal operator in the canonical basis of $\ell^2(I)$, it satisfies $Se_i = s_i e_i$, where $(s_i)_{i \in I}$ are the eigenvalues of A. Define the index sets

$$I_+ := \{i \in I : \ s_i > 0\}, \qquad I_0 := \{i \in I : \ s_i = 0\}, \qquad I_- := \{i \in I : \ s_i < 0\}, \tag{25}$$

so that $I = I_+ \sqcup I_- \sqcup I_0$, and assume, without loss of generality, that $(|s_i|)_{i \in I_+}$ and $(|s_i|)_{i \in I_-}$ are decreasing in their index. The following orthogonal decomposition holds:

$$\ell^2(I) = \ell^2(I_+) \oplus \ell^2(I_0) \oplus \ell^2(I_-), \qquad \ell^2(I_\sigma) := \overline{\operatorname{span}}\{e_i : \ i \in I_\sigma\} \ \ (\sigma \in \{+, -, 0\}). \tag{26}$$

The above decomposition induces a block decomposition on the space $\mathbb{W}$ in the sense that

$$\mathbb{W} = \mathrm{HS}(\ell^2(I_+), \mathbb{R}^k) \ \oplus \ \mathrm{HS}(\ell^2(I_0), \mathbb{R}^k) \ \oplus \ \mathrm{HS}(\ell^2(I_-), \mathbb{R}^k), \tag{27}$$

with corresponding orthonormal bases $\{E_{pi} : \ i \in I_\sigma\}$ for $\sigma \in \{+, -, 0\}$.

Vector fields on $\mathbb{W}$. Let λ be a positive number. We define the family of vector fields over $\mathbb{W}$

$$F^\kappa(\mathbf{w}) = \mathbf{w} \left(I - \left(\mathbf{w}^\top \mathbf{w} + \lambda I\right)^{-1} S (\mathbf{w}^\top \mathbf{w} S)^\kappa \left(\mathbf{w}^\top \mathbf{w} + \lambda I\right)^{-1} \right), \qquad \forall \mathbf{w} \in \mathbb{W}, \tag{28}$$

where κ is a non-negative integer. In particular, we will consider two values: $\kappa = 0$ will recover the coordinate representation of the vector field driving the gradient flow of PEIRA's objective, whereas $\kappa = 1$ will recover the coordinate representations of the vector fields driving the self-distillation dynamics as shown next.

B.2 Reduction to Hilbert–Schmidt coordinate representations

In this section we provide two results, proven in Appendix D, allowing to reduce the analysis to the space of coordinate representations of elements $W \in \mathcal{W}$ in the eigenbasis of A, as we do later on.

Operator expression of PEIRA's gradient. The following proposition allows expressing the PEIRA's objective and its gradient in terms of the operator A, provides an expression of the coordinates of $\nabla \mathcal{E}_{\text{PEIRA}}$ in the eigenbasis of A and establishes the equivalence between the gradient flow dynamics of $\mathcal{E}_{\text{PEIRA}}$ and the dynamics of its components in the eigenbasis of A.

Proposition B.1 (Operator expressions of PEIRA's objective and its gradient)**.** *For given $W = (U, V) \in \mathcal{W}$, the objective $\mathcal{E}_{PEIRA}$ is smooth and admits an L-Lipschitz gradient, with $L = 4\lambda^{-1}$ when $0 < \lambda < 1$. Additionally, the following expressions hold for $\mathcal{E}_{PEIRA}$ and $\nabla \mathcal{E}_{PEIRA}$ in terms of the operator A in Equation (18):*

$$\begin{aligned} \mathcal{E}_{PEIRA}(W) &= -\frac{1}{2} Tr\left(WAW^\top \left(WW^\top + \lambda I\right)^{-1} \right) + \frac{\lambda}{2} Tr\left(WW^\top\right) \\ \nabla \mathcal{E}_{PEIRA}(W) &= \lambda W \left(I - \left(W^\top W + \lambda I\right)^{-1} A \left(W^\top W + \lambda I\right)^{-1} \right). \end{aligned} \tag{29}$$

*Moreover, assuming (**A**) holds, let $\mathbf{w} = (\mathbf{w}_i)_{i \in I} \in \mathbb{W}$ denote the components of an element $W \in \mathcal{W}$ in the eigenbasis $(\zeta_i)_{i \in I}$ of A and let S be the diagonal operator S defined in Equation (24). Then the components of $\nabla \mathcal{E}_{PEIRA}(W)$ in such eigenbasis are of the form:*

$$\mathbf{g} = \lambda F^0(\mathbf{w}), \qquad F^0(\mathbf{w}) = \mathbf{w} \left(I - \left(\mathbf{w}^\top \mathbf{w} + \lambda I\right)^{-1} S \left(\mathbf{w}^\top \mathbf{w} + \lambda I\right)^{-1} \right). \tag{30}$$

Finally, if $(W_t)_{t\geq 0}$ is a trajectory in $\mathcal{W}$ and $(\mathbf{w}_t)_{t\geq 0}$ are its components in the eigenbasis of A, then the following equivalence holds:

$$\dot{W}_t = -\nabla \mathcal{E}_{PEIRA}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^0(\mathbf{w}_t). \tag{31}$$

Operator expression of the Self-distillation vector field. Denote by $\mathcal{F}(U, V)$ the vector field arising from the self-distillation dynamics in Equation (7):

$$(\dot{U}_t, \dot{V}_t) = -\mathcal{F}(U_t, V_t), \qquad \mathcal{F}(U, V) = (\partial_U \mathcal{P}_{U \triangleright V}(P_{U,V}, U, V), \partial_V \mathcal{P}_{V \triangleright U}(P_{U,V}, V, U)). \tag{32}$$

The following proposition expresses $\mathcal{F}$ in terms of the operator A and expresses the coordinates of $\mathcal{F}(W)$ in the eigenbasis of A and establishes the equivalence between the dynamics in Equation (7) and that of its components in the eigenbasis of A.

Proposition B.2. *The vector field $\mathcal{F}(U, V)$ driving the dynamics in Equation (7) admits the expression:*

$$\mathcal{F}(W) = \lambda W \left(I - \left(W^\top W + \lambda I \right)^{-1} A W^\top W A \left(W^\top W + \lambda I \right)^{-1} \right), \qquad \forall W \in \mathcal{W}. \tag{33}$$

*Moreover, assuming **(A)** holds, let $\mathbf{w} = (\mathbf{w}_i)_{i \in I} \in \mathbb{W}$ denote the components of an element $W \in \mathcal{W}$ in the eigenbasis $(\zeta_i)_{i \in I}$ of A and let S be the diagonal operator S defined in Equation (24). Then the components of $\mathcal{F}(W)$ in such eigenbasis are of the form:*

$$\mathbf{g} = \lambda F^1(\mathbf{w}), \qquad F^1(\mathbf{w}) = \mathbf{w} \left(I - \left(\mathbf{w}^\top \mathbf{w} + \lambda I \right)^{-1} S \mathbf{w}^\top \mathbf{w} S \left(\mathbf{w}^\top \mathbf{w} + \lambda I \right)^{-1} \right). \tag{34}$$

Finally, if $(W_t)_{t\geq 0}$ is a trajectory in $\mathcal{W}$ and $(\mathbf{w}_t)_{t\geq 0}$ are its components in the eigenbasis of A, then the following equivalence holds:

$$\dot{W}_t = -\mathcal{F}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^1(\mathbf{w}_t). \tag{35}$$

Reduction to coordinate representation. The above results imply that it suffices to study dynamical systems of the form $\dot{\mathbf{w}}_t = -\lambda F^\kappa(\mathbf{w})$ defined in $\mathbb{W}$, with F^κ being the vector fields defined in Equation (28) for $\kappa \in \{0, 1\}$. Next, we study the properties of such vector fields, including stability of its critical points and convergence of the dynamics arising from them.

B.3 Convergence of Dynamical Systems to Critical Points

Here, we will show convergence to a critical point of F^κ for the following dynamical system:

$$\dot{\mathbf{w}}_t = -\lambda F^\kappa(\mathbf{w}_t), \qquad \mathbf{w}_0 \in \mathbb{W}. \tag{36}$$

The key step is to show that the above system admits a suitable Lyapunov function that makes it a gradient-like system [Chill et al., 2009]. We will then apply a general convergence result for gradient-like systems in infinite dimensional spaces (here $\mathbb{W}$), stated below and proven in Appendix E, after verifying that all conditions of such a result are satisfied by Equation (36).

Proposition B.3. *Let F be a smooth vector field defined on a separable Hilbert space $\mathbb{W}$ and consider the following dynamical system:*

$$\dot{\mathbf{w}}_t = -F(\mathbf{w}_t), \qquad \mathbf{w}_0 \in \mathbb{W}. \tag{37}$$

Assume there exists a real analytic function L defined over $\mathbb{W}$ such that the following holds:

- ***(B) Dissipation.*** *L is a Lyapunov function for the system, i.e. $\dot{L}(\mathbf{w}_t) \leq 0$ for all $t \geq 0$.*
- ***(C) Sufficient decrease.*** *There exists $c_{max} > 0$ such that $\|F(\mathbf{w}_t)\|^2 \leq -c_{max} \dot{L}(\mathbf{w}_t)$ for $t \geq 0$.*
- ***(D) Relative bound.*** *There exists a non-increasing real-valued function $\Omega : [0, +\infty) \to (0, +\infty)$ such that $\Omega(\|\mathbf{w}_t\|) \|\nabla L(\mathbf{w}_t)\|^2 \leq \|F(\mathbf{w}_t)\|^2$.*
- ***(E) Coercivity.*** *L is coercive, i.e. $L(\mathbf{w}) \to +\infty$ when $\|\mathbf{w}\| \to +\infty$.*
- ***(F) Lower bound.*** *L admits a finite lower bound.*

(G) Radial growth. *The vector field F, gradient ∇L and Hessian $\nabla^2 L$ all have radial growth, i.e. there exists a non-decreasing real-valued function $\omega : [0, +\infty) \to [0, +\infty)$ such that* $\max\left(\|F(\mathbf{w})\|, \|\nabla L(\mathbf{w})\|, \left\|\nabla^2 L(\mathbf{w})\right\|\right) \leq \omega(\|\mathbf{w}\|)$ *for all* $\mathbf{w} \in \mathbb{W}$.

(H) Fredholm property. *For any $\mathbf{w} \in \mathbb{W}$, the hessian $\nabla^2 L(\mathbf{w})$ can be decomposed as the sum of a boundedly invertible operator R and a bounded compact operator K.*

(I) Palais-Smale condition. *Any bounded sequence $(\mathbf{w}_n)_{n\geq 1}$ of $\mathbb{W}$ such that $\|\nabla L(\mathbf{w}_n)\| \to 0$ admits a convergent subsequence.*

The dynamical system in Equation (37) is defined at all times and converges to a critical point of F.

To apply Proposition B.3, we need to identify a suitable Lyapunov function L^κ for the system and check that it verifies all conditions of the result. Specifically, we identify the following candidate function L^κ:

$$L^\kappa(\mathbf{w}) = \frac{1}{3} Tr\left(\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^3\right) - \frac{1}{\kappa+1} Tr\left(\left(\mathbf{w} S \mathbf{w}^\top\right)^{\kappa+1}\right). \tag{38}$$

Note that L^1 is the coordinate version of the Lyapunov function $\mathcal{L}$ stated in Equation (8). The following proposition, proven in Appendix E.2, establishes that Equation (38) is indeed a Lyapunov function of the system satisfying the alignment conditions with the vector field F^κ required by Proposition B.3.

Proposition B.4 (Existence of a Lyapunov function)**.** *Let $\kappa \in \{0, 1\}$ and $\lambda > 0$. The function L^κ defined in Equation (38) is a Lyapunov function of the dynamical system in Equation (36), i.e. $\dot{L}^\kappa(\mathbf{w}_t) \leq 0$ for all $t \geq 0$. Moreover, the following inequalities hold along the trajectory:*

$$\frac{1}{4(\|\mathbf{w}_t\|^2 + \lambda)^4} \|\nabla L^\kappa(\mathbf{w}_t)\|^2 \leq \|F^\kappa(\mathbf{w}_t)\|^2 \leq -\frac{1}{2\lambda^3} \dot{L}^\kappa(\mathbf{w}_t), \qquad \forall t \geq 0. \tag{39}$$

The following four lemmas, also proven in Appendix E.2, establish geometric properties of the L^κ and F^κ that are needed for the application of Proposition B.3.

Lemma B.1 (Growth property of F^κ.)**.** *F^κ is radially bounded in the sense that for any $\mathbf{w} \in \mathbb{W}$, it holds that:*

$$\|F^\kappa(\mathbf{w})\| \leq \omega(\|\mathbf{w}\|), \qquad \text{with} \qquad \omega(R) = R\left(1 + \lambda^{-2} \|S\|_{op}^{1+\kappa} R^{2\kappa}\right) \quad \forall R \in [0, +\infty) \tag{40}$$

Lemma B.2 (Growth, coercivity and lower boundedness of L^κ)**.** *The function L^κ defined in Equation (38) is coercive and lower-bounded over $\mathbb{W}$. Moreover, the gradient ∇L^κ and Hessian $\nabla^2 L^\kappa$ both have a radial growth, i.e. there exists a non-decreasing real-valued function $\omega : [0, +\infty) \to [0, +\infty)$ such that:*

$$\|\nabla L^\kappa(\mathbf{w})\| \leq \omega(\|\mathbf{w}\|), \qquad \left\|\nabla^2 L^\kappa(\mathbf{w})\right\| \leq \omega(\|\mathbf{w}\|), \qquad \forall \mathbf{w} \in \mathbb{W}. \tag{41}$$

Lemma B.3 (Hessian decomposition)**.** *The Hessian $\nabla^2 L^\kappa(\mathbf{w})$ at any point admits a decomposition of the form $\nabla^2 L^\kappa(\mathbf{w}) = R + K$, where R is a boundedly invertible operator (bounded with bounded inverse) and K is a bounded compact operator.*

Lemma B.4 (A Palais-Smale condition.)**.** *The function L^κ satisfies the following condition: any bounded sequence $(\mathbf{w}_n)_{n\geq 1}$ in $\mathbb{W}$ satisfying $\|\nabla L^\kappa(\mathbf{w}_n)\| \to 0$ admits a convergent subsequence.*

We are now in position to prove the main result of this section stated below.

Proposition B.5. *The dynamical system Equation (36) converges to a critical point of F^κ.*

Proof. We need to verify that assumptions Assumptions **(B)** to **(I)** in Proposition B.3 hold for $F = \lambda F^\kappa$ and $L = L^\kappa$. Note that F satisfies the same conditions in Proposition B.3 as F^κ up to a positive factor. Hence, it suffices to establish such conditions for F^κ. By Proposition B.4, we know that L^κ is a Lyapunov function of the dynamical system Equation (36). As a polynomial function, L^κ is clearly real analytic. Hence, Assumption **(B)** holds. Moreover, the system is gradient-like, since Equation (39) holds, so that Assumptions **(C)** and **(D)** are fulfilled. Lemma B.2 ensures that L^κ is coercive and bounded below, so that Assumptions **(E)** and **(F)** hold. Moreover, it also ensures that ∇L^κ and $\nabla^2 L^\kappa$ are radially bounded, as required by Assumption **(G)**. By Lemma B.1, we also have that F^κ is radially bounded, as required by Assumption **(G)**. By Lemma B.3, we are guaranteed that Assumption **(H)** on the Hessian decomposition holds. Finally, Lemma B.4 ensures that L^κ satisfies the Palais-Smale condition of Assumption **(I)**. Therefore, we can apply Proposition B.3 to deduce that Equation (36) converges to a critical point $\mathbf{w}^\star$ of F^κ. □

B.4 Characterization of Critical Points of F^κ

The following proposition, proven in Appendix F.1, provides an explicit characterization of all critical points of the vector field F^κ.

Proposition B.6. *Assume all non-zero eigenvalues of S to be distinct. Then, critical points of F^κ are exactly elements of $\mathbb{W}$ of the form:*

$$\mathbf{w} = Q\left(\sum_{p=1}^{r} E_{p,i(p)}\right) M^{\frac{1}{2}} = \sum_{p=1}^{r} m_{i(p)}^{\frac{1}{2}} q_p \otimes e^\star_{i(p)}, \tag{42}$$

where $Q = (q_p)_{1\leq p\leq k}$ is any orthogonal matrix in $\mathbb{R}^{k\times k}$, r is any integer between 0 and k, $I_{sup} = \{i(1), \dots, i(r)\} \subset I_+ \cup I_-$ is any increasing sequence of integers greater than or equal to 1, and M is a PSD operator on $\ell^2(I)$ diagonal in the canonical basis of $\ell^2(I)$ and whose only non-zero diagonal coefficients $(m_{i(p)})_{1\leq p\leq r}$ are positive roots of the equations $(x+\lambda)^2 = x^\kappa s_{i(p)}^{\kappa+1}$ of unknown x for $1 \leq p \leq r$, where $(s_i)_{i\in I}$ are the eigenvalues of the operator A. Specifically,

$$m_{i(p)}^{\frac{1}{2}} = \begin{cases} g(s_{i(p)}, \lambda) & \kappa = 0 \\ f_{\epsilon_p}(s_{i(p)}, \lambda), & \kappa = 1 \end{cases}, \tag{43}$$

where $(\epsilon_p)_{p=1}^r$ can take any values in $\{-1, 1\}$ and g, $f_\epsilon(s, \lambda)$ are given by:

$$g(s,\lambda) = \begin{cases} (\sqrt{s} - \lambda)_+^{\frac{1}{2}} & s \geq 0 \\ 0 & s < 0 \end{cases}, \qquad f_\epsilon(s,\lambda) = \frac{1}{2}\left(|s| + \epsilon\sqrt{s^2 - 4\lambda}\right) \mathbb{1}_{s^2-4\lambda\geq 0}. \tag{44}$$

The functions g and f_ϵ are spectral filters as they can filter out some eigendirections based on the value of the corresponding eigenvalue. For the case $\kappa = 1$, we will call f_1 the *stable filter* and f_{-1} the *unstable filter* for reasons that will be clear in Appendix B.5.

The key idea behind the proof of Proposition B.6 is to obtain an algebraic equation satisfied by the operator $M = (\mathbf{w}^\star)^\top \mathbf{w}^\star$. From such equation, we then show that a certain polynomial of M must commute with the diagonal operator S (containing the eigenvalues of the cross-covariance operator A). The fact that all eigenvalues of S are distinct ensures that such polynomial must also be diagonal. Then, after showing that the polynomial defines a bijective spectral map, we are able to deduce that M must also be diagonal. This diagonal structure implies that the fixed point $\mathbf{w}^\star$ must satisfy an orthogonality constraint which allows decoupling the critical point equation it satisfies into scalar equations. The fact that M is of rank at most k implies that at most k of these equations can have non-zero solution. Solving these equations explicitly yields the desired characterization.

B.5 Stability Analysis of Critical Points of F^κ

We are now interested in characterizing which critical points of the vector field F^κ are stable for a dynamics of the form $\dot{\mathbf{w}} = -\lambda F^\kappa(\mathbf{w})$, i.e. critical points for which the Jacobian of F^κ has only non-negative real parts of the eigenvalues, and those that are unstable (negative real part). The key ingredient to obtain such a characterization is to find a diagonalization of the Jacobian $\nabla F^\kappa(\mathbf{w})$ at any critical point $\mathbf{w}$ of F^κ which allows studying stability: either verify that all eigenvalues are non-negative (stable equilibrium) or identify a negative eigenvalue (unstable equilibrium).

Eigen decomposition of the Jacobian of the vector field F^κ. The following proposition, whose proof is deferred to Appendix F.2, provides the desired decomposition.

Proposition B.7 (Diagonalization of the Jacobian at critical points)**.** *Assume all non-zero eigenvalues of S to be distinct and let $\mathbf{w}$ be a critical point of $F^\kappa(\mathbf{w})$, which, by virtue of Proposition B.6, necessarily has the form:*

$$\mathbf{w} = \underbrace{Q\left(\sum_{p=1}^{r} E_{p,i(p)}\right)}_{\mathbf{w}_0} M^{\frac{1}{2}}, \tag{45}$$

where Q is an orthogonal matrix in $\mathbb{R}^{k\times k}$, r is an integer between 0 and k, $I_{sup} = \{i(1), \dots, i(r)\} \subset I_+ \cup I_-$ an increasing sequence of integers greater or equal to 1 and M is a positive semi-definite

diagonal operator on $\ell^2(I)$ whose only non-zero diagonal coefficients $(m_{i(p)})_{1\leq p\leq r}$ are positive roots of the equations $(x+\lambda)^2 = x^\kappa s_{i(p)}^{\kappa+1}$ of unknown x for $1 \leq p \leq r$. Then the Jacobian $\nabla F^\kappa(\mathbf{w})$ at $\mathbf{w}$ of the vector field admits an eigendecomposition of the form: $(\bar{E}_{p,j}, \mu_{p,j})_{1\leq p\leq k, j\in I}$, satisfying:

$$\bar{E}_{p,j} = \begin{cases} QE_{p,j}, & p > r \text{ or } j \notin I_{sup} \text{ or } (p \leq r \text{ and } j = i(p)) \\ (1-c_{p,p'})^{\frac{1}{2}} QE_{p,i(p')} + c_{p,p'}^{\frac{1}{2}} QE_{p',i(p)}, & p,p' \leq r \text{ and } j = i(p') > i(p) \\ c_{p,p'}^{\frac{1}{2}} QE_{p,i(p')} - (1-c_{p,p'})^{\frac{1}{2}} QE_{p',i(p)}, & p,p' \leq r \text{ and } j = i(p') < i(p), \end{cases} \tag{46}$$

$$\mu_{p,j} = \begin{cases} 1 - \lambda^{-2} s_j \mathbb{1}_{\kappa=0}, & j \notin I_{sup} \text{ and } p > r \\ \delta_{p,j}, & j \notin I_{sup} \text{ and } p \leq r \\ \frac{4m_j}{m_j+\lambda} - 2\mathbb{1}_{\kappa=1}, & j = i(p) \text{ and } p \leq r \\ \gamma_{p,p'}, & p,p' \leq r \text{ and } j = i(p') > i(p) \\ 0, & p' \leq r \text{ and } j = i(p') \text{ and } ((p \leq r \text{ and } j < i(p)) \text{ or } p > r) \end{cases} \tag{47}$$

with $c_{p,p'}$, $\gamma_{p,p'}$ and $\delta_{p,j}$ given by:

$$\begin{aligned} c_{p,p'} &= \frac{m_{i(p')}}{m_{i(p')} + m_{i(p)}}, \\ \gamma_{p,p'} &= \frac{\left(m_{i(p)} + m_{i(p')}\right)\left(\left|s_{i(p)}\right|^{\frac{\kappa+1}{2}} m_{i(p)}^{\frac{\kappa}{2}} + \left|s_{i(p')}\right|^{\frac{\kappa+1}{2}} m_{i(p')}^{\frac{\kappa}{2}} - s_{i(p)} s_{i(p')} \mathbb{1}_{\kappa=1}\right)}{\left(m_{i(p)} + \lambda\right)\left(m_{i(p')} + \lambda\right)} \\ \delta_{p,j} &= \frac{(s_{i(p)} - s_j)}{\lambda(m_{i(p)} + \lambda)}\left(1 + \left(m_{i(p)} s_{i(p)} - 1\right)\mathbb{1}_{\kappa=1}\right). \end{aligned} \tag{48}$$

A striking fact about the above proposition is that most eigenvectors are simply *rotated* versions of the canonical vectors $E_{p,j}$, with rotation Q corresponding to the arbitrary orthogonal matrix defining the critical point $\mathbf{w}$. The remaining ones are linear combinations of two rotated versions of paired canonical vectors $E_{p,i(p')}$ and $E_{p',i(p)}$ on which $\mathbf{w}$ admits non-zero components. The main proof strategy here is to first derive the Jacobian expression at the critical point, and express it in terms of the operator $M = \mathbf{w}^\top \mathbf{w}$ which must be diagonal by virtue of Proposition B.6. The diagonal structure implies some terms involve commuting factors which are key to simplifying the expressions. Applying such Jacobian to each rotated canonical basis element of the form $\tilde{E}_{p,j}$ and employing suitable linear algebra identities results in two types of expressions: either an expression colinear to the original element $\tilde{E}_{p,j}$, or if $j = i(p')$ and $p \leq r$ it is a linear combination of it and another rotated canonical basis element $\tilde{E}_{p',i(p)}$. We then further show that the Jacobian is stable under the subspace spanned by those two elements, which allows recovering the eigenvectors and eigenvalues corresponding to such subspace.

Stability of equilibria. Using the above eigen decomposition, we can study the stability of any critical point $\mathbf{w}$ by investigating the sign of the eigenvalues associated to each eigenvector of the Jacobian of the vector field F^κ at that point. The following propositions, proven in Appendix F.3, provide such characterization for each case $\kappa = 0$ and $\kappa = 1$ separately.

Proposition B.8 (Stability of maximal top-r critical points ($\kappa = 0$))**.** *Assume $\kappa = 0$, $0 < \lambda < \sqrt{c_1}$ (i.e. $\lambda < 1$, since $c_1 = 1$), $k \geq 1$, and that all non-zero eigenvalues of S are distinct. Let $r_{\max}$ be the largest integer such that $1 \leq r_{\max} \leq \min(k, R)$ and $\sqrt{s_{i(r_{\max})}} > \lambda$ and denote by $I_{sup} = \{i(1), \ldots, i(r_{\max})\}$ the set of indices of the $r_{\max}$ largest positive eigenvalues of A by decreasing order. Let M be the positive semi-definite diagonal operator on $\ell^2(I)$ whose only positive diagonal elements $(m_{i(p)})_{1\leq p\leq r_{\max}}$ are given by:*

$$m_{i(p)} = \sqrt{s_{i(p)}} - \lambda. \tag{49}$$

Then, for any orthogonal matrix $Q \in \mathbb{R}^{k\times k}$, the following critical point is stable for F^0:

$$\mathbf{w} = Q\left(\sum_{p=1}^{r_{\max}} E_{p,i(p)}\right) M^{\frac{1}{2}}. \tag{50}$$

Conversely, the only stable critical points of F^0 are of the above form for any orthogonal matrix Q. All other critical points of F^0 are unstable as the Jacobian of F^0 at these points necessarily admits a negative eigenvalue.

The proof requires verifying that all eigenvalues of the identified stable critical points are non-negative. To verify all the other critical points are unstable, we identify precisely unstable directions which are of two types depending on the form of the critical points: type I or II. Type I critical points are those that skip an eigen direction of A of index j_0 associated to a positive eigenvalue s_{j_0} while including another direction $i(p_0)$ with smaller eigenvalue $s_{i(p_0)}$. These possess an unstable direction associated to negative eigenvalue μ_{p_0,j_0} of $\nabla F^\kappa(\mathbf{w})$. Type II critical points contain $r < k$ eigen components without gap, but still miss an eigen direction j_0 with small positive eigenvalue that is not filtered out by the spectral filter, i.e. $g(s_{j_0}, \lambda) > 0$, with $g(s, \lambda) = (\sqrt{s} - \lambda)_+^{\frac{1}{2}}$. These possess an unstable direction associated to negative eigenvalue μ_{r+1,j_0} of $\nabla F^\kappa(\mathbf{w})$.

Proposition B.9 (Stability of the top-r critical points ($\kappa = 1$))**.** *Assume $\kappa = 1$, $\lambda \leq 1$ and that all non-zero eigenvalues of S are distinct. Let r^+ and r^- be any non-negative integers in $[0, R]$ such that $r = r^+ + r^- \leq k$. Denote by $I^+_{sup} = \{i(1), \dots, i(r^+)\}$ and $I^-_{sup} = \{i(r^+ + 1), \dots, i(r)\}$, the sets of indices of the r^+ largest positive eigenvalues and r^- smallest negative eigenvalues of A satisfying $|s_j| \geq 2\sqrt{\lambda}$ for all $j \in I_{sup} = I^+_{sup} \cup I^-_{sup}$. We use the convention that $I^+_{sup} = \emptyset$ if $r^+ = 0$ and $I^-_{sup} = \emptyset$ if $r^- = 0$. Let M be the positive semi-definite diagonal operator on $\ell^2(I)$ whose diagonal elements $m_j = 0$ for $j \notin I_{sup}$, whereas the elements $(m_{i(p)})_{1 \leq p \leq r}$ are given by:*

$$m_{i(p)}^{\frac{1}{2}} = \frac{1}{2}\left(\left|s_{i(p)}\right| + \sqrt{s_{i(p)}^2 - 4\lambda}\right) \mathbb{1}_{s_{i(p)}^2 - 4\lambda \geq 0}. \tag{51}$$

Then, for any orthogonal matrix $Q \in \mathbb{R}^{k \times k}$, the following critical point is stable for F^1:

$$\mathbf{w} = Q\left(\sum_{p=1}^{r} E_{p,i(p)}\right) M^{\frac{1}{2}}. \tag{52}$$

Conversely, the only stable critical points of F^1 are of the above form for any orthogonal matrix Q and any $0 \leq r^+, r^- \leq R$ satisfying $r^+ + r^- \leq k$. All other critical points are unstable.

The proof of Proposition B.9 proceeds similarly as that of Proposition B.8. To verify all the other critical points are unstable, we identify precisely unstable directions which are, again, of two types depending on the form of the critical points: type I or II. Type I critical points are those that include an eigen direction of A of index $i(p_0)$ but weighted by the unstable spectral filter f_{-1} instead of the stable one f_1. These possess an unstable direction associated to negative eigenvalue $\mu_{p_0,i(p_0)}$ of $\nabla F^\kappa(\mathbf{w})$. Type II critical points all use the stable spectral filter f_1 but present a gap in their spectra: they miss an eigen direction of index j_0 of eigenvalue s_{j_0} but contain an eigendirection indexed by $i(p_0)$ of eigenvalue $s_{i(p_0)}$ with the same sign as s_{j_0} but smaller in absolute value: $|s_{j_0}| > \left|s_{i(p_0)}\right|$. These critical points possess an unstable direction associated to negative eigenvalue μ_{p_0,j_0} of $\nabla F^\kappa(\mathbf{w})$.

B.6 Degenerate canonical spectra and shared encoders

Repeated canonical correlations. Propositions 4.3 and 4.4 assume distinct canonical correlations c_i. With repeated correlations, the characterization of critical points still holds: critical points remain in the span of the nonlinear CCA directions, but their expression involves an additional arbitrary rotation within each degenerate canonical subspace. The proof structure is unchanged, except that some diagonal operators, such as M in Proposition B.8, become block-diagonal, with one block per degenerate subspace. We present the non-degenerate case for clarity. For stability, we conjecture that the same conclusions hold; a rigorous proof requires a finer study of the singular vectors of the Hessian operator within each degenerate block, which we leave for future work. The clustered spectrum of Appendix G.11 probes this case empirically.

Shared encoder. Propositions 4.3 and 4.4 are proved for the two-encoder objective, in which U and V vary independently. The shared-encoder variant of Algorithm 1 ($V = U$) is a practical specialization for symmetric views ($\mathbb{P}_X = \mathbb{P}_Y$ and a symmetric joint law). In that case, both encoders must be equal at the equilibria above, and stability follows by a similar analysis restricted to $V = U$. We present the general analysis because it does not rely on the symmetric-views assumption.

C Proofs of the main results

C.1 Gradient flow of the PEIRA objective

C.1.1 Convergence to critical points and characterization of critical points

Proof of Proposition 4.3. By Proposition B.1, we have the following equivalence between the gradient flow $(W_t)_{t\geq 0}$ in $\mathcal{W}$ and an ODE defined on its components $(\mathbf{w}_t)_{t\geq 0}$ in the eigenbasis of A:

$$\dot{W}_t = -\nabla \mathcal{E}_{\text{PEIRA}}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^0(\mathbf{w}_t), \tag{53}$$

where the vector field F^0 is defined in Equation (30). Thus we can apply the results obtained for the dynamics arising from the vector field F^0. Specifically, we directly apply Proposition B.5 to such dynamics, which ensures convergence to a critical point $\mathbf{w}^\star$.

Consider now the re-ordering $\sigma = \{1, \dots, R\} \to I_+$ of indices associated to positive eigenvalues of A from Proposition A.1, so that $(s_{\sigma(i)})_{i=1}^R$ are non-increasing with $R = |I_+|$ possibly infinite. Then, by Proposition B.6, we know that such critical point is necessarily of the form:

$$\mathbf{w}^\star = \sum_{i\in D} \left(\sqrt{s_{\sigma(i)}} - \lambda\right)_+^{\frac{1}{2}} q_i \otimes e_{\sigma(i)} \tag{54}$$

where D is any, possibly empty, subset of $\{1, \dots, R\}$ of at most k elements, $(q_i)_{i\in D}$ is an orthonormal family in $\mathbb{R}^k$ and $(e_i)_{i\in I}$ is the canonical basis of $\ell^2(I)$. It remains to express critical points $W^\star = (U^\star, V^\star)$ of the gradient flow of $\mathcal{E}_{\text{PEIRA}}$ in the eigenbasis of A using the characterization of their components in Equation (54). Let $W^\star$ be an element in $\mathcal{W}$ whose components in the eigenbasis $(\zeta_i)_{i\in I}$ are $\mathbf{w}^\star$. Then, by definition, we directly have:

$$W^\star(x,y) = \sum_{i\in D} q_i \left(\sqrt{s_{\sigma(i)}} - \lambda\right)_+^{\frac{1}{2}} \zeta_{\sigma(i)}(x,y), \tag{55}$$

Moreover, by Proposition A.1, we know that for any $1 \leq i \leq R$, the elements $(\psi_i)_{i=1}^R$ and $(\xi_i)_{i=1}^R$ of $L_2(\mathbb{P}_X)$ and $L_2(\mathbb{P}_Y)$ given by $(\psi_i, \xi_i) = \sqrt{2}\zeta_{\sigma(i)}$ form orthonormal families satisfying $\mathbb{E}\left[\psi_i(x)\xi_j(y)\right] = c_i \mathbb{1}_{i=j}$ with $c_i = s_{\sigma(i)}$. Consequently, we can write $W^\star = (U^\star, V^\star)$ with:

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{i\in D} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \psi_i(x), \qquad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{i\in D} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \xi_i(y). \tag{56}$$

□

C.1.2 Minimizers span thresholded canonical correlation subspaces

Proof of Proposition 4.4. By Proposition B.1, we have the following equivalence between the gradient flow $(W_t)_{t\geq 0}$ in $\mathcal{W}$ and an ODE defined on its components $(\mathbf{w}_t)_{t\geq 0}$ in the eigenbasis of A:

$$\dot{W}_t = -\nabla \mathcal{E}_{\text{PEIRA}}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^0(\mathbf{w}_t), \tag{57}$$

where the vector field F^0 is defined in Equation (30). Thus, we only need to study the stability of critical points of F^0. This is a direct consequence of Proposition B.8 which ensures that the only stable critical points of F^0 are of the form:

$$\mathbf{w}^\star = \sum_{i=1}^{r_{\max}} \left(\sqrt{s_{\sigma(i)}} - \lambda\right)_+^{\frac{1}{2}} q_i \otimes e_{\sigma(i)} \tag{58}$$

with $r_{\max}$ the largest integer such that $\sqrt{s_{\sigma(r_{\max})}} > \lambda$ and $r_{\max} \leq \min(R, k)$. These are the coordinates of the following elements $(U^\star, V^\star) \in \mathcal{W}$ which are the critical points of $\nabla \mathcal{E}_{\text{PEIRA}}$ described in Proposition 4.3:

$$U^\star(x) = \frac{1}{\sqrt{2}} \sum_{i=1}^{r_{\max}} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \psi_i(x), \qquad V^\star(y) = \frac{1}{\sqrt{2}} \sum_{i=1}^{r_{\max}} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \xi_i(y). \tag{59}$$

We will now show that these stable solutions are precisely the global minimizers of $W \mapsto \mathcal{E}_{\text{PEIRA}}(W)$. Since Assumption (A) holds, we know by Proposition A.1 that A is a compact operator and that there exists an orthonormal basis $\mathcal{B} = (\zeta_i)_{i\in I}$ of $\mathcal{H}$ consisting of eigenvectors of A, i.e. $A\zeta_i = s_i\zeta_i$, with $(s_i)_{i\in I}$ being the eigenvalues of A.

We first treat the case where $k \leq dim(\mathcal{H})$. Consider $W \in \mathcal{W}$. Such element can be viewed as a Hilbert-Schmidt operator from $\mathcal{H}$ to $\mathbb{R}^k$. It is therefore compact, and thus, by the spectral theorem for compact operators [Gohberg et al., 2003], it admits a singular value decomposition of the form:

$$W = Q\Sigma T^*, \tag{60}$$

where $Q \in O(k)$, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k)$, and $T : \mathbb{R}^k \to \mathcal{H}$ is an isometric embedding. Then

$$WW^T = Q\,\text{diag}(\sigma_1^2, \dots, \sigma_k^2)Q^*, \tag{61}$$

and, by cyclicity of the trace,

$$\mathcal{E}_{\text{PEIRA}}(W) = -\frac{1}{2}\,\text{tr}\Big(T^*AT\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big) + \frac{\lambda}{2}\sum_{i=1}^{k}\sigma_i^2. \tag{62}$$

Moreover, let us decompose A into the difference of two positive semi-definite operators A_+ and A_-, i.e. $A = A_+ - A_-$, with A_+ containing all non-negative eigendirections, while A_- contains the absolute value of the negative ones. More precisely, $A\zeta_i = A_+\zeta_i = s_i\zeta_i$ and $A_-\zeta_i = 0$ for all indices i such that $s_i \geq 0$, whereas $A\zeta_i = -A_-\zeta_i = s_i\zeta_i$ and $A_+\zeta_i = 0$, for all indices such that $s_i < 0$. It is easy to see that

$$\begin{aligned}\text{tr}\Big(T^*AT\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big) &= \text{tr}\Big(T^*(A_+ - A_-)T\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big)\\ &\leq \text{tr}\Big(T^*A_+T\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big),\end{aligned} \tag{63}$$

since A_- is positive semi-definite. For fixed (σ_i), the numbers $d_i := \frac{\sigma_i^2}{\sigma_i^2+\lambda} \geq 0$, we apply von Neumann's trace inequality [Grigorieff, 1991], which implies that the trace of the product of matrices T^*A_+T and $diag(d_i)$ is upper-bounded by the sum of products of their singular values ordered by non-increasing order:

$$\text{tr}\Big(T^*A_+T\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big) \leq \sum_{i=1}^{k}\tilde{s}_i d_{\beta(i)}, \tag{64}$$

where $(\tilde{s}_i)_{1\leq i\leq k}$ are the singular values of T^*A_+T ordered by non-increasing order, whereas $\beta(1), \dots, \beta(k)$ is a re-indexing to ensure that $d_{\beta(1)} \geq \cdots \geq d_{\beta(k)}$. Since both matrices are PSD, these singular values are also eigenvalues. Thus, by Courant-Fisher's theorem [Bhatia, 2013, Corollary III.1.2], such eigenvalues are upper-bounded by the top-k eigenvalues values of A_+ ordered by non-increasing order, which we denote as $s_1^+ \geq \cdots \geq s_k^+$, so that:

$$\sup_{T^*T=I}\text{tr}\Big(T^*A_+T\ \text{diag}\Big(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\Big)\Big) \leq \sum_{i=1}^{k} s_i^+ d_{\beta(i)}. \tag{65}$$

The above inequality with Equation (63) yields:

$$\inf_W \mathcal{E}_{\text{PEIRA}}(W) \geq \min_{\sigma_1,\dots\sigma_k}\frac{1}{2}\sum_{i=1}^{k}\left(-s_i^+\frac{\sigma_{\beta(i)}^2}{\sigma_{\beta(i)}^2+\lambda} + \lambda\sigma_{\beta(i)}^2\right) = \min_{\sigma_1,\dots\sigma_k}\frac{1}{2}\sum_{i=1}^{k} f_{s_i^+}(\sigma_{\beta(i)}), \tag{66}$$

for $f_s(\sigma)$ defined as follows:

$$f_s(\sigma) := \left(-s\frac{\sigma^2}{\sigma^2+\lambda} + \lambda\sigma^2\right) \qquad \sigma \geq 0. \tag{67}$$

Thus the problem decouples into k scalar minimizations of the functions $\sigma \mapsto f_{s_i^+}(\sigma)$. By studying the variations of such a function, we see that the minimum value is given by $\sigma_i^\star = (\sqrt{s_i^+} - \lambda)_+^{\frac{1}{2}}$,

whenever $s_i^+ > 0$ and by $\sigma_i^\star = 0$ otherwise. Moreover, when $s_i^+ > 0$, this corresponds exactly to the positive eigenvalue s_i of operator A (without loss of generality we assumed that top-k eigenvalues of A are $s_1, \dots, s_k$). We know, by Proposition A.1, that to each of the R positive eigenvalues of A corresponds a negative one with opposite sign and vice versa, with R possibly infinite. Moreover, $R \geq 1$ since we know that 1 is necessarily an eigenvalue of A. Therefore $\mathcal{E}_{\text{PEIRA}}$ satisfies:

$$\inf_{W \in \mathcal{W}} \mathcal{E}_{\text{PEIRA}}(W) \geq \frac{1}{2} \sum_{i=1}^{k} f_{s_i^+}(\sigma_i^\star) = \frac{1}{2} \sum_{i=1}^{\min(k,R)} f_{s_i}(\sigma_i^\star) = -\frac{1}{2} \sum_{i=1}^{\min(k,R)} (\sqrt{s_i} - \lambda)_+^2. \tag{68}$$

Furthermore, denoting by $r_{\max} \leq k$ the number of non-zero components, i.e. those for which $\sqrt{s_i} > \lambda$ and which we know satisfy $r_{\max} \leq \min(R, k)$, we directly get

$$\inf_{W \in \mathcal{W}} \mathcal{E}_{\text{PEIRA}}(W) \geq -\frac{1}{2} \sum_{i=1}^{r_{\max}} (\sqrt{s_i} - \lambda)_+^2. \tag{69}$$

This lower bound is clearly attained for W of the form

$$W^\star(x, y) = \sum_{i=1}^{r_{\max}} q_i (\sqrt{s_i} - \lambda)_+^{\frac{1}{2}} \zeta_i(x, y), \tag{70}$$

where $(q_i)_{i=1}^{r_{\max}}$ are any orthonormal family of $\mathbb{R}^k$. Conversely, any minimizer of $\mathcal{E}_{\text{PEIRA}}$ is of the form in Equation (70). To see this, let W be a minimizer of $\mathcal{E}_{\text{PEIRA}}$. Since $\mathcal{E}_{\text{PEIRA}}$ is differentiable, then W must be a critical point of $\mathcal{E}_{\text{PEIRA}}$. By Proposition 4.3, we know such critical points must be of the form $W = (U, V)$ with

$$U(x) = \frac{1}{\sqrt{2}} \sum_{i \in D} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \psi_i(x), \qquad V(y) = \frac{1}{\sqrt{2}} \sum_{i \in D} q_i \left(\sqrt{c_i} - \lambda\right)_+^{\frac{1}{2}} \xi_i(y), \tag{71}$$

where D is any, possibly empty, subset of $\{1, \dots, R\}$ of at most k elements, $(q_i)_{i \in D}$ is an orthonormal family in $\mathbb{R}^k$. Evaluating $\mathcal{E}_{\text{PEIRA}}$ at such point yields:

$$\mathcal{E}_{\text{PEIRA}}(W) = -\frac{1}{2} \sum_{i \in D} (\sqrt{c_i} - \lambda)_+^2. \tag{72}$$

The above quantity is minimized exactly when $(c_i)_{i \in D}$ corresponds to the top $r_{\max}$-singular values of A, since these are assumed to be distinct. These are also equal to the top-$r_{\max}$ positive eigenvalues of A by recalling the structure of the eigenvalues of A from Proposition A.1. Since, we assumed, without loss of generality that $s_1, \dots, s_k$ are also non-increasing, then we may set $c_i = s_i$ and $(\psi_i, \xi_i) = \sqrt{2}\zeta_i$ for $1 \leq i \leq r_{\max}$ by application of Proposition A.1, thus recovering an expression of the form in Equation (70) for the minimizer W.

The case when $k > \dim(\mathcal{H})$ is treated similarly with the difference that the SVD of W contains at most $\dim(\mathcal{H})$ non-zero singular values, so that k is replaced by $\dim(\mathcal{H})$ in Equation (66) and the same conclusion follows.

□

C.1.3 The gradient of PEIRA's objective matches that of the auxiliary loss $\mathcal{L}_{\text{aux}}$

Proof of Proposition 5.1. Recall the auxiliary loss defined in the proposition, which we can express in terms of the operator A:

$$\begin{aligned} \mathcal{L}_{\text{aux}}(U, V; P, Q) =& \frac{1}{2} Tr\left((QP + \lambda I)\, \mathbb{E}\left[U(x)U(x)^\top + V(y)V(y)^\top\right]\right) \\ & - Tr\left(Q \mathbb{E}\left[U(x)V(y)^\top + V(y)U(x)^\top\right]\right) \\ =& \frac{1}{2} Tr\left((QP + \lambda I)\, WW^\top - QWAW^\top\right), \end{aligned} \tag{73}$$

where we replaced expectations by $WW^\top$ and $WAW^\top$. For P and Q such that Q and QP are symmetric, we get by direct differentiation w.r.t. W that:

$$\langle \partial_W \mathcal{L}_{\text{aux}}(W; P, Q), \delta W \rangle_{\mathcal{W}} = -Tr\left(\delta W \left(QWA - (QP + \lambda I)W\right)^\top\right). \tag{74}$$

In particular, the above expression holds when $P = P_W$ and $Q = Q_W$, since Q_W and $Q_W P_W$ are symmetric, by definition. Hence, we have shown that $\partial_W \mathcal{L}_{\text{aux}}(W, P_W, Q_W) = -Q_W W A + (Q_W P_W + \lambda I)W$. Using the property $\left(WW^\top + \lambda I\right)^{-1} W = W \left(W^\top W + \lambda I\right)^{-1}$ and recalling that A is bounded directly yields:

$$\begin{aligned} \partial_W \mathcal{L}_{\text{aux}}(W, P_W, Q_W) &= -Q_W W A + (Q_W P_W + \lambda I)W \\ &= \lambda W \left(I - \left(W^\top W + \lambda I\right)^{-1} A \left(W^\top W + \lambda I\right)^{-1}\right). \end{aligned} \tag{75}$$

The above expression matches that of $\nabla \mathcal{E}_{\text{PEIRA}}(W)$ from Proposition B.1. Thus, we have shown that $\partial_W \mathcal{L}_{\text{aux}}(W, P_W, Q_W) = \nabla \mathcal{E}_{\text{PEIRA}}(W)$. Finally, by the same proposition we know that the objective $\mathcal{E}_{\text{PEIRA}}$ is smooth, has a Lipschitz gradient with Lipschitz constant $\frac{4}{\lambda}$. □

C.2 Self-distillation dynamics

Here we provide a proof of Proposition 4.1 following a similar scheme as for PEIRA. We will first express the vector field driving the self-distillation dynamics in terms of the operator A, then proceed to prove the main result.

C.2.1 Convergence of the self-distillation dynamics and characterization of equilibria

Using the vector field expression in Proposition B.2, we will be able to prove Proposition 4.1 by reducing the problem to the components of the dynamics in the eigenbasis of the operator A.

Proof of Proposition 4.1. Denote by $\mathbf{w} = (\mathbf{w}_i)_{i \in I}$ the components of an element $W \in \mathcal{W}$ in the eigenbasis $(\zeta_i)_{i \in I}$ of A (see Equation (18)). By Proposition B.2, we have the following equivalence between the dynamics $(W_t)_{t \geq 0}$ in $\mathcal{W}$ driven by the vector field $\mathcal{F}$ and an ODE defined on its components $(\mathbf{w}_t)_{t \geq 0}$ in the eigenbasis of A driven by a vector field F^1 defined in Equation (33):

$$\dot{W}_t = -\mathcal{F}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^1(\mathbf{w}_t). \tag{76}$$

Thus it suffices to study the dynamics of the components, which is precisely the analysis in Appendix B.

Specifically, we directly apply Proposition B.5 to the dynamics driven by the vector field F^1, which ensures convergence to a critical point $\mathbf{w}^\star$.

Then, by Proposition B.6, we know that such critical point is necessarily of the form:

$$\mathbf{w}^\star = \sum_{i=1}^{r} f_{\epsilon_i}(s_{\alpha(i)}, \lambda) q_i \otimes e_{\alpha(i)}, \qquad f_\epsilon(s, \lambda) = \frac{1}{2}\left(|s| + \epsilon\sqrt{s^2 - 4\lambda}\right) \mathbb{1}_{|s|^2 - 4\lambda \geq 0}, \tag{77}$$

where r is any non-negative integer such that $0 \leq r \leq k$, the vectors $(q_i)_{i=1}^r$ form an orthonormal family in $\mathbb{R}^k$, $(e_i)_{i \in I}$ is the canonical basis of $\ell^2(I)$, $(\alpha(i))_{i=1}^r$, are distinct indices associated to non-zero eigenvalues $(s_{\alpha(i)})_{i=1}^r$ and ϵ_i are numbers taking two possible values -1 or 1. Hence, $\mathbf{w}^\star$ are the components of the following critical point $W^\star$ of the vector field $\mathcal{F}$:

$$W^\star(x, y) = \sum_{i=1}^{r} q_i f_{\epsilon_i}(s_{\alpha(i)}, \lambda) \zeta_{\alpha(i)}(x, y). \tag{78}$$

This is equivalent to the characterization in the statement of the result after noticing that $\left|s_{\alpha(i)}\right| = c_{j(i)}$ for the corresponding canonical index $j(i) \in \{1, \ldots, R\}$, and that eigenvectors associated to positive eigenvalues are of the form $\zeta_{\alpha(i)} = \frac{1}{\sqrt{2}}(\psi_{j(i)}, \xi_{j(i)})$, whereas those associated to negative eigenvalues are of the form $\frac{1}{\sqrt{2}}(\psi_{j(i)}, -\xi_{j(i)})$. □

C.2.2 Stability of equilibria of the self-distillation dynamics

Proof of Proposition 4.2. By Proposition B.2, we have the following equivalence between the dynamics $(W_t)_{t \geq 0}$ in $\mathcal{W}$ driven by the vector field $\mathcal{F}$ and an ODE defined on its components $(\mathbf{w}_t)_{t \geq 0}$ in the eigenbasis of A driven by a vector field F^1 defined in Equation (33):

$$\dot{W}_t = -\mathcal{F}(W_t) \iff \dot{\mathbf{w}}_t = -\lambda F^1(\mathbf{w}_t). \tag{79}$$

In particular, a point $W^\star \in \mathcal{W}$ is stable for $\mathcal{F}$ iff its coordinate representation $\mathbf{w}^\star$ is stable for F^1.

Stability of critical points of F^1 is studied in Proposition B.9, which implies that, amongst all critical points of the vector field F^1, the only ones that are stable are of the form:

$$\mathbf{w}^\star = \sum_{i=1}^{r^+} f_1(s_{\alpha^+(i)}, \lambda) q_i^+ \otimes e_{\alpha^+(i)} + \sum_{i=1}^{r^-} f_1(s_{\alpha^-(i)}, \lambda) q_i^- \otimes e_{\alpha^-(i)} \tag{80}$$

where r^+ and r^- are non-negative integers in $[0, R]$ satisfying $0 \leq r^+ + r^- \leq k$, $(q_i^\pm)_{\sigma\in\{+,-\}, i\in\{1,\dots,r^\sigma\}}$ are any orthonormal vectors in $\mathbb{R}^k$, and where $\alpha^+(1), \dots, \alpha^+(r^+)$ are the indices of the top-r^+ eigenvalues of A by decreasing order, whereas $\alpha^-(1), \dots, \alpha^-(r^-)$ are the indices of the bottom-r^- eigenvalues of A by increasing order. We get the desired result by expressing the element $W^\star$ based on its coordinates $\mathbf{w}$.

□

D Proofs of the reduction to coordinate representations

D.1 Proof of Proposition B.1: Expressions of the gradient of PEIRA's objective

Proof of Proposition B.1. Let (U, V) be in $\mathcal{U} \times \mathcal{V}$. We start by expressing the objective in terms of the operator A, then its gradient.

Expression of the objective. Recall the expression of $\mathcal{E}_{\text{PEIRA}}(U, V)$:

$$\mathcal{E}_{\text{PEIRA}}(U, V) = -\frac{1}{2} Tr\big(P_{U,V}\big) + \frac{\lambda}{2} \mathbb{E}\big[\|U(x)\|_2^2 + \|V(y)\|_2^2\big]\,. \tag{81}$$

A first step is to express $P_{U,V}$ in closed form in terms of W and operator A. As the solution of the linear regression problem in Equation (3), $P_{U,V}$ admits the following expression:

$$\begin{aligned} P_{U,V} &= \left(\mathbb{E}\left[U(x)V(y)^\top + V(y)U(x)^\top\right]\right)\left(\mathbb{E}\left[U(x)U(x)^\top + V(y)V(y)^\top\right] + \lambda I\right)^{-1} \\ &= WAW^\top \left(WW^\top + \lambda I\right)^{-1}. \end{aligned} \tag{82}$$

Therefore, $\mathcal{E}_{\text{PEIRA}}(U, V)$ can be written as:

$$\mathcal{E}_{\text{PEIRA}}(U, V) = \mathcal{E}_{\text{PEIRA}}(W) = -\frac{1}{2} Tr\left(WAW^\top \left(WW^\top + \lambda I\right)^{-1}\right) + \frac{\lambda}{2} Tr\left(WW^\top\right). \tag{83}$$

Hence, $\mathcal{E}_{\text{PEIRA}}$ is a rational function of W. To prove smoothness, recall that by definition, A is a bounded operator. Hence, $WAW^\top$ is a smooth function of W, i.e. infinitely differentiable. Moreover, since $\lambda > 0$, then $WW^\top + \lambda I$ is smooth and invertible for any W, therefore $W \mapsto Q_W = (WW^\top + \lambda I)^{-1}$ is smooth by composition with the inverse map. Finally, $\mathcal{E}_{\text{PEIRA}}$ is also smooth by composition of smooth functions.

Gradient expression. The expression of the gradient is obtained by direct differentiation along any direction $\delta W \in \mathcal{W}$

$$\begin{aligned} \langle \nabla \mathcal{E}_{\text{PEIRA}}(W), \delta W\rangle_{\mathcal{W}} &= -\frac{1}{2} Tr\left(\left(\delta W AW^\top + WA\delta W^\top\right) Q_W\right) + \lambda Tr(\delta W W^\top) \\ &\quad + \frac{1}{2} Tr\left(WAW^\top Q_W \left(\delta W W^\top + W\delta W^\top\right) Q_W\right) \\ &= -Tr\left(\delta W AW^\top Q_W\right) + \lambda Tr(\delta W W^\top) \\ &\quad + Tr\left(\delta W W^\top Q_W W AW^\top Q_W\right) \\ &= -Tr\left(\delta W \left(AW^\top Q_W - W^\top Q_W P_W - \lambda W^\top\right)\right) \\ &= -Tr\left(\delta W \left(Q_W WA - (Q_W P_W + \lambda I) W\right)^\top\right). \end{aligned} \tag{84}$$

This shows that the gradient is given by:

$$\nabla \mathcal{E}_{\text{PEIRA}}(W) = -Q_W WA + Q_W P_W W + \lambda W. \tag{85}$$

To get the second expression, recall that $P_W = WAW^\top(WW^\top + \lambda I)^{-1} = WAW^\top Q_W$. Hence,

$$\begin{aligned}
\nabla \mathcal{E}_{\text{PEIRA}}(W) &= -Q_W WA + Q_W P_W W + \lambda W \\
&= Q_W WA\left(W^\top (WW^\top + \lambda I)^{-1} W - I\right) + \lambda W \\
&= Q_W WA\left((W^\top W + \lambda I)^{-1} W^\top W - I\right) + \lambda W \\
&= -\lambda Q_W WA(W^\top W + \lambda I)^{-1} + \lambda W \\
&= -\lambda W\left(W^\top W + \lambda I\right)^{-1} A\left(W^\top W + \lambda I\right)^{-1} + \lambda W \\
&= \lambda W\left(I - \left(W^\top W + \lambda I\right)^{-1} A\left(W^\top W + \lambda I\right)^{-1}\right),
\end{aligned} \tag{86}$$

where we used the following property $(WW^\top + \lambda I)^{-1} W = W(W^\top W + \lambda I)^{-1}$ in the third and fifth lines.

To show that the gradient is Lipschitz, we will show that its Hessian is bounded over $\mathcal{W}$. Let us first express the action of the Hessian on a direction δW by differentiating the second expression of the gradient along such direction. For conciseness we define $R = (W^\top W + \lambda I)^{-1}$ and get:

$$\begin{aligned}
\nabla^2 \mathcal{E}_{\text{PEIRA}}(W)[\delta W] =& \lambda \delta W\left(I - RAR\right) + \lambda WR\left(\delta W^\top W + W^\top \delta W\right) RAR \\
&+ \lambda WRAR\left(\delta W^\top W + W^\top \delta W\right) R.
\end{aligned} \tag{87}$$

Let us now find an upper bound C on the operator norm of the Hessian. To this end it suffices to find an upper bound $\left\|\nabla^2 \mathcal{E}_{\text{PEIRA}}(W)[\delta W]\right\|$ of the form $\left\|\nabla^2 \mathcal{E}_{\text{PEIRA}}(W)[\delta W]\right\| \leq C \left\|\delta W\right\|$. Standard operator norm inequalities yield:

$$\begin{aligned}
\left\|\nabla^2 \mathcal{E}_{\text{PEIRA}}(W)[\delta W]\right\| \leq& \lambda\left(\left\|RAR\right\|_{op} + 1\right)\left\|\delta W\right\| \\
&+ \lambda \left\|WR\right\|_{op} \left\|WRAR\right\|_{op} \left\|\delta W\right\| + \lambda \left\|WRW^\top\right\|_{op} \left\|RAR\right\|_{op} \left\|\delta W\right\| \\
&+ \lambda \left\|WRAR\right\|_{op} \left\|WR\right\|_{op} \left\|\delta W\right\| + \lambda \left\|WRARW^\top\right\|_{op} \left\|R\right\|_{op} \left\|\delta W\right\|.
\end{aligned} \tag{88}$$

Define C_0 as:

$$\begin{aligned}
C_0 =& \lambda\left(1 + \left\|RAR\right\|_{op} + 2\left\|WR\right\|_{op}\left\|WRAR\right\|_{op}\right) \\
&+ \lambda\left(\left\|WRW^\top\right\|_{op}\left\|RAR\right\|_{op} + \left\|WRARW^\top\right\|_{op}\left\|R\right\|_{op}\right)
\end{aligned} \tag{89}$$

We have shown, so far that $\left\|\nabla^2 \mathcal{E}_{\text{PEIRA}}(W)\right\|_{op} \leq C_0$. It remains to show that C_0 is uniformly bounded w.r.t. W. To this end, let us further upper-bound C_0:

$$C_0 \leq \lambda\left(1 + \left\|A\right\|_{op}\left(\left\|R\right\|_{op}^2 + 3\left\|WR\right\|_{op}^2\left\|R\right\|_{op} + \left\|R\right\|_{op}^2\left\|WRW^\top\right\|_{op}\right)\right) \tag{90}$$

Recalling that $R = (W^\top W + \lambda I)^{-1}$, we directly have that $\left\|R\right\|_{op} \leq \lambda^{-1}$. Moreover, since W has finite dimensional range (of dimension k), it must be a compact operator, so that it admits a singular-value decomposition, by the spectral theorem for compact operators [Gohberg et al., 2003]. Using such a decomposition, we can show that the singular values of WR and $WRW^\top$ are given by $\frac{w_i}{w_i^2+\lambda}$ and $\frac{w_i^2}{w_i^2+\lambda}$ where w_i are the singular values of W. These quantities are uniformly bounded by $\frac{1}{2\sqrt{\lambda}}$ and 1 respectively. Consequently, we have $\left\|WR\right\|_{op} \leq \frac{1}{2\sqrt{\lambda}}$ and $\left\|WRW^\top\right\|_{op} \leq 1$. Finally, recalling that $\left\|A\right\|_{op} \leq 1$ by Equation (19) it follows that:

$$C_0 \leq \lambda\left(1 + \frac{11}{4\lambda^2}\right) \leq \frac{4}{\lambda}, \tag{91}$$

where we used that $\lambda < 1$. This allows us to conclude that the gradient $\nabla \mathcal{E}_{\text{PEIRA}}$ is Lipschitz, with Lipschitz constant $\frac{4}{\lambda}$.

Components of the gradient in the eigenbasis of A. Using the decomposition in Equation (24) of $WAW^\top$ and $WW^\top$ in such basis and recalling the expression of the gradient $\nabla \mathcal{E}_{\text{PEIRA}}(W)$ above, we directly get the desired expression for its components $\mathbf{g}$ in the basis $(\zeta_i)_{i\in I}$ in terms of $\mathbf{w}$ and the diagonal operator S which is given by the vector field F^0. Equivalence between the ODE in $\mathcal{W}$ and the ODE of its coordinates in $\mathbb{W}$ follows directly. □

D.2 Proof of Proposition B.2: Expressions of the self-distillation vector field

Proof of Proposition B.2. Let us first express the partial derivatives of $\mathcal{P}_{U \triangleright V}(P, U, V)$ along some direction δU:

$$\begin{aligned}\langle \partial_U \mathcal{P}_{U\triangleright V}(P,U,V), \delta U\rangle_{\mathcal{U}} =& \mathbb{E}\left[(P\delta U(x))^\top (PU(x) - V(y))\right] + \lambda \mathbb{E}\left[\delta U(x)^\top U(x)\right] \\ =& Tr\left(P^\top P \mathbb{E}\left[U(x)(\delta U(x))^\top\right] - \mathbb{E}\left[V(y)(\delta U(x))^\top\right] P^\top\right) \\ &+ \lambda Tr\left(\mathbb{E}\left[U(x)(\delta U(x))^\top\right]\right)\end{aligned} \tag{92}$$

A similar expression can be obtained for $\mathcal{P}_{V \triangleright U}(P, V, U)$ along some perturbation δV, by exchanging the roles of U and V. Hence, denoting by $W = (U, V)$ and $\delta W = (\delta U, \delta V)$ and summing both expressions above yields:

$$\begin{aligned}\langle \partial_U \mathcal{P}_{U\triangleright V}(P,U,V), \delta U\rangle_{\mathcal{U}} + & \langle \partial_V \mathcal{P}_{V\triangleright U}(P,V,U), \delta V\rangle_{\mathcal{V}} \\ =& Tr\left(P^\top P \mathbb{E}\left[U(x)(\delta U(x))^\top + V(y)(\delta V(y))^\top\right]\right) \\ &- Tr\left(\mathbb{E}\left[V(y)(\delta U(x))^\top + U(x)(\delta V(y))^\top\right] P^\top\right) \\ &+ \lambda Tr\left(\mathbb{E}\left[U(x)(\delta U(x))^\top + V(y)(\delta V(y))^\top\right]\right) \\ =& Tr\left(P^\top PW(\delta W)^\top\right) - Tr\left(WA(\delta W)^\top P^\top\right) \\ &+ \lambda Tr\left(W(\delta W)^\top\right),\end{aligned} \tag{93}$$

where we viewed W as a Hilbert-Schmidt operator from $\mathcal{H}$ to $\mathbb{R}^k$. Therefore, the vector field $(\partial_U \mathcal{P}_{U\triangleright V}(P,U,V), \partial_V \mathcal{P}_{V\triangleright U}(P,V,U))$ is given by:

$$(\partial_U \mathcal{P}_{U\triangleright V}(P,U,V), \partial_V \mathcal{P}_{V\triangleright U}(P,V,U)) = P^\top PW - P^\top WA + \lambda W. \tag{94}$$

Thus we only need to evaluate the above expression at optimal regressor $P_{U,V}$ to get an expression for the vector field $\mathcal{F}(W)$ driving the self-distillation dynamics. To this end, note that $P_{U,V}$ can be expressed in closed form as follows:

$$\begin{aligned}P_{U,V} &= \left(\mathbb{E}\left[U(x)V(y)^\top + V(y)U(x)^\top\right]\right)\left(\mathbb{E}\left[U(x)U(x)^\top + V(y)V(y)^\top\right] + \lambda I\right)^{-1} \\ &= WAW^\top \left(WW^\top + \lambda I\right)^{-1},\end{aligned} \tag{95}$$

Consequently, using the above expression along with the one on partial derivatives $(\partial_U \mathcal{P}_{U\triangleright V}, \partial_V \mathcal{P}_{V\triangleright U})$, yields an expression for the vector field $\mathcal{F}(W)$:

$$\begin{aligned}\mathcal{F}(W) &= (P_{U,V})^\top WA\left(W^\top \left(WW^\top + \lambda I\right)^{-1} W - I\right) + \lambda W \\ &= (P_{U,V})^\top WA\left(\left(W^\top W + \lambda I\right)^{-1} W^\top W - I\right) + \lambda W \\ &= -\lambda (P_{U,V})^\top WA\left(W^\top W + \lambda I\right)^{-1} + \lambda W \\ &= -\lambda \left(WW^\top + \lambda I\right)^{-1} WA^\top W^\top WA\left(W^\top W + \lambda I\right)^{-1} + \lambda W \\ &= -\lambda W\left(W^\top W + \lambda I\right)^{-1} AW^\top WA\left(W^\top W + \lambda I\right)^{-1} + \lambda W \\ &= \lambda W\left(I - \left(W^\top W + \lambda I\right)^{-1} AW^\top WA\left(W^\top W + \lambda I\right)^{-1}\right).\end{aligned} \tag{96}$$

In the above, we used that A is symmetric, i.e. $A^\top = A$ in the fifth line. Moreover, we used the property that $W^\top \left(WW^\top + \lambda I\right)^{-1} = \left(W^\top W + \lambda I\right)^{-1} W^\top$ in the second line and that $\left(WW^\top + \lambda I\right)^{-1} W = W\left(W^\top W + \lambda I\right)^{-1}$ in the fifth line.

Components of the vector field in the eigenbasis of A. Using the decomposition in Equation (24) of $WAW^\top$ and $WW^\top$ in such basis and recalling the expression of the vector field $\mathcal{F}(W)$ above, we directly get the desired expression for its components $\mathbf{g}$ in the basis $(\zeta_i)_{i\in I}$ in terms of $\mathbf{w}$ and the diagonal operator S which is given by the vector field F^1. Equivalence between the ODE in $\mathcal{W}$ and the ODE of its coordinates in $\mathbb{W}$ follows directly. □

E Proof of convergence to critical points

E.1 Proof of convergence to critical points for general gradient-like dynamical systems

For convenience, let us recall here the statement of the general convergence result before proceeding with the proof.

Proposition E.1. *Let F be a smooth vector field defined on a separable Hilbert space $\mathbb{W}$ and consider the following dynamical system:*

$$\dot{\mathbf{w}}_t = -F(\mathbf{w}_t), \qquad \mathbf{w}_0 \in \mathbb{W}. \tag{97}$$

Assume there exists a real analytic function L defined over $\mathbb{W}$ such that the following holds:

- ***(A) Dissipation.*** *L is a Lyapunov function for the system, i.e. $\dot{L}(\mathbf{w}_t) \leq 0$ for all $t \geq 0$.*
- ***(B) Sufficient decrease.*** *There exists $c_{max} > 0$ such that $\|F(\mathbf{w}_t)\|^2 \leq -c_{max}\dot{L}(\mathbf{w}_t)$ for $t \geq 0$.*
- ***(C) Relative bound.*** *There exists a non-increasing real-valued function $\Omega : [0, +\infty) \to (0, +\infty)$ such that $\Omega(\|\mathbf{w}_t\|)\|\nabla L(\mathbf{w}_t)\|^2 \leq \|F(\mathbf{w}_t)\|^2$.*
- ***(D) Coercivity.*** *L is coercive, i.e. $L(\mathbf{w}) \to +\infty$ when $\|\mathbf{w}\| \to +\infty$.*
- ***(E) Lower bound.*** *L admits a finite lower bound.*
- ***(F) Radial growth.*** *The vector field F, gradient ∇L and Hessian $\nabla^2 L$ all have radial growth, i.e. there exists a non-decreasing real-valued function $\omega : [0, +\infty) \to [0, +\infty)$ such that $\max\left(\|F(\mathbf{w})\|, \|\nabla L(\mathbf{w})\|, \left\|\nabla^2 L(\mathbf{w})\right\|\right) \leq \omega(\|\mathbf{w}\|)$ for all $\mathbf{w} \in \mathbb{W}$.*
- ***(G) Fredholm property.*** *For any $\mathbf{w} \in \mathbb{W}$, the hessian $\nabla^2 L(\mathbf{w})$ can be decomposed as the sum of a boundedly invertible operator R and a bounded compact operator K.*
- ***(H) Palais-Smale condition.*** *Any bounded sequence $(\mathbf{w}_n)_{n\geq 1}$ of $\mathbb{W}$ such that $\|\nabla L(\mathbf{w}_n)\| \to 0$ admits a convergent subsequence.*

Then the dynamical system in Equation (97) is defined at all times and converges to a critical point of the vector field F.

Proof of Proposition E.1. Let $\mathbf{w}_t$ be a solution to the ODE and L be a Lyapunov function, which exists by Assumption **(A)**.

Local existence. Since F is a smooth vector field, it is therefore locally Lipschitz. Hence, by Cauchy-Lipschitz theorem [Teschl, 2000, Theorem 1.9], the ODE Equation (97) admits a unique maximal solution starting from a point $\mathbf{w}_0$, continuous in time and defined on a maximal interval $[0, T)$ with $T > 0$ possibly equal to $+\infty$. We will later show that $T = +\infty$.

Bounded trajectories and velocities. We know by Assumption **(A)** that L is a Lyapunov function for the ODE Equation (97). Therefore, $\mathbf{w}_t$ remains in the level set $\{\mathbf{w} \in \mathbb{W} | L(\mathbf{w}) \leq L(\mathbf{w}_0)\}$. Moreover, since L is coercive by Assumption **(D)**, such level set must be bounded so that $\mathbf{w}_t$ is also bounded, i.e. there exists $b > 0$ so that $\|\mathbf{w}_t\| \leq b$ for all $0 \leq t < T$. To show that $\dot{\mathbf{w}}_t$ is bounded, recall that $F(\mathbf{w})$ is bounded over the centered ball of radius b by Assumption **(F)**. In particular, $t \mapsto \|F(\mathbf{w}_t)\|$ is also bounded, since we have just established that $\mathbf{w}_t$ belongs to such a ball for all $0 \leq t < T$. Consequently, $t \mapsto \dot{\mathbf{w}}_t$ remains also bounded for all $0 \leq t < T$, since, by definition $\dot{\mathbf{w}}_t = -F(\mathbf{w}_t)$.

Global existence. Let us now show that $T = +\infty$. By contradiction, assume that T is finite. We have shown that $\dot{\mathbf{w}}_t$ is bounded for all $0 \leq t < T$. Hence, there exists a positive constant M such that $\|\dot{\mathbf{w}}_t\| \leq M$ for all $0 \leq t < T$. Hence, for any $0 \leq s < t < T$,

$$\|\mathbf{w}_t - \mathbf{w}_s\| \leq \int_s^t \|\dot{\mathbf{w}}_\tau\| \, \mathrm{d}\tau \leq M(t - s). \tag{98}$$

Thus $(\mathbf{w}_t)_{t<T}$ is Cauchy as $t \uparrow T$. Since $\mathbb{W}$ is complete, there exists $\mathbf{w}_T \in \mathbb{W}$ such that$\mathbf{w}_t \to \mathbf{w}_T$ as $t \uparrow T$. Applying the local existence theorem again with initial condition $\mathbf{w}_T$ extends the solution beyond time T, contradicting the maximality of T. Therefore, $T = +\infty$.

Convergence of the energy gradient $\nabla L(\mathbf{w}_t)$ to 0. First, by Assumption (C), we know that $\Omega(\|\mathbf{w}\|_t)\,\|\nabla L(\mathbf{w}_t)\|^2 \leq \|F(\mathbf{w}_t)\|^2$. Since $\|\mathbf{w}_t\|$ is bounded by $b > 0$ and Ω is non-increasing, it follows that $0 < \Omega(b) \leq \Omega(\|\mathbf{w}\|_t)$. Denoting by $c_{min} = \Omega(b)$, we have shown that:

$$c_{min}\,\|\nabla L(\mathbf{w}_t)\|^2 \leq \|F(\mathbf{w}_t)\|^2\,. \tag{99}$$

Moreover, further using Assumption (B), it follows that $\|\nabla L(\mathbf{w}_t)\|^2 \leq -\frac{c_{max}}{c_{min}}\dot{L}(\mathbf{w}_t)$. Therefore, by integrating such inequality between 0 and any positive time T, it follows that:

$$\int_0^T \|\nabla L(\mathbf{w}_t)\|^2\,\mathrm{d}t \leq \frac{c_{max}}{c_{min}}\left(L(\mathbf{w}_0) - L(\mathbf{w}_T)\right). \tag{100}$$

Now recall that, by Assumptions (A) and (E), $t \mapsto L(\mathbf{w}_t)$ is non-increasing with L being lower-bounded. Therefore, $L(\mathbf{w}_t)$ must converge to a finite limit L_∞ satisfying $L(\mathbf{w}_t) \geq L_\infty$ for all $t \geq 0$. Using that $L(\mathbf{w}_T) \geq L_\infty$ in Equation (100) implies that $\int_0^{+\infty} \|\nabla L(\mathbf{w}_t)\|^2\,\mathrm{d}t$ is finite. Such an integrability condition alone is not enough to deduce that $\|\nabla L(\mathbf{w}_t)\|^2$ converges to 0. However, provided that $\|\nabla L(\mathbf{w}_t)\|^2$ is uniformly continuous, we can deduce convergence to 0 of $\|\nabla L(\mathbf{w}_t)\|^2$ by application of Barbalat's theorem [Barbălat, 1959]. We obtain uniform continuity by Lemma E.1 which applies here, as $\mathbf{w}_t$ and $\dot{\mathbf{w}}_t$ are bounded and $\mathbf{w} \mapsto \nabla L(\mathbf{w})$ and $\mathbf{w} \mapsto \nabla^2 L(\mathbf{w})$ are radially bounded by Assumption (F).

Existence of cluster points. We wish to establish the existence of a cluster point for the dynamics, i.e., a point $\mathbf{w}^\star$ such that there exists a subsequence along the trajectory $\mathbf{w}_t$ converging towards it. Let $(t_n)_{n\geq 1}$ be any increasing sequence of positive real numbers with $t_n \to +\infty$. We know that $(\mathbf{w}_{t_n})_{n\geq 1}$ is bounded since we established that $t \mapsto \mathbf{w}_t$ is bounded. By Assumption (H), we deduce that $(\mathbf{w}_{t_n})$ admits a convergent subsequence, thus establishing the existence of a cluster point as the limit point of such a subsequence. If $\mathbf{w}^\star$ is a cluster point, then it must be a critical point of ∇L. Indeed, we have already established that $\nabla L(\mathbf{w}_t) \to 0$. Hence, by continuity of ∇L, we directly get that $\nabla L(\mathbf{w}^\star) = 0$.

A Lojasiewicz-Simon's inequality. For any $\mathbf{w} \in \mathbb{W}$, we know, by Assumption (G), that the Hessian $\nabla^2 L(\mathbf{w})$ can be decomposed as the sum of a boundedly invertible operator R and a bounded compact operator K. The operator R is necessarily a Fredholm operator of index 0, i.e. $ker(R) = \{0\}$ and $ker(R^*) = \{0\}$ and R has a closed range. Moreover, by stability of Fredholm operators to compact perturbations, e.g. [Conway, 2019, Theorem 3.11], it follows that $\nabla^2 L(\mathbf{w}) = R + K$ is also Fredholm with index 0. Moreover, by assumption, L is analytic. Therefore, by direct application of [Feehan and Maridakis, 2020, Theorem 1], it follows that L satisfies a desired Lojasiewicz–Simon gradient inequality near its critical points. In other words, if $\mathbf{w}^\star$ is a critical point of ∇L, then there exist finite constants $Z > 0$, $0 < \sigma < 1$ and $\frac{1}{2} \leq \theta < 1$ so that for any $\mathbf{w}$ satisfying $\|\mathbf{w} - \mathbf{w}^\star\| \leq \sigma$, it must hold that:

$$\|\nabla L(\mathbf{w})\| \geq Z\,|L(\mathbf{w}) - L(\mathbf{w}^\star)|^\theta\,. \tag{101}$$

Trapping near a cluster point. Let $\mathbf{w}^\star$ be an accumulation point of $\mathbf{w}_t$, which was shown to exist. We have also shown that $L(\mathbf{w}_t)$ converges to a finite value L_∞ with $L(\mathbf{w}_t) \geq L_\infty$. Here we are only interested in the case where $L(\mathbf{w}_t) > L_\infty$ at all times. The other will be treated separately. To show that the trajectory remains close to the same cluster points, we will use a trapping argument near $\mathbf{w}^\star$ following the scheme from [Bolte et al., 2014] for Lyapunov functions satisfying a Lojasiewicz-Simon inequality. Fix $\epsilon > 0$. Since $\mathbf{w}^\star$ is an accumulation point of $\mathbf{w}_t$, and that $L(\mathbf{w}_t) \to L_\infty > -\infty$, there exists $T > 0$ so that $\|\mathbf{w}_T - \mathbf{w}^\star\| \leq \epsilon$ and $0 \leq L(\mathbf{w}_t) - L_\infty \leq \epsilon$ for all $t \geq T$. By choosing $\epsilon < \sigma$, we can guarantee that Equation (101) holds. Let us show that $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \sigma$ for any $t \geq T$. Consider T_{max}, the largest time greater than T for which $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \sigma$ for all $t \in (T, T_{max})$. We need to show that $T_{max} = \infty$. It is clear by continuity of $\mathbf{w}_t$ that $T_{max} > T$ so that the interval

(T, T_{max}) is non-empty. Recalling that $\dot{\mathbf{w}}_t = -F(\mathbf{w}_t)$, the following holds, for all $t \in (T, T_{max})$:

$$\begin{aligned}\|\dot{\mathbf{w}}_t\| = \|F(\mathbf{w}_t)\| &\leq \sqrt{c_{min}^{-1}} \frac{\|F(\mathbf{w}_t)\|^2}{\|\nabla L(\mathbf{w}_t)\|} \\ &\leq -\sqrt{c_{min}^{-1}} c_{max} \frac{\dot{L}(\mathbf{w}_t)}{\|\nabla L(\mathbf{w}_t)\|} \\ &\leq -Z^{-1}\sqrt{c_{min}^{-1}} c_{max} \frac{\dot{L}(\mathbf{w}_t)}{|L(\mathbf{w}_t) - L(\mathbf{w}^\star)|^\theta}, \\ &= -Z^{-1}\sqrt{c_{min}^{-1}} c_{max} \frac{\dot{L}(\mathbf{w}_t)}{(L(\mathbf{w}_t) - L_\infty)^\theta},\end{aligned} \tag{102}$$

where we used Equation (99) for the first line, then Assumption (B) for the second line. The third inequality, uses Equation (101), whereas the last one uses that $L(\mathbf{w}_t) \geq L_\infty = L(\mathbf{w}^\star)$, where the equality $L_\infty = L(\mathbf{w}^\star)$ follows by continuity of L and uniqueness of the limit L_∞. As we are in the case where $L(\mathbf{w}_t) - L_\infty > 0$, we must have $\|\nabla L(\mathbf{w}_t)\| > 0$ thanks to Equation (101), so that the above upper-bounds are always finite. Integrating the above inequality over an interval (t, s) with $T \leq t \leq s < T_{max}$ yields:

$$\begin{aligned}\int_t^s \|\dot{\mathbf{w}}_\tau\| \, \mathrm{d}\tau &\leq \alpha \left((L(\mathbf{w}_t) - L_\infty)^{1-\theta} - (L(\mathbf{w}_s) - L_\infty)^{1-\theta} \right) \\ &\leq \alpha \, (L(\mathbf{w}_t) - L_\infty)^{1-\theta}\end{aligned} \tag{103}$$

where $\alpha = \frac{1}{1-\theta} Z^{-1} \sqrt{c_{min}^{-1}} c_{max} < \infty$. Consequently, the above inequality implies, in particular, that:

$$\begin{aligned}\|\mathbf{w}_t - \mathbf{w}^\star\| &\leq \|\mathbf{w}_T - \mathbf{w}^\star\| + \int_T^t \|\dot{\mathbf{w}}_\tau\| \, \mathrm{d}\tau \\ &\leq \|\mathbf{w}_T - \mathbf{w}^\star\| + \alpha \, (L(\mathbf{w}_T) - L_\infty)^{1-\theta} \leq \epsilon + \alpha \epsilon^{1-\theta}.\end{aligned} \tag{104}$$

We can choose ϵ small enough so that $\epsilon + \alpha\epsilon^{1-\theta} \leq \frac{\sigma}{2}$, thus yielding $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \frac{\sigma}{2}$ for all $T \leq t < T_{max}$. If, by contradiction, T_{max} were finite, then, by continuity of $\mathbf{w}_t$ we would also have $\|\mathbf{w}_{T_{max}} - \mathbf{w}^\star\| \leq \frac{\sigma}{2}$ contradicting the definition of T_{max}. Therefore $T_{max} = +\infty$ and $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \frac{\sigma}{2}$ for all $t \geq T$.

Convergence to an accumulation point. To prove convergence towards $\mathbf{w}^\star$, we consider two separate cases depending on whether $L(\mathbf{w}_t)$ reaches its lower bound L_∞ or not.

- **Case 1**: There exists $t_0 \geq 0$ so that $L(\mathbf{w}_{t_0}) = L_\infty$.

Let $t \geq t_0$. Since $L(\mathbf{w}_t)$ is non-increasing by Assumption (A), it holds that $L_\infty = L(\mathbf{w}_{t_0}) \geq L(\mathbf{w}_t)$. On the other hand, we have $L(\mathbf{w}_t) \geq L_\infty$ so that $L(\mathbf{w}_t) = L_\infty$. As this equality holds for any $t \geq t_0$, it must be that $\dot{L}(\mathbf{w}_t) = 0$ for all $t \geq t_0$. Along with Assumption (B), we deduce that $\|F(\mathbf{w}_t)\|^2 \leq -c_{max}\dot{L}(\mathbf{w}_t) = 0$ for all $t \geq t_0$. Consequently, $\mathbf{w}_t$ remains constant, but since we know that $\mathbf{w}^\star$ is an accumulation point of the trajectory, it must hold that $\mathbf{w}_t = \mathbf{w}^\star$ for all $t \geq t_0$, which, a fortiori, implies convergence of $\mathbf{w}_t$ towards $\mathbf{w}^\star$.

- **Case 2**: $L(\mathbf{w}_t) > L_\infty$ for all $t \geq 0$.

We have already established that $\|\mathbf{w}_t - \mathbf{w}^\star\| \leq \sigma$ for all $t \geq T$, so that Equation (101) always holds and we can use Equation (103). Let $t \geq T$ and consider a sequence $\mathbf{w}_{s_n}$ converging towards $\mathbf{w}^\star$ with s_n increasing and greater than t. Such sequence always exists as $\mathbf{w}^\star$ is an accumulation point of $\mathbf{w}_t$. Recalling that $\mathbf{w}_{s_n} - \mathbf{w}_t = \int_t^{s_n} \dot{\mathbf{w}}_{t'} \, \mathrm{d}t'$ and using Equation (103), it follows that for any $n \geq 1$:

$$\|\mathbf{w}_{s_n} - \mathbf{w}_t\| \leq \int_t^{s_n} \|\dot{\mathbf{w}}_{t'}\| \, \mathrm{d}t' \leq \alpha \, (L(\mathbf{w}_t) - L_\infty)^{1-\theta}. \tag{105}$$

Taking the limit $n \to +\infty$ yields: $\|\mathbf{w}^\star - \mathbf{w}_t\| \leq \alpha \left(L(\mathbf{w}_t) - L_\infty\right)^{1-\theta}$, which holds for any $t \geq T$. Furthermore, since $L(\mathbf{w}_t) - L_\infty \to 0$, $\alpha > 0$ and $1-\theta > 0$, we deduce that $\mathbf{w}_t \to \mathbf{w}^\star$. By continuity of ∇L and F, we also have that $\nabla L(\mathbf{w}_t) \to \nabla L(\mathbf{w}^\star)$ and $F(\mathbf{w}_t) \to F(\mathbf{w}^\star)$. Recall also that $\mathbf{w}^\star$ is a critical point of ∇L, i.e. $\nabla L(\mathbf{w}^\star) = 0$. Finally, by Assumption **(B)**, we get

$$\|F(\mathbf{w}_t)\|^2 \leq -c_{max}\dot{L}(\mathbf{w}_t) = c_{max}\langle \nabla L(\mathbf{w}_t), F(\mathbf{w}_t)\rangle, \tag{106}$$

taking the limit to infinity yields $\|F(\mathbf{w}^\star)\|^2 \leq 0$. Thus $\mathbf{w}^\star$ is a critical point of F. □

Lemma E.1. *Consider a differentiable map $t \mapsto \mathbf{w}_t$ with values in a separable Hilbert space $\mathbb{W}$ with $\mathbf{w}_t$ and $\dot{\mathbf{w}}_t$ bounded. Let G be a smooth map on $\mathbb{W}$ so that G and ∇G have radial growth, i.e. there exists a non-decreasing function $\omega : [0, +\infty) \to [0, +\infty)$ so that $\|G(\mathbf{w})\|, \|\nabla G(\mathbf{w})\| \leq \omega(\|\mathbf{w}\|)$ for all $\mathbf{w} \in \mathbb{W}$. Then $t \mapsto \|G(\mathbf{w}_t)\|^2$ is uniformly continuous.*

Proof. We know that $\mathbf{w}_t$ is bounded, hence, there exists a positive real number b so that $\mathbf{w}_t$ is included in a centered ball of radius b for all t. As G and ∇G have radial growth, by assumption, they must be bounded on such a ball. Therefore, one can find a positive constant L so that the following holds:

$$\|G(\mathbf{w}_t) - G(\mathbf{w}_s)\| \leq \sup_{\substack{\mathbf{w}\in\mathbb{W} \\ \|\mathbf{w}\|\leq b}} \|\nabla G(\mathbf{w})\| \, \|\mathbf{w}_t - \mathbf{w}_s\| \leq L \sup_{t'\geq 0} \|\dot{\mathbf{w}}_{t'}\| \, |t-s|, \qquad \|G(\mathbf{w}_t)\| \leq L \tag{107}$$

Since, by assumption, $\sup_{t'\geq 0} \|\dot{\mathbf{w}}_{t'}\| < +\infty$, we deduce that $t \mapsto G(\mathbf{w}_t)$ is Lipschitz and bounded. We can therefore deduce that $t \mapsto \|G(\mathbf{w}_t)\|^2$ is Lipschitz thanks to the inequalities:

$$\begin{aligned} \left| \|G(\mathbf{w}_t)\|^2 - \|G(\mathbf{w}_s)\|^2 \right| &\leq \|G(\mathbf{w}_t) - G(\mathbf{w}_s)\| \left(\|G(\mathbf{w}_t)\| + \|G(\mathbf{w}_s)\|\right) \\ &\leq 2L^2 \sup_{t'\geq 0} \|\dot{\mathbf{w}}_{t'}\| \, |t-s|. \end{aligned} \tag{108}$$

A fortiori, $\|G(\mathbf{w}_t)\|^2$ is uniformly continuous. □

E.2 Proof of the auxiliary results of Proposition B.4 and Lemmas B.1 to B.4

Existence of a Lyapunov function

Proof of Proposition B.4 (Existence of a Lyapunov function). Differentiating L^κ w.r.t. $\mathbf{w}$ yields:

$$\nabla L^\kappa(\mathbf{w}) = 2\mathbf{w}\Phi(\mathbf{w}^\top\mathbf{w}), \qquad \Phi(M) = (M+\lambda I)^2 - S(MS)^\kappa. \tag{109}$$

Note the vector field F^κ can be expressed in terms of Φ as follows:

$$F^\kappa(\mathbf{w}) = \mathbf{w}B(M)\Phi(M)B(M), \qquad M = \mathbf{w}^\top\mathbf{w}, \qquad B(M) = (M+\lambda I)^{-1}. \tag{110}$$

Introducing $M_t = \mathbf{w}_t^\top\mathbf{w}_t$ for conciseness and computing the time derivatives of $L^\kappa(\mathbf{w}_t)$ yields:

$$\begin{aligned} \dot{L}^\kappa(\mathbf{w}_t) &= Tr\left(\nabla L^\kappa(\mathbf{w}_t)^\top \dot{\mathbf{w}}_t\right) \\ &= -2\lambda Tr\left(\Phi(M_t)M_tB(M_t)\Phi(M_t)B(M_t)\right) \\ &= -2\lambda Tr\left(\Phi(M_t)B(M_t)M_t\Phi(M_t)B(M_t)\right) \\ &= -2\lambda Tr\left(B(M_t)^{\frac{1}{2}}M_t\Phi(M_t)B(M_t)\Phi(M_t)B(M_t)^{\frac{1}{2}}\right) \\ &= -2\lambda Tr\left(M_tB(M_t)^{\frac{1}{2}}\Phi(M_t)B(M_t)\Phi(M_t)B(M_t)^{\frac{1}{2}}\right) \\ &= -2\lambda Tr\left(M_tQ_t^2\right) = -2\lambda\|\mathbf{w}_tQ_t\|^2, \end{aligned} \tag{111}$$

where we defined $Q_t = B(M_t)^{\frac{1}{2}}\Phi(M_t)B(M_t)^{\frac{1}{2}}$ and used that M_t always commutes with $B(M_t)$.

Sufficient decrease. First, recall that by definition of F^κ, we have that:

$$\|F^\kappa(\mathbf{w})\|^2 = Tr\left(B\Phi BMB\Phi B\right), \tag{112}$$

where, we used that $M = \mathbf{w}^\top\mathbf{w}$, $B = B(M) = (M+\lambda I)^{-1}$ and $\Phi = \Phi(M)$ for conciseness. Recalling that M and B are PSD and commute, by construction, and that $B \leq \frac{1}{\lambda}I$, it follows:

$$\begin{aligned}
\|F^\kappa(\mathbf{w})\|^2 &= Tr\left(B\Phi B^{\frac{1}{2}}M^{\frac{1}{2}}BM^{\frac{1}{2}}B^{\frac{1}{2}}\Phi B\right)\\
&\leq \lambda^{-1}Tr\left(B\Phi B^{\frac{1}{2}}M^{\frac{1}{2}}M^{\frac{1}{2}}B^{\frac{1}{2}}\Phi B\right)\\
&= \lambda^{-1}Tr\left(M^{\frac{1}{2}}B^{\frac{1}{2}}\Phi B^{\frac{1}{2}}BB^{\frac{1}{2}}\Phi B^{\frac{1}{2}}M^{\frac{1}{2}}\right)\\
&\leq \lambda^{-2}Tr\left(M^{\frac{1}{2}}B^{\frac{1}{2}}\Phi B^{\frac{1}{2}}B^{\frac{1}{2}}\Phi B^{\frac{1}{2}}M^{\frac{1}{2}}\right)\\
&= \lambda^{-2}Tr\left(MB^{\frac{1}{2}}\Phi B\Phi B^{\frac{1}{2}}\right) = -\frac{1}{2\lambda^3}\dot{L}^\kappa(\mathbf{w}_t).
\end{aligned} \tag{113}$$

Relative error. Similarly, using again that B and M commute:

$$\begin{aligned}
\|F^\kappa(\mathbf{w})\|^2 &= Tr\left(B\Phi M^{\frac{1}{2}}B^2M^{\frac{1}{2}}\Phi B\right)\\
&\geq \sigma_{min}(B^2)Tr\left(B\Phi M^{\frac{1}{2}}M^{\frac{1}{2}}\Phi B\right)\\
&\geq \sigma_{min}(B^2)Tr\left(M^{\frac{1}{2}}\Phi B^2\Phi M^{\frac{1}{2}}\right)\\
&\geq \sigma_{min}(B^2)^2Tr\left(M^{\frac{1}{2}}\Phi\Phi M^{\frac{1}{2}}\right)\\
&= \sigma_{min}(B^2)^2Tr\left(\Phi M\Phi\right)\\
&= \frac{1}{4}\sigma_{min}(B^2)^2\left\|\nabla L^\kappa(\mathbf{w})\right\|^2,
\end{aligned} \tag{114}$$

where we used that $B^2 \geq \sigma_{min}(B^2)I$ the smallest singular value of B^2. Moreover, simple estimates show that $\sigma_{min}(B^2) \geq \frac{1}{(\|\mathbf{w}\|^2+\lambda)^2}$. Hence, we have established the relative error bound:

$$\|F^\kappa(\mathbf{w})\|^2 \geq \frac{1}{4(\|\mathbf{w}\|^2+\lambda)^4}\left\|\nabla L^\kappa(\mathbf{w})\right\|^2.$$

□

Growth of the vector field F^κ.

Proof of Lemma B.1. The bound holds by direct application of operator norm properties. Specifically, for $\mathbf{w} \in \mathbb{W}$, and denoting by $R = \left(\mathbf{w}^\top\mathbf{w} + \lambda I\right)^{-1}$, we have:

$$\begin{aligned}
\|F^\kappa(\mathbf{w})\| &\leq \|\mathbf{w}\|\left(1 + \left\|RS(\mathbf{w}^\top\mathbf{w}S)^\kappa R\right\|_{op}\right)\\
&\leq \|\mathbf{w}\|\left(1 + \underbrace{\|R\|_{op}^2}_{\leq\lambda^{-2}}\|S\|_{op}^{1+\kappa}\left\|\mathbf{w}^\top\mathbf{w}\right\|_{op}^{\kappa}\right)\\
&\leq \|\mathbf{w}\|\left(1 + \lambda^{-2}\|S\|_{op}^{1+\kappa}\|\mathbf{w}\|^{2\kappa}\right).
\end{aligned} \tag{115}$$

This establishes that F^κ is radially bounded by the function $\omega(\|\mathbf{w}\|) = \|\mathbf{w}\|\left(1 + \lambda^{-2}\|S\|_{op}^{1+\kappa}\|\mathbf{w}\|^{2\kappa}\right)$. □

Growth, coercivity and lower boundedness of the energy L^κ.

Proof of Lemma B.2 (Finite lower bound on the energy). Let us first show that L^κ is lower-bounded. Let $\mathbf{w} \in \mathbb{W}$. Then $\mathbf{w}\mathbf{w}^\top$ is a $k \times k$ PSD matrix. Denote by $\sigma_1, \dots, \sigma_k$ its singular values. Then

$$Tr\left((\mathbf{w}\mathbf{w}^\top + \lambda I)^3\right) = \sum_{i=1}^{k} (\sigma_i + \lambda)^3 \;\geq\; \sum_{i=1}^{k} \sigma_i^3 + k\lambda^3 \;\geq\; \frac{1}{k^2}\|\mathbf{w}\|_{\mathbb{W}}^6 + k\lambda^3, \tag{116}$$

using $\sum \sigma_i^3 \geq k^{-2}(\sum \sigma_i)^3 = k^{-2}\|\mathbf{w}\|_{\mathbb{W}}^6$ by Jensen's inequality. For the second term of L^κ, we use the following upper-bound:

$$\left|Tr\left(\left(\mathbf{w}S\mathbf{w}^\top\right)^{\kappa+1}\right)\right| \leq \|S\|_{op}^{\kappa+1} \|\mathbf{w}\|_{\mathbb{W}}^{2(\kappa+1)} \tag{117}$$

Hence, with $r = \|\mathbf{w}\|_{\mathbb{W}}$,

$$L^\kappa(\mathbf{w}) \geq \frac{1}{3k^2} r^6 - \frac{1}{\kappa+1} \|S\|_{op}^{\kappa+1} r^{2(\kappa+1)} + \frac{k\lambda^3}{3}. \tag{118}$$

The right-hand side is bounded below for $r \geq 0$, as long as $\kappa < 2$ (positive leading coefficient), so L^κ is bounded below. From the above inequality, we also get that L^κ is coercive.

To show that ∇L^κ and $\nabla^2 L^\kappa$ are radially bounded, recall that L^κ is a smooth polynomial, so that ∇L^κ and $\nabla^2 L^\kappa$ are also smooth polynomials and are thus, a fortiori, radially bounded. □

Hessian decomposition.

Proof of Lemma B.3 (Hessian decomposition). First, by direct differentiation, we have that:

$$\nabla L^\kappa(\mathbf{w}) = 2\mathbf{w}\left(\left(\mathbf{w}^\top\mathbf{w} + \lambda I\right)^2 - S\left(\mathbf{w}^\top\mathbf{w}S\right)^\kappa\right) = 2\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^2\mathbf{w} - 2\mathbf{w}S\left(\mathbf{w}^\top\mathbf{w}S\right)^\kappa. \tag{119}$$

Hence, differentiating the above along a direction v yields:

$$\begin{aligned}\nabla^2 L^\kappa(\mathbf{w})(\mathbf{w}') =& \underbrace{2\left(\left(\mathbf{w}'\mathbf{w}^\top + \mathbf{w}\mathbf{w}'^\top\right)\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right) + \left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)\left(\mathbf{w}'\mathbf{w}^\top + \mathbf{w}\mathbf{w}'^\top\right)\right)}_{Q(\mathbf{w}')}\mathbf{w} \\ &+ 2\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^2\mathbf{w}' - \underbrace{\left(2\mathbf{w}'S\left(\mathbf{w}^\top\mathbf{w}S\right)^\kappa + 2\delta_{\kappa=1}\mathbf{w}S\left(\mathbf{w}^\top\mathbf{w}' + \mathbf{w}'^\top\mathbf{w}\right)S\right)}_{K_1(\mathbf{w}')} \\ =& 2\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^2\mathbf{w}' + Q(\mathbf{w}')\mathbf{w} - K_1(\mathbf{w}')\end{aligned} \tag{120}$$

By assumption, we know that S is a bounded compact operator, hence, $\mathbf{w}' \mapsto K_1(\mathbf{w}')$ is also bounded and compact as a sum of compositions of bounded and compact operators. Moreover $Q(\mathbf{w}')$ is a $k \times k$ matrix, so that $K_2 : \mathbf{w}' \mapsto Q(\mathbf{w}')\mathbf{w}$ lives in the range of the k components of $\mathbf{w}$. Hence, $\mathbf{w}' \mapsto Q(\mathbf{w}')\mathbf{w}$ has a finite range and is thus, a fortiori, compact. Finally, note that $2\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^2$ is an invertible $k \times k$ matrix. Hence, the map $R : \mathbf{w}' \mapsto 2\left(\mathbf{w}\mathbf{w}^\top + \lambda I\right)^2\mathbf{w}'$ is invertible with a bounded inverse. We have shown, so far, that $\nabla^2 L^\kappa(\mathbf{w})$ is decomposed as the sum of a boundedly invertible operator R and a bounded compact operator $K_1 + K_2$.

□

Palais-Smale condition.

Proof of Lemma B.4 (Palais-Smale condition). Consider a bounded sequence $(\mathbf{w}_n)_{n\geq 1}$ in $\mathbb{W}$ satisfying $\|\nabla L^\kappa(\mathbf{w}_n)\| \to 0$. We will show that it must admit a cluster point. By direct calculations, we know that $\nabla L^\kappa(\mathbf{w}_n)$ admits the following expression:

$$r_n = \nabla L^\kappa(\mathbf{w}_n) = 2\mathbf{w}_n\left((M_n + \lambda I)^2 - S(M_n S)^\kappa\right), \qquad M_n = \mathbf{w}_n^\top\mathbf{w}_n. \tag{121}$$

We will first prove that M_n admits a convergent subsequence, then deduce that $\mathbf{w}_n$ also does.

Existence of a convergent subsequence of M_n. Left-multiplying the above equation by $\mathbf{w}_n^\top$ we get:

$$\mathbf{w}_n^\top r_n = 2M_n(M_n + \lambda I)^2 - 2(M_n S)^{\kappa+1} \tag{122}$$

Define $X_n = 2M_n(M_n + \lambda I)^2$ and $B_n = 2(M_n S)^\kappa M_n$. Then since $(\mathbf{w}_n)_{n\geq 1}$ is bounded, it follows that $(X_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ form bounded sequences of bounded operators on $\ell^2(I)$. Moreover, X_n is self-adjoint by construction and $X_n - B_n S = \mathbf{w}_n^\top r_n \to 0$, since, by assumption $r_n \to 0$ and $\mathbf{w}_n$ is bounded. Thus, recalling that S is a bounded compact operator, we can apply Lemma E.2 to directly deduce that X_n admits a convergent subsequence in operator norm.

Consider the scalar function $\phi : t \mapsto 2t(t+\lambda)^2$ which is strictly increasing on $[0, +\infty)$ and thus admits a continuous inverse ϕ^{-1} defined on $[0, +\infty)$. Consequently, the map $\Phi : M \mapsto 2M(M + \lambda I)^2$ defined over bounded self-adjoint semi-definite operators on $\ell^2(I)$ is a homeomorphism, by continuous functional calculus for bounded PSD self-adjoint operators, and thus admits a continuous inverse Φ^{-1}. Let now $(X_{n_l})_{l\geq 1}$ be a convergent subsequence of $(X_n)_{n\geq 1}$, then we have that $X_{n_l} = \Phi(M_{n_l})$ for all $l \geq 1$. Moreover, $M_{n_l} = \Phi^{-1}(X_{n_l})$ converges by continuity of Φ^{-1}. Therefore, $(M_n)_{n\geq 1}$ admits a convergent subsequence in operator norm.

Existence of a convergent subsequence for $(\mathbf{w}_n)_{n\geq 1}$. Multiplying Equation (121) on the right by $(M_n + \lambda I)^{-2}$ and re-arranging the equality yields:

$$\mathbf{w}_n = \frac{1}{2}\left(r_n(M_n + \lambda I)^{-2}\right) + \mathbf{w}_n S(M_n S)^\kappa (M_n + \lambda I)^{-2}. \tag{123}$$

Since $\mathbf{w}_n$ is bounded, so is M_n, hence, the first terms $r_n(M_n + \lambda I)^{-2}$ converges to 0 as $r_n \to 0$ by assumption. For the second term, we know that M_n admits a convergent subsequence $(M_{n_l})_{l\geq 1}$. Moreover, by compactness of S, we know that $(\mathbf{w}_{n_l} S)_{l\geq 1}$ admits a convergent subsequence. Thus, we have shown that one can extract a subsequence of $\mathbf{w}_n$ so that the right hand side of Equation (123) converges by continuity. Denote this limit by $\mathbf{w}^\star$. This precisely means that such subsequence of $(\mathbf{w}_n)_{n\geq 1}$ converges itself to $\mathbf{w}^\star$ thanks to the identity in Equation (123). □

Lemma E.2. *Let S be a bounded compact operator on $\ell^2(I)$, and $(B_n)_{n\geq 1}$ a bounded sequence of bounded operators on $\ell^2(I)$. Let $(X_n)_{n\geq 1}$ be a family of self-adjoint bounded operators satisfying:*

$$\|X_n - B_n S\| \to 0.$$

Then (X_n) admits a convergent subsequence in operator norm.

Proof. Let $(P_m)_{m\geq 1}$ be a family of finite-rank orthogonal projections in $\ell^2(I)$ such that P_m converges to the identity as m increases. Since S is compact, necessarily, $\|S(I - P_m)\| \to 0$. Moreover, we have that $X_n - B_n S \to 0$ and (B_n) is bounded. Hence, (X_n) is bounded and

$$\limsup_{n\to\infty} \|X_n(I - P_m)\| \leq \sup_n \|B_n\| \, \|S(I - P_m)\| \to 0 \qquad (m \to \infty). \tag{124}$$

Moreover, by assumption X_n is self-adjoint, i.e., $X_n = X_n^*$, so that

$$\|(I - P_m)X_n\| = \|X_n(I - P_m)\|,$$

and

$$\begin{aligned}\|X_n - P_m X_n P_m\| &= \|X_n(I - P_m) + (I - P_m)X_n P_m\| \\ &\leq \|X_n(I - P_m)\| + \|(I - P_m)X_n\| \, \|P_m\| \\ &\leq 2\,\|X_n(I - P_m)\|\,.\end{aligned}$$

Therefore, combining the above upper-bound with Equation (124) yields:

$$\lim_{m\to\infty} \limsup_{n\to\infty} \|X_n - P_m X_n P_m\| = 0. \tag{125}$$

We now show that the sequence $(X_n)_{n\geq 1}$ is relatively compact in the operator norm, meaning that its closure is a compact set. To achieve this, we only need to show that $(X_n)_{n\geq 1}$ admits a finite ϵ-cover for any $\epsilon > 0$. Fix $\epsilon > 0$, then Equation (125) ensures there exists $m > 0$ so that $\limsup_{n\to\infty} \|X_n - P_m X_n P_m\| \leq \epsilon/4$. Consequently, there exists $N > 0$ so that for all $n \geq N$:

$$\|X_n - P_m X_n P_m\| \leq \epsilon/2 \tag{126}$$

Denote by $P_m \ell^2(I)$ the finite dimensional space of projections of $\ell^2(I)$ by the operator P_m. Then the restrictions Q_n of $P_m X_n P_m$ to $P_m \ell^2(I)$ are stable, i.e., Q_n are operators from $P_m \ell^2(I)$ to itself. Moreover, $(Q_n)_{n\geq 1}$ is a bounded family of bounded operators on the finite-dimensional space

$P_m\ell^2(I)$. Therefore, $(Q_n)_n$ is relatively compact and thus admits a finite $\epsilon/2$-cover $\tilde{Q}_1, \dots, \tilde{Q}_K$, for some $K > 0$ with $\tilde{Q}$ bounded operators on $P_m\ell^2(I)$ satisfying:

$$\min_{1\leq l\leq K} \left\| Q_n - \tilde{Q}_l \right\| \leq \frac{\epsilon}{2}. \tag{127}$$

By abuse of notation, we identify $\tilde{Q}_l$ with its unique extension on $\ell^2(I)$ satisfying $\tilde{Q}_l \mathbf{w} = \tilde{Q}_l P_m \mathbf{w}$ for any $\mathbf{w} \in \ell^2(I)$. Hence, combining Equations (126) and (127) ,we have shown that for all $n \geq N$:

$$\min_{1\leq l\leq K} \left\| X_n - \tilde{Q}_l \right\| \leq \|X_n - P_m X_n P_m\| + \min_{1\leq l\leq K} \left\| Q_n - \tilde{Q}_l \right\| \leq \epsilon. \tag{128}$$

Consequently, $\{\tilde{Q}_1, \dots, \tilde{Q}_K\} \cup \{X_1, \dots, X_N\}$ is a finite ϵ-cover of $(X_n)_{n\geq 1}$ w.r.t. the operator norm. This establishes that $(X_n)_{n\geq 1}$ is relatively compact and a fortiori that X_n admits a convergent subsequence. □

F Proof of the characterization of equilibria and their stability

F.1 Proof of Proposition B.6 on characterization of equilibria

Proof of Proposition B.6. Let $\mathbf{w}$ be a critical point of F^κ, i.e. $F^\kappa(\mathbf{w}) = 0$. Left-multiplying $F^\kappa(\mathbf{w})$ by $\mathbf{w}^\top$ and defining $M = \mathbf{w}^\top \mathbf{w}$, the following condition must hold for M:

$$M\left(I - (M + \lambda I)^{-1} G(M) (M + \lambda I)^{-1}\right) = 0, \qquad G(M) = S(MS)^\kappa. \tag{129}$$

Recalling that M necessarily commutes with $(M + \lambda I)^{-1}$ and multiplying both sides by $(M + \lambda I)$, we directly deduce that:

$$(M + \lambda I)M(M + \lambda I) = MG(M). \tag{130}$$

Since the l.h.s. of the above expression is always symmetric, so must be its r.h.s., i.e. $MG(M) = (MG(M))^\top$. We will use such symmetry to prove that $MG(M)$ must be diagonal.

$MG(M)$ is diagonal. Using that $G(M) = S(MS)^\kappa$, the symmetry condition on $MG(M)$ becomes $MS(MS)^\kappa = (MS(MS)^\kappa)^\top$. Using the associativity of the product, it is clear that $(MS)^\kappa MS = SM(SM)^\kappa$. Furthermore, by setting $T = (MS)^\kappa M$, and noting that $T = M(SM)^\kappa$ by associativity of the product, we have shown that $TS = ST$, i.e. T commutes with S. Moreover, we know, by assumption, S is compact and diagonal in the canonical orthonormal basis of $\ell^2(I)$, thus it admits the following block representation with respect to the orthogonal decomposition of $\ell^2(I)$ in Equation (26):

$$S \equiv \begin{pmatrix} S^+ & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S^- \end{pmatrix}, \tag{131}$$

where $\equiv$ denotes operator similarity, S^+ and S^- collect the positive and negative eigenvalues of S in decreasing order w.r.t. their absolute values. For any canonical basis element e_i, the following relation holds $STe_i = TSe_i = s_i Te_i$ for all $i \in I$. The previous equation implies Te_i is an eigenvector of S associated to eigenvalue s_i. By assumption, eigenspaces of S associated to non-zero eigenvalues have dimension 1. Therefore, Te_i belongs to the span of e_i. On the other hand, the null space of S is invariant under T. In other words, T admits the following block diagonal structure w.r.t. the orthogonal decomposition of $\ell^2(I)$ in Equation (26):

$$T \equiv \begin{pmatrix} T^+ & 0 & 0 \\ 0 & T^0 & 0 \\ 0 & 0 & T^- \end{pmatrix}, \tag{132}$$

where T^+ and T^- are diagonal, while T^0 is a block associated to 0 eigenvalues of S. Hence, we have shown that M satisfies an equation of the form:

$$(M + \lambda I)M(M + \lambda I) = ST \equiv \begin{pmatrix} S^+T^+ & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & S^-T^- \end{pmatrix}. \tag{133}$$

The above equation establishes that $(M + \lambda I)M(M + \lambda I)$ is diagonal in the canonical basis. It remains to show that M is itself diagonal.

M is positive semi-definite (PSD), and is compact, as an operator of rank at most k. Therefore, by the spectral theorem, it admits an eigenvalue decomposition of the form: $M = R^\top \Lambda R$, where R is orthogonal ($R^\top R = I$) and Λ is diagonal with non-negative components. Consequently, $(M+\lambda I)M(M+\lambda I) = \Gamma(M) = R^\top \Gamma(\Lambda) R$, where $\Gamma(\Lambda)$ is a diagonal matrix whose i-th diagonal element is given by $\gamma(\ell_i)$ where ℓ_i is the i-th diagonal element of Λ and $\gamma(\ell) = \ell(\ell+\lambda)^2$ for any ℓ in $[0, +\infty)$. Since γ is continuous and strictly increasing over $[0, +\infty)$, it must have an inverse γ^{-1}. Applying continuous functional calculus for bounded self-adjoint PSD operators, we can define the inverse Γ^{-1} of Γ which applies γ^{-1} to the eigenvalues of any PSD matrix D without affecting the eigenvectors. Consequently, we can write $M = \Gamma^{-1}(\Gamma(M)) = \Gamma^{-1}(ST)$. Since we established that ST is diagonal, $\Gamma^{-1}(ST)$ must also be a diagonal matrix. Therefore, M is diagonal and admits the following block representation w.r.t. the orthogonal decomposition of $\ell^2(I)$ from Equation (26):

$$M \equiv \begin{pmatrix} M^+ & 0 & 0 \\ 0 & M^0 & 0 \\ 0 & 0 & M^- \end{pmatrix}. \tag{134}$$

Characterizing fixed points. Since M is diagonal, we only need to compute its diagonal coefficients. From Equation (133), we directly see that $M^0 = 0$. For M^+ and M^-, we use Equation (129), which implies the following equations on its diagonal components $(m_i)_{i \in I}$:

$$m_i \left(1 - \frac{m_i^\kappa s_i^{\kappa+1}}{(m_i + \lambda)^2}\right) = 0, \qquad m_i \geq 0. \tag{135}$$

In particular, we can solve in closed form for $\kappa \in \{0, 1\}$.

- **Case $\kappa = 0$.** In this case, m_i has the following admissible solutions:

$$m_i \in \{0, (g(s_i, \lambda))^2\}, \qquad g(s, \lambda) = (\sqrt{s} - \lambda)_+^{\frac{1}{2}} \mathbb{1}_{s \geq 0}. \tag{136}$$

- **Case $\kappa = 1$.** In this case, m_i has the following admissible solutions:

$$m_i \in \{0, (f_1(s_i, \lambda))^2, (f_{-1}(s_i, \lambda))^2\}, \quad f_\epsilon(s, \lambda) = \frac{1}{2}\left(|s| + \epsilon\sqrt{s^2 - 4\lambda}\right) \mathbb{1}_{|s|^2 - 4\lambda \geq 0}. \tag{137}$$

Recalling that $M = \mathbf{w}^\top \mathbf{w}$, it follows that M is of rank at most k, as $\mathbf{w}$ is itself of rank at most k, and therefore there can be at most k non-zero diagonal elements in M. Let r be the number of such non-zero elements. We have that $0 \leq r \leq k$. If $r > 0$, let $\mathcal{K} = \{i(1), \ldots, i(r)\}$ be the set of indices corresponding to such non-zero elements and set $\mathcal{K} = \emptyset$ if $r = 0$. When $r = 0$, clearly $\mathbf{w} = 0$. Otherwise, when $r > 0$, the condition $M = \mathbf{w}^\top \mathbf{w}$ along with the constraints on M imply that $\mathbf{w}$ is of the form:

$$\mathbf{w} = Q \left(\sum_{p=1}^{r} E_{p,i(p)}\right) M^{\frac{1}{2}}, \text{ with } 1 \leq r \leq k, \tag{138}$$

where Q is an orthogonal matrix in $\mathbb{R}^{k \times k}$ and $i(1), \ldots, i(r)$ is an increasing sequence of positive integers. Conversely, we can verify that any element $\mathbf{w}$ of the form Equation (42) is a critical point of the vector field F^κ. □

F.2 Proof of the diagonalization of the Jacobian of the vector field F^κ at critical points

Proof of Proposition B.7. Let $\mathbf{w}$ be a critical point of F^κ, which is necessarily of the form:

$$\mathbf{w} = Q \underbrace{\left(\sum_{p=1}^{r} E_{p,i(p)}\right)}_{\mathbf{w}_0} M^{\frac{1}{2}}, \tag{139}$$

Let us first express the Jacobian of F^κ at $\mathbf{w}$. By direct differentiation, one gets that the vector field F^κ defined in Equation (28) admits the following Jacobian representation when acting on elements $\mathbf{w}'$ in $\mathbb{W}$:

$$\begin{aligned}\nabla F^\kappa(\mathbf{w})[\mathbf{w}'] &= \mathbf{w}'\left(I - RG^\kappa(M)R\right) + \mathbf{w}R\left(DRG^\kappa(M) + G^\kappa(M)RD - \nabla G^\kappa[M](D)\right)R,\\ \text{with}\quad M &= \mathbf{w}^\top\mathbf{w}, \qquad R = (M+\lambda I)^{-1}, \qquad D = \mathbf{w}'^\top\mathbf{w} + \mathbf{w}^\top\mathbf{w}',\\ \text{and } G^\kappa(M) &= S(MS)^\kappa, \quad \nabla G^0[M](M') = 0, \qquad \nabla G^1[M](M') = SM'S.\end{aligned} \tag{140}$$

We can further express $\nabla F^\kappa(\mathbf{w})[\mathbf{w}']$ as the sum of three terms:

$$\begin{aligned}\nabla F^\kappa(\mathbf{w})[\mathbf{w}'] = &\underbrace{\mathbf{w}'\left(I - RG^\kappa(M)R\right)}_{H_1}\\ &+ \underbrace{\mathbf{w}R\left(\mathbf{w}'^\top\mathbf{w}RG^\kappa(M) + G^\kappa(M)R\mathbf{w}'^\top\mathbf{w} - \nabla G^\kappa[M](\mathbf{w}'^\top\mathbf{w})\right)R}_{H_2}\\ &+ \underbrace{\mathbf{w}R\left(\mathbf{w}^\top\mathbf{w}'RG^\kappa(M) + G^\kappa(M)R\mathbf{w}^\top\mathbf{w}' - \nabla G^\kappa[M](\mathbf{w}^\top\mathbf{w}')\right)R}_{H_3}.\end{aligned} \tag{141}$$

We will establish a general diagonal decomposition of $\nabla F^\kappa(\mathbf{w})$. Let $\mathbf{w}$ be a critical point of F^κ as in the statement. By definition of the critical point $\mathbf{w}$ in Equation (45), $M = \mathbf{w}^\top\mathbf{w}$ is diagonal, so is R and $G^\kappa(M)$. Let us denote by m_i, r_i and g_i the corresponding i-th diagonal coefficient for $i \in I$ with r_i and g_i satisfying:

$$r_i = (m_i + \lambda)^{-1}, \qquad g_i = s_i^{\kappa+1}m_i^\kappa \tag{142}$$

We will show that the desired decomposition can be obtained using the orthonormal basis $(QE_{p,j})_{1\leq p\leq k, j\in I}$ of $\mathbb{W}$. To this end, we will need to compute the action of $\nabla F^\kappa(\mathbf{w})$ on an element $\tilde{E}_{p,j} = QE_{p,j}$ for a given $1 \leq p \leq k$ and $j \in I$. As a first step, we need to compute $\tilde{E}_{p,j}^\top\mathbf{w}$ as it appears in both H_2 and H_3. Using the decomposition of $\mathbf{w}$ from Equation (45) and properties of the canonical basis, it holds that:

$$E_{p,j}^\top\mathbf{w}_0 = \sum_{q=1}^{r} E_{p,j}^\top E_{q,i(q)} = E_{p,j}^\top E_{p,i(p)}\mathbb{1}_{1\leq p\leq r}, \quad \tilde{E}_{p,j}^\top\mathbf{w} = m_{i(p)}^{\frac{1}{2}}E_{p,j}^\top E_{p,i(p)}\mathbb{1}_{1\leq p\leq r}. \tag{143}$$

Furthermore, recall that right-multiplication of $E_{p,j}$ with a diagonal operator T yields $E_{p,j}T = t_jE_{p,j}$, where t_j is the j-th diagonal coefficient of T. Hence, we can express H_1, H_2 and H_3 as:

$$\begin{aligned}H_1 &= \left(1 - r_j^2g_j\right)\tilde{E}_{p,j}\\ H_2 &= m_j^{\frac{1}{2}}m_{i(p)}^{\frac{1}{2}}\beta_{i(p),j}Q\mathbf{w}_0E_{p,j}^\top E_{p,i(p)}\mathbb{1}_{1\leq p\leq r}\\ H_3 &= m_{i(p)}\beta_{i(p),j}Q\mathbf{w}_0E_{p,i(p)}^\top E_{p,j}\mathbb{1}_{1\leq p\leq r},\end{aligned} \tag{144}$$

where we introduced $\beta_{i,j}$ for conciseness:

$$\beta_{i,j} = r_ir_j^2g_j + r_jr_i^2g_i - r_jr_is_js_i\mathbb{1}_{\kappa=1}. \tag{145}$$

Moreover, defining $i^{-1}(i(p)) = p$ for $1 \leq p \leq r$, the following identity holds:

$$\mathbf{w}_0E_{p,j}^\top E_{p,i(p)} = E_{i^{-1}(j),i(p)}\mathbb{1}_{j\in I_{sup}}, \qquad \mathbf{w}_0E_{p,i(p)}^\top E_{p,j} = E_{p,j}. \tag{146}$$

Using the above in Equation (144) and writing $q = i^{-1}(j)$ and $l = i(p)$ allows us to express the Jacobian as follows:

$$\nabla F^\kappa(\mathbf{w})(\tilde{E}_{p,j}) = \left(1 - r_j^2g_j + m_l\beta_{l,j}\mathbb{1}_{p\leq r}\right)\tilde{E}_{p,j} + m_j^{\frac{1}{2}}m_l^{\frac{1}{2}}\beta_{l,j}\tilde{E}_{q,l}\mathbb{1}_{p\leq r, j\in I_{sup}}. \tag{147}$$

We distinguish two cases depending on whether $j \in I_{sup}$ and $p \leq r$ or not.

Case $j \notin I_{sup}$ or $p > r$. The second term in Equation (147) vanishes and we get that $\tilde{E}_{p,j}$ is an eigenvector of $\nabla F^\kappa(\mathbf{w})$ with eigenvalue $\mu_{p,j}$ given by:

$$\begin{aligned}\mu_{p,j} &= \left(1 - r_j^2 g_j + m_{i(p)}\beta_{i(p),j}\mathbb{1}_{1\leq p\leq r}\right) \\ &= \begin{cases} 0, & j \in I_{sup}, p > r \\ 1 - \lambda^{-2} s_j^{\kappa+1} m_j^\kappa, & j \notin I_{sup}, p > r \\ 1 - \lambda^{-1}\frac{s_j^{\kappa+1} m_j^\kappa}{m_{i(p)}+\lambda} + \lambda^{-1} m_{i(p)}\left(1 - \frac{s_{i(p)}s_j}{m_{i(p)}+\lambda}\mathbb{1}_{\kappa=1}\right), & j \notin I_{sup}, p \leq r \end{cases},\end{aligned} \tag{148}$$

where we used that $m_j = 0$ and $m_{i(p)}$ satisfies the equation $(m_{i(p)} + \lambda)^2 = s_{i(p)}^{\kappa+1} m_{i(p)}^\kappa$.

Case $j \in I_{sup}$ and $p \leq r$. If $j = i(p)$, then $j' = j$ and $p' = p$ so that $\tilde{E}_{p,j} = \tilde{E}_{p',j'}$. Hence, $\tilde{E}_{p,j}$ is again an eigenvector of $\nabla F^\kappa(\mathbf{w})$ associated to eigenvalue $\mu_{p,j}$ given by:

$$\begin{aligned}\mu_{p,j} &= 1 - \underbrace{r_j^2 g_j}_{=1} + 2m_j\beta_{j,j} = 2m_j\left(2r_j^3 g_j - r_j^2 s_j^2 \mathbb{1}_{\kappa=1}\right) \\ &= 4r_j m_j - 2\underbrace{r_j^2 m_j s_j^2 \mathbb{1}_{\kappa=1}}_{=\mathbb{1}_{\kappa=1}} = \frac{4m_j}{m_j+\lambda} - 2\mathbb{1}_{\kappa=1}.\end{aligned} \tag{149}$$

If $j \neq i(p)$, then $\tilde{E}_{p,j}$ and $\tilde{E}_{i^{-1}(j),i(p)}$ are two different orthogonal elements of $\mathbb{W}$. Moreover, setting $p' = i^{-1}(j)$ and $j' = i(p)$, and noticing that $i^{-1}(j') = p$ and $i(p') = j$, it follows that $\tilde{E}_{i^{-1}(j'),i(p')} = \tilde{E}_{p,j}$. Therefore, applying $\nabla F^\kappa(\mathbf{w})$ to $\tilde{E}_{p',j'}$ yields the following expression:

$$\begin{aligned}\nabla F^\kappa(\mathbf{w})[\tilde{E}_{p',j'}] &= \left(1 - r_{j'}^2 g_{j'} + m_{i(p')}\beta_{i(p'),j'}\right)\tilde{E}_{p',j'} + m_{j'}^{\frac{1}{2}} m_{i(p')}^{\frac{1}{2}}\beta_{i(p'),j'}\tilde{E}_{i^{-1}(j'),i(p')} \\ &= \left(1 - r_{j'}^2 g_{j'} + m_{i(p')}\beta_{i(p'),j'}\right)\tilde{E}_{i^{-1}(j),i(p)} + m_{j'}^{\frac{1}{2}} m_{i(p')}^{\frac{1}{2}}\beta_{i(p'),j'}\tilde{E}_{p,j}.\end{aligned} \tag{150}$$

We have shown so far that $span\left(\tilde{E}_{p,j}, \tilde{E}_{p',j'}\right)$ is a stable subspace of dimension 2 for $\nabla F^\kappa(\mathbf{w})$. Moreover, $i^{-1}(j) \leq r$ and $p \leq r$, so that, by definition, m_j and $m_{j'}$ are positive and satisfy the equation $r_i^{-2} = (m_i + \lambda)^2 = s_i^{\kappa+1} m_i^\kappa = g_i$ for $i \in \{j, j'\}$. Consequently, the terms $1 - r_{j'}^2 g_{j'}$ and $1 - r_j^2 g_j$ appearing in the expressions of $\nabla F^\kappa(\mathbf{w})[\tilde{E}_{p',j'}]$ and $\nabla F^\kappa(\mathbf{w})[\tilde{E}_{p,j}]$ vanish and $\nabla F^\kappa(\mathbf{w})$ restricted to the 2-dimensional subspace $span\left(\tilde{E}_{p,j}, \tilde{E}_{p',j'}\right)$ admits the matrix representation:

$$\beta_{j',j}\begin{pmatrix} m_{j'}, & m_{j'}^{\frac{1}{2}} m_j^{\frac{1}{2}} \\ m_{j'}^{\frac{1}{2}} m_j^{\frac{1}{2}} & m_j \end{pmatrix}. \tag{151}$$

The above matrix admits eigenvalues $\mu_{p,j}^0 = 0$ and $\mu_{p,j}^1 = \beta_{j',j}\left(m_j + m_{j'}\right) = \gamma_{p,p'}$ with associated eigenvectors $\bar{E}_{p,j}^0$ and $\bar{E}_{p,j}^1$:

$$\bar{E}_{p,j}^0 = \frac{1}{(m_j + m_{j'})^{\frac{1}{2}}}\left(m_j^{\frac{1}{2}}\tilde{E}_{p,j} - m_{j'}^{\frac{1}{2}}\tilde{E}_{p',j'}\right), \quad \bar{E}_{p,j}^1 = \frac{1}{(m_j + m_{j'})^{\frac{1}{2}}}\left(m_{j'}^{\frac{1}{2}}\tilde{E}_{p,j} + m_j^{\frac{1}{2}}\tilde{E}_{p',j'}\right). \tag{152}$$

□

F.3 Proof of Propositions B.8 and B.9 on stability of equilibria

Stability of critical point of F^0.

Proof of Proposition B.8. We will first prove that for any critical point $\mathbf{w}$ of the form in Equation (50), the Jacobian $\nabla F^0(\mathbf{w})$ has only non-negative eigenvalues, whereas for any other critical point, there

exists a negative eigenvalue. We will heavily rely on the eigen-decomposition of $\nabla F^0(\mathbf{w})$ from Proposition B.7 which establishes that all its eigenvalues are of the form:

$$\mu_{p,j} = \begin{cases} 1 - \lambda^{-2} s_j, & j \notin I_{sup} \text{ and } p > r \\ \delta_{p,j}, & j \notin I_{sup} \text{ and } p \leq r \\ \frac{4m_j}{m_j + \lambda}, & j = i(p) \text{ and } p \leq r \\ \gamma_{p,p'}, & p, p' \leq r \text{ and } j = i(p') > i(p) \\ 0, & p' \leq r \text{ and } j = i(p') \text{ and } ((p \leq r \text{ and } j < i(p)) \text{ or } p > r), \end{cases} \tag{153}$$

with $\delta_{p,j}$ and $\gamma_{p,p'}$ given by:

$$\begin{aligned} \delta_{p,j} &= \frac{(s_{i(p)} - s_j)}{\lambda(m_{i(p)} + \lambda)} \\ \gamma_{p,p'} &= \frac{\left(m_{i(p)} + m_{i(p')}\right)\left(\sqrt{s_{i(p)}} + \sqrt{s_{i(p')}}\right)}{\left(m_{i(p)} + \lambda\right)\left(m_{i(p')} + \lambda\right)}. \end{aligned} \tag{154}$$

Critical points of the form in Equation (50) are stable. To see this, it suffices to check that all eigenvalues $\mu_{p,j}$ of the Jacobian of F^0 in Equation (153) are non-negative. We treat each case separately. We only need to check the cases where $j \notin I_{sup}$ as, for all the other ones, $\mu_{p,j}$ is clearly non-negative.

- $j \notin I_{sup}$ **and** $p > r$**.** In this case we have $\mu_{p,j} = 1 - \lambda^{-2} s_j$. We need to show that necessarily $s_j \leq \lambda^2$. Recall that r is the maximal integer such that $s_{i(r)} > \lambda^2$ and $0 \leq r \leq k$. We distinguish two cases, either $r = k$ or $r < k$. If $r < k$, then having $s_j > \lambda^2$ contradicts the maximality of r, since we could include s_j to I_{sup}. The case $r = k$ is excluded here, since we would have $p > k$, which is impossible since p must always be smaller than k. Consequently, $\mu_{p,j} \geq 0$.
- $j \notin I_{sup}$ **and** $p \leq r$**.** In this case $\mu_{p,j} = \delta_{p,j}$. We need to show that $s_{i(p)} \geq s_j$. Indeed, $s_{i(1)}, \ldots, s_{i(r)}$ consists of the r highest eigenvalues, hence if $s_j > s_{i(p)}$ then j must be in I_{sup} which contradicts the condition $j \notin I_{sup}$.

Strict saddles characterization. Consider a critical point $\mathbf{w}$ of F^0. By Proposition B.6, it admits a general decomposition of the form:

$$\mathbf{w} = Q \left(\sum_{p=1}^{r} E_{p,i(p)} \right) M^{\frac{1}{2}}, \tag{155}$$

where Q is an orthogonal matrix in $\mathbb{R}^{k \times k}$, r is an integer between 0 and k, $I_{sup} = \{i(1), \ldots, i(r)\}$ is an increasing sequence of integers greater than or equal to 1, and M is a positive semi-definite diagonal operator on $\ell^2(I)$ whose only non-zero diagonal coefficients $(m_{i(p)})_{1 \leq p \leq r}$ are given by:

$$m_{i(p)}^{\frac{1}{2}} = g(s_{i(p)}, \lambda), \qquad g(s, \lambda) = (\sqrt{s} - \lambda)_+^{\frac{1}{2}}. \tag{156}$$

We are interested in critical points that are not of the form of Equation (50). Two possibilities arise:

- **Case 1**: There exists $j_0 \notin I_{sup}$ and $1 \leq p_0 \leq r$ with $s_{j_0} > s_{i(p_0)}$.
- **Case 2**: There exists $j_0 \notin I_{sup}$ so that $s_{i(p)} > s_{j_0} > \lambda^2$ for all p such that $1 \leq p \leq r$ and the rank r satisfies $r < k$.

We will treat each case separately by identifying a negative eigenvalue of the Jacobian $\nabla F^0(\mathbf{w})$.

Case 1. Consider the eigenvalue μ_{p_0,j_0}. Since $p_0 \leq r$ and $j_0 \notin I_{sup}$, μ_{p_0,j_0} must take the form:

$$\mu_{p_0,j_0} = \delta_{p_0,j_0} = \frac{(s_{i(p_0)} - s_{j_0})}{\lambda(m_{i(p_0)} + \lambda)} < 0. \tag{157}$$

μ_{p_0,j_0} is negative since by assumption $s_{j_0} > s_{i(p_0)}$. Therefore, the considered critical points cannot be stable since the Jacobian admits a negative eigenvalue.

Case 2. Set $p_0 = r + 1 \leq k$ and consider the eigenvalue μ_{p_0,j_0} of the Jacobian $\nabla F^0(\mathbf{w})$. Since $j_0 \notin I_{sup}$ and $p_0 > r$, then μ_{p_0,j_0} must be equal to:

$$\mu_{p_0,j_0} = 1 - \lambda^{-2} s_{j_0}. \tag{158}$$

By assumption $s_{j_0} > \lambda^2$, therefore $\mu_{p_0,j_0} < 0$. Therefore, such critical point cannot be stable. □

Stability of critical point of F^1.

Proof of Proposition B.9. We will first prove that for any critical point $\mathbf{w}$ of the form in Equation (52), the Jacobian $\nabla F^1(\mathbf{w})$ has only non-negative eigenvalues, whereas for any other critical point, there exists a negative eigenvalue. We will heavily rely on the eigen-decomposition of $\nabla F^1(\mathbf{w})$ from Proposition B.7 which establishes that all its eigenvalues are of the form:

$$\mu_{p,j} = \begin{cases} 1, & j \notin I_{sup} \text{ and } p > r \\ \delta_{p,j}, & j \notin I_{sup} \text{ and } p \leq r \\ \frac{4m_j}{m_j+\lambda} - 2, & j = i(p) \text{ and } p \leq r \\ \gamma_{p,p'}, & p, p' \leq r \text{ and } j = i(p') > i(p) \\ 0, & p' \leq r \text{ and } j = i(p') \text{ and } \left((p \leq r \text{ and } j < i(p)) \text{ or } p > r\right), \end{cases} \tag{159}$$

with $\delta_{p,j}$ and $\gamma_{p,p'}$ given by:

$$\begin{aligned} \delta_{p,j} &= \frac{1}{\lambda(m_{i(p)} + \lambda)} m_{i(p)} s_{i(p)} (s_{i(p)} - s_j), \\ \gamma_{p,p'} &= \frac{\left(m_{i(p)} + m_{i(p')}\right) \left(\left|s_{i(p)}\right| m_{i(p)}^{\frac{1}{2}} + \left|s_{i(p')}\right| m_{i(p')}^{\frac{1}{2}} - s_{i(p)} s_{i(p')}\right)}{\left(m_{i(p)} + \lambda\right)\left(m_{i(p')} + \lambda\right)}. \end{aligned} \tag{160}$$

Critical points of the form in Equation (52) are stable. Here $I_{sup} = I_{sup}^+ \cup I_{sup}^-$. To see why these critical points are stable, it suffices to check that all eigenvalues $\mu_{p,j}$ of the Jacobian of F^1 in Equation (159) are non-negative. We treat each case separately.

- $j \notin I_{sup}$ **and** $p > r$**.** In this case $\mu_{p,j} = 1 > 0$.
- $j \in I_{sup}$ **and** $(p > r$ **or** $(p \leq r$ **and** $j < i(p)))$**.** This corresponds to the last case in Equation (159) for which $\mu_{p,j} = 0 \geq 0$.
- $j = i(p)$ **and** $p \leq r$**.** By definition of the critical point, $m_{i(p)}$ is assumed to be positive and of the form $m_{i(p)}^{\frac{1}{2}} = \frac{1}{2}\left(\left|s_{i(p)}\right| + \sqrt{s_{i(p)}^2 - 4\lambda}\right)$. This is only possible if $\left|s_{i(p)}\right| \geq 2\sqrt{\lambda}$, which, in turn, can only happen if $\lambda \leq \frac{1}{4}$ since $|s_i| \leq 1$ for all $i \in I$. In this case, $m_{i(p)}^{\frac{1}{2}} \geq \frac{1}{2}\left|s_{i(p)}\right| \geq \sqrt{\lambda}$, so that $m_{i(p)} \geq \lambda$. Recalling that $\mu_{p,j} = \frac{4m_j}{m_j+\lambda} - 2 = 2\frac{m_{i(p)}-\lambda}{m_{i(p)}+\lambda}$ it follows that $\mu_{p,j} \geq 0$.
- $j \notin I_{sup}$ **and** $p \leq r$**.** In this case, by assumption $m_{i(p)} > 0$ which is only possible if $s_{i(p)} \neq 0$. If $s_{i(p)} s_j \leq 0$, then necessarily $s_{i(p)}(s_{i(p)} - s_j) > 0$ so that $\mu_{p,j} = \delta_{p,j} > 0$. If $s_{i(p)} s_j > 0$, then s_j cannot be such that $|s_j| > \left|s_{i(p)}\right|$. Otherwise, j would belong to the set I_{sup}, which contains the r^+ highest positive and r^- smallest negative eigenvalues. This is not possible by assumption since $j \notin I_{sup}$. Thus, necessarily $|s_j| < \left|s_{i(p)}\right|$ (non-zero eigenvalues are assumed to be distinct). Therefore, $s_{i(p)}(s_{i(p)} - s_j) > 0$ and $\mu_{p,j} = \delta_{p,j} > 0$.
- $p, p' \leq r$ **and** $j = i(p') > i(p)$**.** In this case, $\mu_{p,j} = \gamma_{p,p'}$ as defined in Equation (160). By definition of the critical point, $m_{i(p)}$ and $m_{i(p')}$ are positive and given by an equation of the form in Equation (51), so that $m_{i(p)}^{\frac{1}{2}} \geq \frac{1}{2}\left|s_{i(p)}\right|$ and $m_{i(p')}^{\frac{1}{2}} \geq \frac{1}{2}\left|s_{i(p')}\right|$. Consequently, it follows that:

$$\mu_{p,j} = \gamma_{p,p'} \geq \frac{\left(m_{i(p)} + m_{i(p')}\right)\left(\left|s_{i(p)}\right|^2 + \left|s_{i(p')}\right|^2 - 2 s_{i(p)} s_{i(p')}\right)}{2\left(m_{i(p)} + \lambda\right)\left(m_{i(p')} + \lambda\right)} \geq 0.$$

Strict saddles characterization. Consider a critical point $\mathbf{w}$ of F^1. By Proposition B.6, it admits a general decomposition of the form:

$$\mathbf{w} = Q \left(\sum_{p=1}^{r} E_{p,i(p)} \right) M^{\frac{1}{2}}, \tag{161}$$

where Q is an orthogonal matrix in $\mathbb{R}^{k \times k}$, r is an integer between 0 and k, $I_{sup} = \{i(1), \dots, i(r)\}$ is an increasing sequence of integers greater than or equal to 1, and M is a positive semi-definite diagonal operator on $\ell^2(I)$ whose only non-zero diagonal coefficients $(m_{i(p)})_{1 \leq p \leq r}$ are given by:

$$m_{i(p)}^{\frac{1}{2}} = f_{\epsilon_p}(s_{i(p)}, \lambda), \qquad f_\epsilon(s, \lambda) = \frac{1}{2} \left(|s| + \epsilon \sqrt{s^2 - 4\lambda} \right) \mathbb{1}_{|s|^2 - 4\lambda \geq 0}. \tag{162}$$

where $(\epsilon_p)_{p=1}^r$ can take any values in $\{-1, 1\}$. We are interested in critical points that are not of the form of Equation (52). Two possibilities arise:

- **Case 1**: There exists $1 \leq p_0 \leq r$ so that $m_{i(p_0)}^{\frac{1}{2}} = f_{-1}(s_{i(p_0)}, \lambda)$ and $s_{i(p_0)}^2 > 4\lambda$.
- **Case 2**: For all $1 \leq p \leq r$, we have $m_{i(p)}^{\frac{1}{2}} = f_1(s_{i(p)}, \lambda)$ and there exists j_0 and p_0 so that $|s_{j_0}| > \left|s_{i(p_0)}\right|$, $s_{j_0} s_{i(p_0)} > 0$ and $j_0 \notin I_{sup}$.

We will treat each case separately by identifying a negative eigenvalue of the Jacobian $\nabla F^1(\mathbf{w})$.

Case 1. Set $j_0 = i(p_0)$ and consider the eigenvalue μ_{p_0,j_0}. Since $p_0 \leq r$ and $j_0 = i(p_0)$, μ_{p_0,j_0} must take the form:

$$\begin{aligned}
\mu_{p_0,j_0} &= \frac{4 m_{j_0}}{m_{j_0} + \lambda} - 2 \\
&= 2 \frac{m_{j_0}^{\frac{1}{2}} + \sqrt{\lambda}}{m_{j_0} + \lambda} \left(m_{j_0}^{\frac{1}{2}} - \sqrt{\lambda} \right) \\
&= \frac{m_{j_0}^{\frac{1}{2}} + \sqrt{\lambda}}{m_{j_0} + \lambda} \left(|s_{j_0}| - 2\sqrt{\lambda} - \sqrt{s_{j_0}^2 - 4\lambda} \right) \\
&= \frac{m_{j_0}^{\frac{1}{2}} + \sqrt{\lambda}}{m_{j_0} + \lambda} \left(|s_{j_0}| - 2\sqrt{\lambda} \right)^{\frac{1}{2}} \left(\left(|s_{j_0}| - 2\sqrt{\lambda} \right)^{\frac{1}{2}} - \left(|s_{j_0}| + 2\sqrt{\lambda} \right)^{\frac{1}{2}} \right) < 0.
\end{aligned} \tag{163}$$

In the above, we used that $m_{j_0} = f_{-1}^2(s_{j_0}, \lambda)$ for the third line and that $|s_{j_0}| > 2\sqrt{\lambda}$ and $\lambda > 0$ to ensure the last inequality is strict. Since the Jacobian admits a negative eigenvalue, the considered critical points cannot be stable.

Case 2. Consider the eigenvalue μ_{p_0,j_0} of the Jacobian $\nabla F^1(\mathbf{w})$. Since $j_0 \notin I_{sup}$ and $p_0 \leq r$, μ_{p_0,j_0} must be equal to:

$$\mu_{p_0,j_0} = \frac{1}{\lambda(m_{i(p_0)} + \lambda)} m_{i(p_0)} s_{i(p_0)} (s_{i(p_0)} - s_{j_0}). \tag{164}$$

Moreover, since $s_{i(p_0)}$ and s_{j_0} have the same sign, it holds that $s_{i(p_0)}(s_{i(p_0)} - s_{j_0}) = \left|s_{i(p_0)}\right| \left(\left|s_{i(p_0)}\right| - |s_{j_0}| \right)$. Moreover, by assumption, $|s_{j_0}| > \left|s_{i(p_0)}\right|$, so that $\mu_{p_0,j_0} < 0$. This establishes that such critical point cannot be stable. □

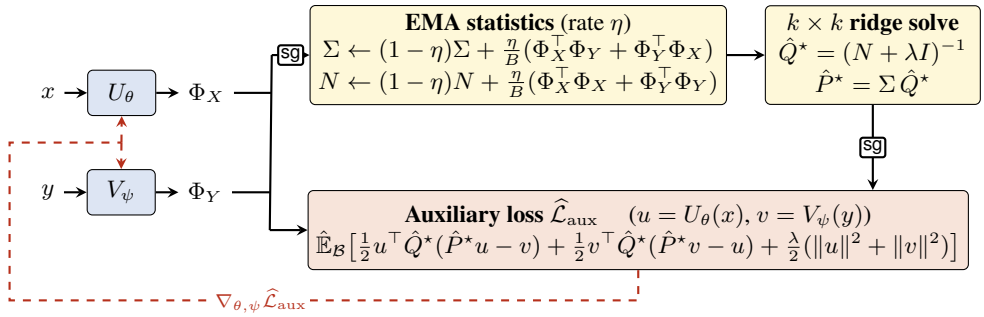


Figure G.1: **One step of Algorithm 1.** Both views of a mini-batch $\mathcal{B}$ are encoded into features $\Phi_X, \Phi_Y \in \mathbb{R}^{B\times k}$. Through a stop-gradient (sg), their second moments update EMA estimates of the signal $\Sigma_{U,V}$ and noise $N_{U,V}$, which give $\hat{P}^\star$ and $\hat{Q}^\star$ in closed form by a $k \times k$ inversion. With $\hat{P}^\star, \hat{Q}^\star$ held fixed, the auxiliary loss has the same gradient as $\mathcal{E}_{\text{PEIRA}}$ (Proposition 5.1); only the encoder parameters θ, ψ are updated (dashed). There is no target encoder and no backpropagation through the inverse; with a shared encoder, $V_\psi = U_\theta$.

G Experimental details and additional results

G.1 Experimental details

ImageNet-1K (ResNet-50, 100 epochs)

This subsection details the pretraining and evaluation pipeline used to produce the linear-probe top-1 accuracies reported in Table 1 of the main text; the corresponding hyperparameters are summarized in Table G.1.

Setup. We pretrain on the 1.28M-image ImageNet-1K training set (ILSVRC2012 split, distributed under the ImageNet Terms of Access, a custom non-commercial research license), using the two-view asymmetric augmentation stack of VICReg / BYOL [Bardes et al., 2022, Grill et al., 2020]: two 224^2 crops per image with asymmetric Gaussian blur (p=1.0 on view 1, p=0.1 on view 2) and solarization (p=0.0 on view 1, p=0.2 on view 2). All three methods use a three-layer MLP projector of hidden width 8192 (2048 for SIGReg); the output-width disparity reported in Table G.1 is theoretically motivated: SIGReg's isotropic-Gaussian target is unsatisfiable in too-high-dimensional spaces, whereas VICReg's decorrelation objective benefits from wider projectors.

Optimization. All three methods follow the VICReg LARS recipe of Bardes et al. [2022] (Layer-wise Adaptive Rate Scaling, LARS [You et al., 2017], with linear warm-up followed by cosine decay; full values in Table G.1). Gradient clipping is enabled for PEIRA only. For PEIRA, the EMA rate η of Algorithm 1 is cosine-annealed from 0.99 to 0.95 (Remark 5.2) for both k=256 and k=1024. We use a *decoupled* online linear probe trained with SGD on stop-gradient features; coupling the online probe to the backbone's cosine-decayed LR under-reads the final top-1 accuracy by 1–5 pp in our measurements, so we decouple for all methods and rely on an offline probe as the canonical comparison metric.

Evaluation. Following Appendix E of Bardes et al. [2022], we train an offline linear classifier on frozen features for 100 epochs with SGD (batch size 1024, base learning rate 0.25, cosine schedule, weight decay 10^{-6}). Augmentations at probe-training time are random-resized crop + horizontal flip; evaluation uses bicubic resize to 256 followed by a 224^2 center crop. We report top-1 accuracy on the ImageNet-1K validation set.

Compute. A single 100-epoch ResNet-50 pretraining run takes 17 to 23 hours on one node of 8× NVIDIA H200 GPUs (Table G.8). We report 3 seeds per configuration, and preliminary hyperparameter exploration consumed roughly 20 times the compute budget of the final reported runs.

Table G.1: ImageNet-1K training configurations (RN50, 100 epochs). All three methods use LARS, batch size 2048, and a separate online linear probe (SGD, fixed LR 0.3, no weight decay, no schedule).

Hyperparameter	VICReg	SIGReg	PEIRA (ours)
Optimizer / scheduler (LARS)			
Base LR	2.4	2.4	2.4
Weight decay	10^{-6}	10^{-6}	10^{-6}
Momentum	0.9	0.9	0.9
LARS trust coeff.	10^{-3}	10^{-3}	10^{-3}
Warmup epochs	10	10	10
Cosine min LR	2×10^{-3}	2×10^{-3}	2×10^{-3}
Gradient clip	–	–	1.0
Projector			
Hidden dim	8192	2048	8192
Output dim	8192	128	256 / 1024
Final-layer bias	False	True	True
Loss-specific			
Loss type	VICReg	SIGReg	PEIRA
Coefficients	$\alpha = 25, \beta = 25, \gamma = 1$	$\lambda = 0.01$	$\lambda = 0.1, \eta_{\text{init}} = 0.99, \eta_{\min} = 0.95$

CIFAR-10 (ResNet-18, 1000 epochs)

This subsection details the pretraining and evaluation pipeline used to produce the CIFAR-10 linear-probe accuracies reported in Table 1; the corresponding hyperparameters are summarized in Table G.2. The VICReg and SIGReg baselines use the default hyperparameters of the `eb_jepa` codebase [Terver et al., 2026a].

Setup. We pretrain a CIFAR-style ResNet-18 backbone (3×3 first convolution, no max-pool) on the 50k-image CIFAR-10[6] training set [Krizhevsky and Hinton, 2009], using a symmetric two-view augmentation pipeline shared by all three methods: a 32^2 random-resized crop with scale $\in [0.2, 1.0]$, color jitter (p=0.8, brightness/contrast 0.4, saturation 0.2, hue 0.1), random grayscale (p=0.2), random solarization (p=0.1), and horizontal flip (p=0.5); no Gaussian blur is used at this resolution. All three methods share a three-layer MLP projector of hidden width 2048; the per-method output widths are reported in Table G.2.

Optimization. All three methods use LARS [You et al., 2017] with linear warm-up followed by cosine decay; full values in Table G.2. Gradient clipping is disabled for VICReg and SIGReg[7], and set to 1.0 from epoch 4 onwards for PEIRA: applying clipping from the start plateaus performance, so a 4-epoch warm-up period without clipping lets the gradient estimator stabilize before the constraint is enforced.

Evaluation. Following Appendix E of Bardes et al. [2022], we train an offline linear classifier on frozen features and report top-1 accuracy on the 10k-image CIFAR-10 validation set. Numbers in Table 1 are mean ± one-sigma sample standard deviation over 3 seeds.

Compute. A single 1000-epoch ResNet-18 CIFAR-10 run takes approximately 3.5 hours on one H100 GPU. We report 3 seeds per configuration, and preliminary hyperparameter exploration consumed roughly 10× the compute budget of the final reported runs.

Additional experiments

Table G.3 summarizes the additional experiments of Appendices G.6 to G.12. Each ImageNet-1K ablation changes a single axis of the PEIRA recipe and keeps the pretraining pipeline and the offline

[6]The CIFAR-10 dataset [Krizhevsky and Hinton, 2009] is released by the University of Toronto with no explicit license and is freely available for research use.

[7]Our current SIGReg implementation multiplies the Epps–Pulley statistic by the batch size, in accordance with Balestriero and LeCun [2025]; the λ=0.1 reported in Table G.2 is therefore comparable, after this rescaling, to the λ=10 reported in Table 5 of Terver et al. [2026a].

Table G.2: CIFAR-10 training configurations (ResNet-18, 1000 epochs). All three methods share the backbone, batch size 256, the two-view symmetric augmentation pipeline above, and the offline linear-probe protocol of Appendix E of Bardes et al. [2022].

Hyperparameter	VICReg	SIGReg	PEIRA (ours)
Optimizer / scheduler (LARS)			
Base LR	0.3	0.3	0.04
Weight decay	10^{-4}	10^{-4}	10^{-4}
Warmup epochs	10	10	10
Warmup start LR	3×10^{-5}	3×10^{-5}	3×10^{-5}
Cosine min LR	0	0	0
Gradient clip	–	–	1.0 from epoch 4
Projector			
Hidden dim	2048	2048	2048
Output dim	2048	128	256 / 1024
Loss-specific			
Loss type	VICReg	SIGReg	PEIRA
Coefficients	$\alpha = 1, \beta = 1, \gamma = 80$	$\lambda = 0.1$	$\lambda = 0.7, \eta_{\text{init}} = 0.99, \eta_{\min} = 0.95$

linear-probe protocol above; seeds are drawn from $\{1, 1000, 10000\}$. Effective ranks use one of two estimators. The *in-training* estimator computes erank of backbone embeddings accumulated over 10 steps of the per-GPU batch ($N = 10\,B/8$ vectors), so its ceiling $\min(N, 2048)$ depends on the batch size; we normalize it by the erank of i.i.d. Gaussian features under the same estimator. The *held-out* estimator computes erank of backbone and projector embeddings of the final checkpoint on the first 20,480 ImageNet-1K validation images, which exceeds every width k we use. Protocol details specific to each experiment are given in its section.

Setting of the ImageNet-1K ablations. The ImageNet-1K ablations of PEIRA (projector width, EMA schedule, batch size, longer pretraining and the λ study of Figure 2 and Section G.3) anneal η from 0.9 to 0.5 and use the mini-batch estimate of Σ in $\hat{P}^\star$; everything else follows Table G.1. Their k=256 baseline matches Table 1 within seed noise.

Table G.3: **Additional experiments.** Base recipe, varied factor, number of seeds per configuration and reported metrics. IN1K: ImageNet-1K, ResNet-50, 100 epochs, recipes of Table G.1 (PEIRA ablations: setting of Section G.1); top-1: offline linear probe.

Experiment	Base recipe	Varied factor (seeds)	Metrics
Width (Appendix G.6)	IN1K PEIRA	$k \in \{64, \dots, 2048\}$ (1–3)	top-1, erank, transfer
Long training (Appendix G.7)	IN1K PEIRA, VICReg	300 epochs (1)	top-1
EMA schedule (Appendix G.8)	IN1K PEIRA	$(\eta_{\text{init}}, \eta_{\min})$, 6 cells (1)	top-1, erank
Batch size (Appendix G.8)	IN1K PEIRA	$B \in \{512, \dots, 4096\}$, LR $\propto B$ (1)	top-1, erank
Wall-clock (Appendix G.9)	IN1K, all three	none	s / epoch
Transfer (Appendix G.10)	IN1K, all three	8 target datasets (1–3)	accuracy
Synthetic (Appendix G.11)	linear encoders	λ; init. variance (1)	recovered rank
CIFAR-10 (Appendix G.12)	PEIRA, SD	init. variance 10^{-2}–10 (1)	accuracy, erank, align.

G.2 Per-epoch correlation of $\mathcal{E}_{\text{PEIRA}}$ with downstream accuracy

This subsection records the closed-form values that the auxiliary loss $\mathcal{L}_{\text{aux}}$ and the trace-form objective $\mathcal{E}_{\text{PEIRA}}$ take at equilibrium, and explains why Figure 2 (right) plots $\mathcal{E}_{\text{PEIRA}}$ rather than $\mathcal{L}_{\text{aux}}$ against linear-probe top-1.

Lemma G.1 (Closed-form value of $\mathcal{L}_{\text{aux}}$ at the closed-form regressor)**.** *For any $(U, V) \in \mathcal{U} \times \mathcal{V}$, evaluating $\mathcal{L}_{\text{aux}}$ of Proposition 5.1 at the closed-form $P^\star = \Sigma_{U,V}(N_{U,V} + \lambda I)^{-1}$ and $Q^\star = (N_{U,V} + \lambda I)^{-1}$ of Equation (5) gives*

$$\mathcal{L}_{\text{aux}}(U, V; P^\star, Q^\star) = \tfrac{\lambda}{2}\big[Tr(N_{U,V}) - Tr(\Sigma_{U,V}(Q^\star)^2)\big]. \tag{165}$$

Proof. From Equation (73), $\mathcal{L}_{\text{aux}}(U,V;P,Q) = \frac{1}{2}Tr\big((QP+\lambda I)N_{U,V} - Q\Sigma_{U,V}\big)$. Setting $P = P^\star = \Sigma_{U,V}Q^\star$ and using the algebraic identity $N_{U,V}Q^\star = I - \lambda Q^\star$ (which follows from $N_{U,V} = (N_{U,V}+\lambda I) - \lambda I$),

$$\begin{aligned} Tr(Q^\star P^\star N_{U,V}) &= Tr\big(Q^\star\Sigma_{U,V}Q^\star N_{U,V}\big) = Tr\big(Q^\star\Sigma_{U,V}(I-\lambda Q^\star)\big) \\ &= Tr(Q^\star\Sigma_{U,V}) - \lambda Tr(\Sigma_{U,V}(Q^\star)^2), \end{aligned}$$

where the last equality uses trace cyclicity ($Q^\star$ symmetric). Substituting,

$$\mathcal{L}_{\text{aux}}(U,V;P^\star,Q^\star) = \tfrac{1}{2}\big[Tr(Q^\star\Sigma_{U,V}) - \lambda Tr(\Sigma_{U,V}(Q^\star)^2) + \lambda Tr(N_{U,V}) - Tr(Q^\star\Sigma_{U,V})\big], \quad (166)$$

which simplifies to (165). □

Equilibrium values of $\mathcal{L}_{\text{aux}}$ and $\mathcal{E}_{\text{PEIRA}}$. The identity (165) of Lemma G.1 holds for any (U,V). At the stable equilibrium of Proposition 4.3, the spectral filter of Proposition 4.4 forces $N_{U,V}$ and $\Sigma_{U,V}$ to be simultaneously diagonalizable, with eigenvalues $(\sqrt{c_i}-\lambda)$ and $c_i(\sqrt{c_i}-\lambda)$ respectively on every active mode i and zero elsewhere; in particular, $\Sigma_{U,V}(Q^\star)^2$ shares its eigenvalues with $N_{U,V}$ on the active subspace and both matrices vanish on its orthogonal complement, so $Tr(N_{U,V}) = Tr(\Sigma_{U,V}(Q^\star)^2)$ and $\mathcal{L}_{\text{aux}}(U,V;P^\star,Q^\star) = 0$. The trace-form objective of Equation (6) reaches its minimum (14), $\mathcal{E}_{\text{PEIRA}} = -\frac{1}{2}\sum_{i=1}^{r_{\max}}(\sqrt{c_i}-\lambda)^2 < 0$. Figure 2 (right) plots $\mathcal{E}_{\text{PEIRA}}$ (evaluated post-hoc at the closed-form regressor $P_{U,V}$ from the running covariance buffers of Algorithm 1) against online linear-probe top-1 across our ResNet-50 / IN1K runs.

SIGReg counterpart. For comparison, Figure G.3 (right) shows the analogous per-epoch correlation between the negative SIGReg total training loss $-\mathcal{L}_{\text{train}}$ and online linear-probe top-1, across the SIGReg ResNet-50 / IN1K runs of Section G.4; the negative loss is also strongly correlated with downstream accuracy in the λ range covered by our sweep, mirroring the convention used for $-\mathcal{E}_{\text{PEIRA}}$ in the main paper.

G.3 Effective rank vs. λ on ImageNet-1K

This subsection tracks the entropy-based effective rank of PEIRA's backbone and projector embeddings during ImageNet-1K pretraining as a function of λ, both at end of training and along the trajectory.

Metric. We track the entropy-based effective rank as a basis- and label-free summary of representational diversity, sensitive to partial collapse without requiring a labelled probe; this is the standard SSL diagnostic introduced by Garrido et al. [2023a]. Concretely, for an embedding batch $Z \in \mathbb{R}^{N\times d}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_d$ and $p_i = \sigma_i / \sum_j \sigma_j$,

$$\text{erank}(Z) \;=\; \exp\left(-\sum_{i:p_i>0} p_i \log p_i\right) \;\in\; [1,d]. \quad (167)$$

Since $\sigma_i(Z) = \sqrt{N\,\lambda_i(\hat{\Sigma}_Z)}$ with $\hat{\Sigma}_Z = Z^\top Z/N$, erank is a one-number summary of the empirical covariance eigenspectrum.

What is reported. Figure G.2 reports end-of-training (epoch $\geq$ 95) erank of the backbone (d=2048) and the projector (d=256) across $\lambda \in \{0.025, 0.05, 0.1, 0.2, 0.5\}$ on the PEIRA configuration of Table G.1 (1–4 seeds per cell, 13 runs total). The per-epoch backbone-rank trajectory is in Figure 2 (middle) of the main text.

Observations.

- Projector erank sits at $\sim$248 across every λ, within $\sim$3% of the 256-d cap.
- Backbone erank is $\sim$ 1314 at $\lambda \in \{0.025, 0.05\}$ and $\sim$ 1283–1290 at $\lambda \in \{0.1, 0.2, 0.5\}$, a $\sim$30-unit ($\sim$2%) drop, then a plateau.
- erank stays clear of any collapse regime across the full 20$\times$ λ span.

Interpretation. Recall from Section 4 that R denotes the number of canonical correlations of the data distribution (Definition 4.1) and $r_{\max} \leq \min(k, R)$ is the largest rank supporting a non-collapsed equilibrium of Proposition 4.4. We draw three conclusions from the erank measurements above. First, the projector (d=256) sits within $\sim 3\%$ of its dimensionality cap, i.e. the projector output dimension $k = 256$ that bounds $r_{\max}$: the projector spectrum is set by this cap rather than by λ whenever k binds, which is the regime our sweep operates in. Second, for λ small relative to the leading $\sqrt{c_i}$, Proposition 4.4 predicts a gradual contraction of the active eigenvalues $\frac{1}{4}(\sqrt{c_i} - \lambda)^2$ without a strict-rank drop; the mild $\sim 2\%$ backbone contraction we observe across the $20\times$ λ span is consistent with this regime, although we do not claim to observe a change in $r_{\max}$ itself. Third, Proposition 4.4 together with Proposition 4.3 rules out collapse as a stable equilibrium of the PEIRA dynamics, and the absence of any erank collapse on real data corroborates this. Two caveats: the λ window stays well below the regime in which Proposition 4.4 would predict visible mode killing ($\lambda \to \sqrt{c_1}$), so these numbers do not test the strict-rank prediction $r_{\max} = \#\{i : \sqrt{c_i} > \lambda\}$; and erank is a one-number entropy summary, so we do not resolve the full eigenspectrum here.

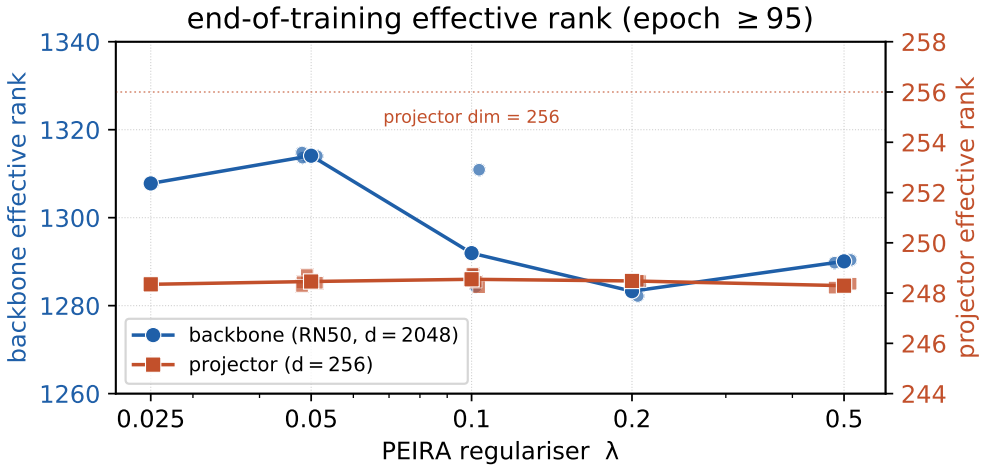


Figure G.2: **Effective rank vs. λ on ImageNet-1K.** End-of-training erank [Garrido et al., 2023a] of the encoder during ImageNet-1K pretraining (RN50, 100 epochs, projector recipe of Table G.1; epoch ≥ 95, mean across seeds, scatter shows individual seeds) for the backbone (left axis, blue, d=2048) and projector (right axis, orange, d=256) across $\lambda \in \{0.025, 0.05, 0.1, 0.2, 0.5\}$. The per-epoch backbone-rank trajectory is in Figure 2 (middle) of the main text.

G.4 Comparison with SIGReg on ImageNet-1K

We report the SIGReg counterparts to the three PEIRA diagnostics of Figure 2 of the main text in a single 3-panel figure (Figure G.3). The runs use the SIGReg ablation configuration (p_h=2048, p_o=128), which we use as SIGReg's frozen baseline on ResNet-50 / IN1K.

Sensitivity to λ. Figure G.3 (left) sweeps SIGReg's coefficient λ over the range we tested on ResNet-50 / IN1K. Within the tested span $\lambda \in [0.005, 0.02]$ the mean linear-probe top-1 swings by 0.80 pp; values $\lambda \geq 0.1$ collapsed in our preliminary tests and were excluded from the reported sweep. Outside the tested range, Balestriero and LeCun [2025] report empirical stability over $\lambda \in [10^{-3}, 5{\cdot}10^{-1}]$ on ImageNet-100 with V=8 multi-crop views.

Loss as a model-selection signal. Figure G.3 (right) plots the per-epoch negative SIGReg total training loss $-\mathcal{L}_{\text{train}}$ against online linear-probe top-1; the negative loss is strongly correlated with downstream accuracy in the λ range covered by our sweep, mirroring the convention used for $-\mathcal{E}_{\text{PEIRA}}$ in Figure 2 / Section G.2.

Backbone effective-rank trajectory. Figure G.3 (middle) tracks the entropy-based effective rank of the RN50 backbone embeddings throughout pretraining for $\lambda \in \{0.005, 0.01, 0.02\}$. The narrow span tested here (a $4\times$ ratio) is too small to expose a clear separation between curves; this stands in

contrast to the PEIRA trajectory of Figure 2 (middle), where the $20\times$ span $\lambda \in [0.025, 0.5]$ produces a visible ordering of early-epoch ranks. We report this for completeness rather than as evidence of a qualitative gap between the two methods.

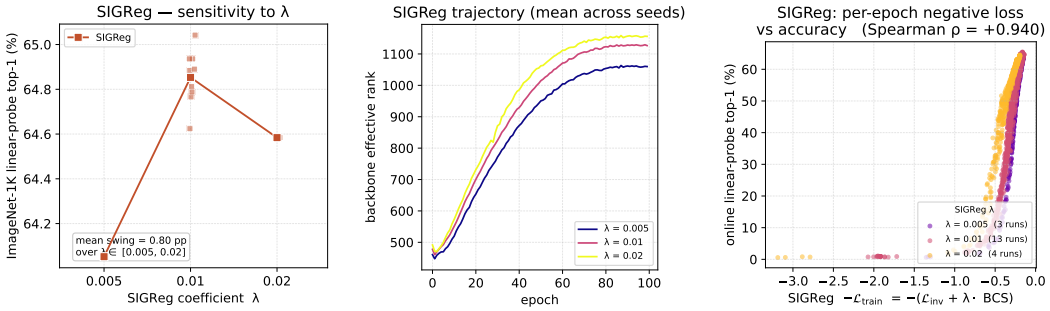


Figure G.3: **ImageNet-1K diagnostics for SIGReg.** ResNet-50, 100 epochs, frozen ablation projector p_h=2048, p_o=128; 1–3 random seeds per cell. Mirrors Figure 2 of the main text. *Left:* sensitivity of the online linear-probe top-1 to λ at end of training; each dot is one sweep cell, $\lambda \geq 0.1$ collapsed in preliminary tests and is not shown. *Middle:* backbone entropy-based effective rank [Garrido et al., 2023a] versus epoch, colored by λ. *Right:* per-epoch negative SIGReg total training loss $-\mathcal{L}_{\text{train}}$ versus online linear-probe top-1 (one point per epoch per run); per-run Spearman correlation reported in the legend.

G.5 Comparison with VICReg on ImageNet-1K

We compare PEIRA's diagnostics with VICReg's at the published RN50 / 100-epoch recipe, restricted to the analogues of the middle and right panels of Figure 2 (Figure G.4, 18 runs: 3 seeds across 6 design-axis ablations of the frozen recipe, all using the `vicreg_ref` loss with $\lambda = \mu = 25, \nu = 1$).

Coefficient-sensitivity ablation. We do not include a sensitivity panel: our reproduction did not sweep any of (λ, μ, ν) at the published recipe. Bardes et al. [2022]'s Table 7 (App. D.4) reports a joint coefficient ablation at the same backbone and budget: turning on the covariance term lifts top-1 by $+11.1$ pp ($57.5\% \to 68.6\%$), while the diagonal sweep $\lambda = \mu \in \{5, 10, 25, 50\}$ at fixed $\nu = 1$ produces only a $\sim$0.5 pp spread, consistent with VICReg being largely insensitive to its coefficients in the stable regime $\nu < \mu$, $\lambda \approx \mu$.

Loss as a model-selection signal. Backbone effective rank (Figure G.4, left) climbs to $\sim$1300–1400 across all recipes without collapsing; per-epoch negative training loss $-\mathcal{L}_{\text{train}}$ is strongly correlated with online linear-probe top-1 within every recipe (per-recipe Spearman $\rho \geq +0.998$, Figure G.4, right), mirroring PEIRA (Section G.2) and SIGReg (Section G.4) and supporting use of the total loss as a label-free model-selection signal within a recipe.

G.6 Projector output width

Table G.4 varies the projector output width k in the ablation setting (Section G.1), all other hyperparameters fixed. ImageNet-1K top-1 rises with k up to k=1024, where it approaches VICReg, and eases at k=2048. In this setting, k=1024 transfers slightly worse on average than k=256; with the recipe of Table G.1, it is the strongest PEIRA model on both ImageNet-1K and transfer (Tables 1 and G.9). The other ImageNet-1K results use k=256.

The projector effective rank stays close to its cap k at every width, consistent with $r_{\max} \leq \min(k, R)$ of Proposition 4.4 with k binding, and the backbone effective rank grows monotonically with k. The $k \times k$ system solved by Algorithm 1 stayed well conditioned at every width.

G.7 Longer pretraining

To test whether the ImageNet-1K gap to VICReg is an artifact of the 100-epoch budget, we pretrain PEIRA (ablation setting, Section G.1) and VICReg for 300 epochs with the cosine schedule stretched

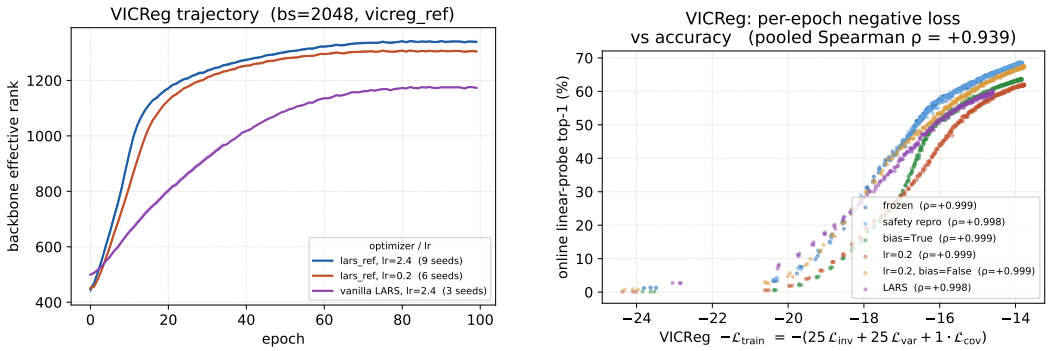


Figure G.4: **ImageNet-1K diagnostics for VICReg.** ResNet-50, 100 epochs, batch size 2048, projector p_h=p_o=8192, VICReg loss with $\lambda = \mu = 25, \nu = 1$; 3 seeds across 6 design-axis ablations of the frozen recipe. Mirrors the middle and right panels of Figure 2 and Figure G.3. *Left:* backbone effective rank [Garrido et al., 2023a] versus epoch (faint lines: per-seed; bold: per-group mean). The six recipes are pooled into three groups along the true axis of variation (optimizer / lr); the bias and probe-protocol toggles within each group leave backbone effective rank unchanged within seed-spread. *Right:* per-epoch negative training loss $-\mathcal{L}_{\text{train}}$ versus online linear-probe top-1, colored by recipe, with per-recipe Spearman in the legend.

Table G.4: **Projector output width.** PEIRA on ImageNet-1K (ResNet-50, 100 epochs, output width k varied). Top-1: offline linear probe, mean $\pm$ sample standard deviation over seeds $\{1, 1000, 10000\}$ (1 seed at k=64). Projector and backbone erank [Garrido et al., 2023a]: end-of-training checkpoints, 20,480 held-out ImageNet-1K validation images, mean over seeds. Normalized erank: backbone erank of the in-training estimator divided by that of i.i.d. Gaussian features under the same estimator. Transfer: mean linear-probe accuracy over the eight datasets of Table G.9 (1 seed at k=256, 2 seeds at k=1024). All rows use the ablation setting (Section G.1). VICReg (k=8192) reaches $68.81 \pm 0.09\%$.

k	Top-1 (%)	Seeds	Proj. erank $/k$	Backbone erank	Norm. erank	Transfer (%)
64	61.74	1	–	–	0.70	–
256	66.50 ± 0.02	3	0.987	1461	0.73	70.72
512	67.88 ± 0.10	3	0.983	1479	0.73	–
1024	68.50 ± 0.13	3	0.976	1495	0.74	69.63
2048	68.14 ± 0.15	3	0.965	1504	–	–

accordingly, and probe offline at epochs 200 and 300. Both methods gain about as much from the longer schedule, so the gap stays nearly constant (Table G.5): it is an offset, not a budget effect.

Table G.5: **Longer pretraining on ImageNet-1K.** ResNet-50, VICReg recipe of Table G.1, PEIRA ablation setting (Section G.1), offline linear-probe top-1 (%). The 100-epoch row is the 3-seed mean of each recipe. The 200- and 300-epoch rows come from one 300-epoch run per method (seed 1); the 200-epoch checkpoint is taken mid-schedule, with the learning rate still at 0.63 (peak 2.4), so it is comparable across methods but understates both.

Epochs	PEIRA	VICReg	Gap (pp)
100	66.5	68.8	2.3
200	66.1	68.4	2.3
300	68.3	70.7	2.4

G.8 EMA schedule and batch size

EMA schedule. PEIRA tracks the covariance buffers that define $\hat{P}^\star$ and $\hat{Q}^\star$ with an EMA whose rate η is cosine-annealed from η_{init} to $\eta_{\min}$ (Remark 5.2). Top-1 changes little across the schedules of Table G.6 and the backbone erank does not move: PEIRA is robust to the quality of the predictor estimate over this range.

Batch size. Table G.7 varies the global batch size with the learning rate scaled linearly ($0.3\times B/256$). Top-1 peaks at moderate batch sizes (1024–2048) and eases at both extremes under this fixed scaling rule. The raw backbone erank grows with the batch size because the in-training estimator accumulates $N = 10\,B/8$ embeddings, so its ceiling $\min(N, 2048)$ grows with B; normalized by the erank of i.i.d. Gaussian features under the same estimator, it is flat (0.71–0.75).

Table G.6: **EMA schedule.** PEIRA, ImageNet-1K, ResNet-50, 100 epochs, ablation setting (Section G.1) with $(\eta_{\text{init}}, \eta_{\min})$ varied; offline top-1 (%), seed 1. Default: $(0.9, 0.5)$. The backbone erank stays at ~ 1280 (normalized 0.72) in every cell.

η_{init}	$\eta_{\min}$=0.3	$\eta_{\min}$=0.5
0.5	65.88	66.05
0.7	66.48	66.20
0.9	66.30	66.21

Table G.7: **Batch size.** Same setting as Table G.6 with the global batch size B varied and the learning rate scaled as $0.3 \times B/256$. Raw erank: backbone, in-training estimator, with its ceiling (i.i.d. Gaussian features, same estimator); normalized erank = raw / ceiling.

B (LR)	512 (0.6)	1024 (1.2)	2048 (2.4)	4096 (4.8)
Top-1 (%)	63.85	66.11	66.23	64.50
Raw erank	446	823	1284	1441
Ceiling	612	1155	1770	1931
Norm. erank	0.73	0.71	0.73	0.75

G.9 Wall-clock cost and conditioning of the $k \times k$ solve

Table G.8 reports the training time of the three ImageNet-1K recipes of Table G.1 on one node of 8× NVIDIA H200 GPUs at batch size 2048. The per-epoch times include the regularized linear solve of Algorithm 1. This solve acts on a $k \times k$ system, with k the SSL latent dimension, and is a negligible fraction of each step. PEIRA is the fastest of the three methods per epoch; for VICReg, the 8192-d expander dominates the difference. The system $\hat{N}_{U,V} + \lambda I$ is symmetric positive definite with smallest eigenvalue at least λ, so its condition number is bounded by $1 + \|\hat{N}_{U,V}\|/\lambda$; we observed no inversion-related numerical failure in any run, including the wider projectors of Appendix G.6. For much larger k, the joint updates of Remark 5.3 remove the inversion altogether.

Table G.8: **Training cost on ImageNet-1K.** ResNet-50, batch size 2048, one node of 8× NVIDIA H200, recipes of Table G.1. Per-epoch time includes data loading, the forward and backward passes, and, for PEIRA, the $k \times k$ regularized solve. PEIRA is 24% faster per epoch than VICReg and 23% faster than SIGReg.

Method	Projector output k	s / epoch	100-epoch wall-clock
SIGReg	128	809	$\sim$22.5 h
PEIRA	256	626	$\sim$17.4 h
VICReg	8192	822	$\sim$22.8 h

G.10 Transfer to the LeJEPA-8 suite

We evaluate frozen ImageNet-1K ResNet-50 backbones on the eight classification datasets used by Balestriero and LeCun [2025]: CIFAR-10 and CIFAR-100 [Krizhevsky and Hinton, 2009], DTD [Cimpoi et al., 2014], Flowers102 [Nilsback and Zisserman, 2008], Oxford-IIIT Pets [Parkhi et al., 2012], Food101 [Bossard et al., 2014], FGVC-Aircraft [Maji et al., 2013] and Stanford Cars [Krause et al., 2013].

Protocol. Images are resized to 256 and center-cropped to 224^2 with ImageNet normalization; the 2048-d pooled backbone features are extracted once without augmentation. A linear head is trained on the frozen features with Adam (learning rate 10^{-3}, cosine schedule, 100 epochs); the weight decay is selected in $\{0, 10^{-4}, 10^{-2}\}$ on a 10% held-out split of the training set, and the head is refit on the full training set. We report top-1 accuracy on CIFAR-10, CIFAR-100 and Food101 and mean per-class recall on the other five datasets. Checkpoints: PEIRA k=256 and k=1024, SIGReg k=128, VICReg k=8192 (Table G.1).

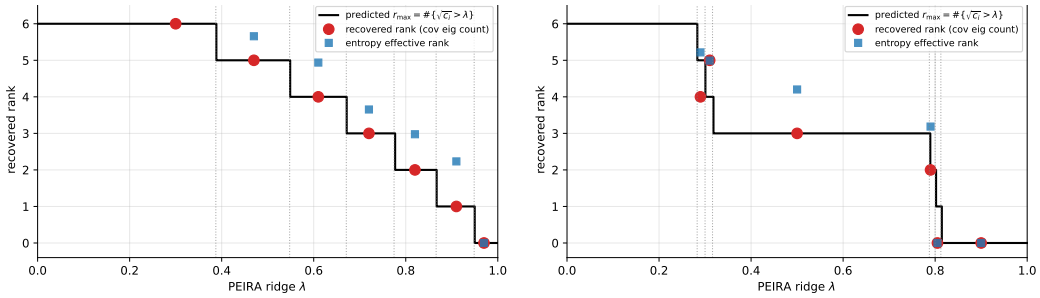
Results. PEIRA (k=256) exceeds SIGReg by 2.6 pp on average and on 7 of 8 datasets (all but DTD), and trails VICReg by 1.2 pp on average, less than its 2.4 pp ImageNet-1K gap (Table G.9). The k=1024 PEIRA checkpoints, stronger on ImageNet-1K, also transfer slightly better on average (71.09% vs. 70.89%) and are best on CIFAR-10, Pets and Food101.

Table G.9: **Linear transfer from ImageNet-1K.** Frozen ResNet-50 backbones pretrained for 100 epochs on ImageNet-1K (recipes of Table G.1; VICReg and SIGReg: seed 1; PEIRA: mean over seeds $\{1, 1000, 10000\}$); linear probe protocol of Appendix G.10. IN1K: offline ImageNet-1K top-1 of the same checkpoints. Metric: top-1 (%) for CIFAR-10/100 and Food101, mean per-class recall (%) otherwise. Bottom row: PEIRA with a 1024-d projector (Appendix G.6). Best per column in bold.

Method (k)	IN1K	C10	C100	DTD	Flowers	Pets	Food	Aircraft	Cars	Mean
VICReg (8192)	68.80	90.55	**73.98**	68.24	**85.80**	82.94	69.92	**54.79**	**50.35**	**72.07**
SIGReg (128)	66.46	90.16	71.94	**70.05**	82.56	77.44	66.87	47.86	39.25	68.27
PEIRA (256)	66.45	90.46	72.89	68.30	85.76	79.52	68.47	53.11	48.64	70.89
PEIRA (1024)	**69.17**	**90.63**	73.54	69.08	84.62	**83.20**	**70.02**	50.95	46.70	71.09

G.11 Synthetic two-view Gaussian data

Strict-rank staircase. On real data λ stays far below $\sqrt{c_1}$, so the mode filtering of Proposition 4.4, $r_{\max} = \#\{i : \sqrt{c_i} > \lambda\}$, is not observable (Section G.3). We therefore train two linear encoders $\mathbb{R}^{32} \to \mathbb{R}^{12}$ with PEIRA for 250 epochs on 20,000 pairs of two-view Gaussian data with six prescribed canonical correlations, sweeping λ with one value between each pair of consecutive thresholds $\sqrt{c_i}$. The recovered rank steps down by one each time λ crosses a $\sqrt{c_i}$ and equals $r_{\max}$ at every λ (Figure G.5a). With clustered correlations (Figure G.5b), it drops at each cluster and is resolved up to ± 1 inside a cluster, where thresholds differ by only 0.01–0.02.



(a) (b)

Figure G.5: **Strict-rank prediction on synthetic two-view Gaussian data.** Linear encoders trained with PEIRA on two-view Gaussian data with 6 prescribed canonical correlations c_i, swept over λ. Black: predicted $r_{\max} = \#\{i : \sqrt{c_i} > \lambda\}$; dotted grey: thresholds $\sqrt{c_i}$; red: recovered rank (covariance eigenvalues above $5 \cdot 10^{-3}$); blue: entropy effective rank. *(a)* Well-separated spectrum $c = (0.90, 0.75, 0.60, 0.45, 0.30, 0.15)$. *(b)* Clustered spectrum $c = (0.66, 0.64, 0.62, 0.10, 0.09, 0.08)$.

Recovery of CCA directions across initialization scales. We consider a linear Gaussian two-view model with known CCA directions and spectra, using a fixed, orthogonalized batch to remove sampling noise. We train linear encoders for 400,000 updates from Gaussian initializations with standard deviation $\rho_0 \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. Both methods use exact predictors; self-distillation follows the fast-predictor dynamics of Section 4. We track ground-truth directional alignment and spectral ratios relative to the analytical target, which retains four canonical modes. PEIRA recovers all four target directions and suppresses inactive modes at every scale (Figure G.6). Self-distillation recovers four directions at $\rho_0 = 10^{-2}$ and 10^{-3}, three at 10^{-4}, one at 10^{-5}, and none at 10^{-6}. Thus, even with an exact predictor, self-distillation exhibits initialization-dependent dimensional collapse, whereas PEIRA consistently recovers the target subspace within the training budget.

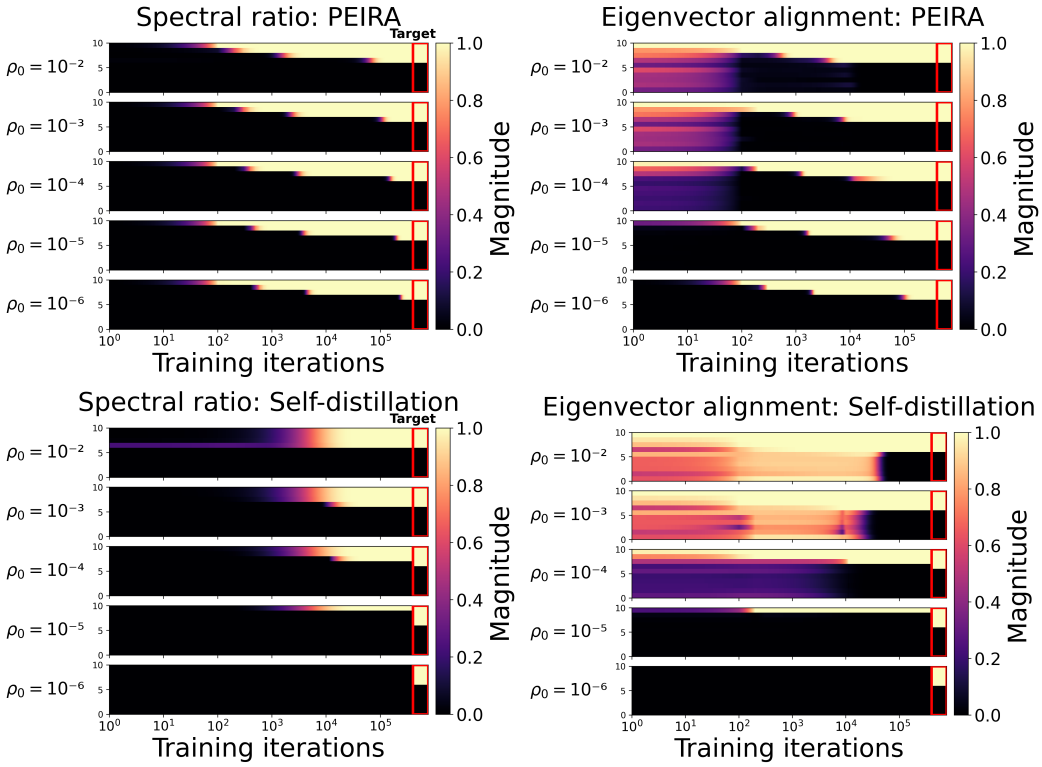


Figure G.6: **Learning dynamics on deterministic toy data.** Spectral ratios and eigenvector alignment for PEIRA and self-distillation across initialization scales $\rho_0 \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, ordered from top to bottom. Each heatmap shows the evolution of individual components over training iterations on a logarithmic horizontal scale; color indicates magnitude. The red-outlined block shows the reference target, with four active components. PEIRA recovers the target spectral structure across the initialization scales, whereas self-distillation exhibits increasing dimensional collapse as initialization becomes smaller.

G.12 CIFAR-10: robustness to initialization scale

We study the sensitivity of PEIRA and self-distillation to the initial weight scale on CIFAR-10. We vary the multiplier applied to convolutional and linear weights over three orders of magnitude, $\sigma_0 \in \{10^{-2}, 10^{-1}, 1, 10\}$. Both methods use ResNet-18, a 1024-dimensional projector, the same augmentations, batch size 256, and 1000 training epochs. The self-distillation baseline uses stop-gradient targets and a learned predictor updated by an implicit regression step after each encoder update. We retain method-specific optimization settings: PEIRA uses λ=0.7 with the two covariance-estimation settings described in the caption of Figure G.7, an encoder learning rate of 0.04, gradient clipping, and float16, whereas self-distillation uses a learning rate of 0.001, implicit step-size of 0.01, no gradient clipping, and bfloat16 as we found this setting to yield improved results. Results are obtained for three seeds.

We evaluate classification accuracy using an online linear probe trained on detached backbone features. To characterize dimensional collapse, we compute the entropy-based effective rank of the projector's uncentered second-moment matrix and the mean alignment across its 1024 directions. These diagnostics are computed separately for each seed before averaging.

Figure G.7 shows that PEIRA learns useful representations across all four initialization scales. Its mean final accuracy ranges from 89.56% to 91.65%, with effective rank 364–467 and mean alignment 0.44–0.54. Larger initialization delays the growth of alignment and effective rank, but these quantities recover during training. Self-distillation is substantially more sensitive to initialization: its mean accuracy is 77.69% and 81.10% at scales 0.1 and 1, respectively, but falls to 49.59% at 0.01 and 46.00% at 10. Across scales, its final effective rank remains approximately 1–61 and mean alignment

0.009–0.070, indicating pronounced dimensional collapse. The smallest initialization also produces substantial seed-to-seed variability in accuracy ($49.59 \pm 24.09\%$, mean $\pm$ sample standard deviation). Under these training settings, PEIRA therefore exhibits more consistent accuracy and avoids the severe dimensional collapse observed for self-distillation.

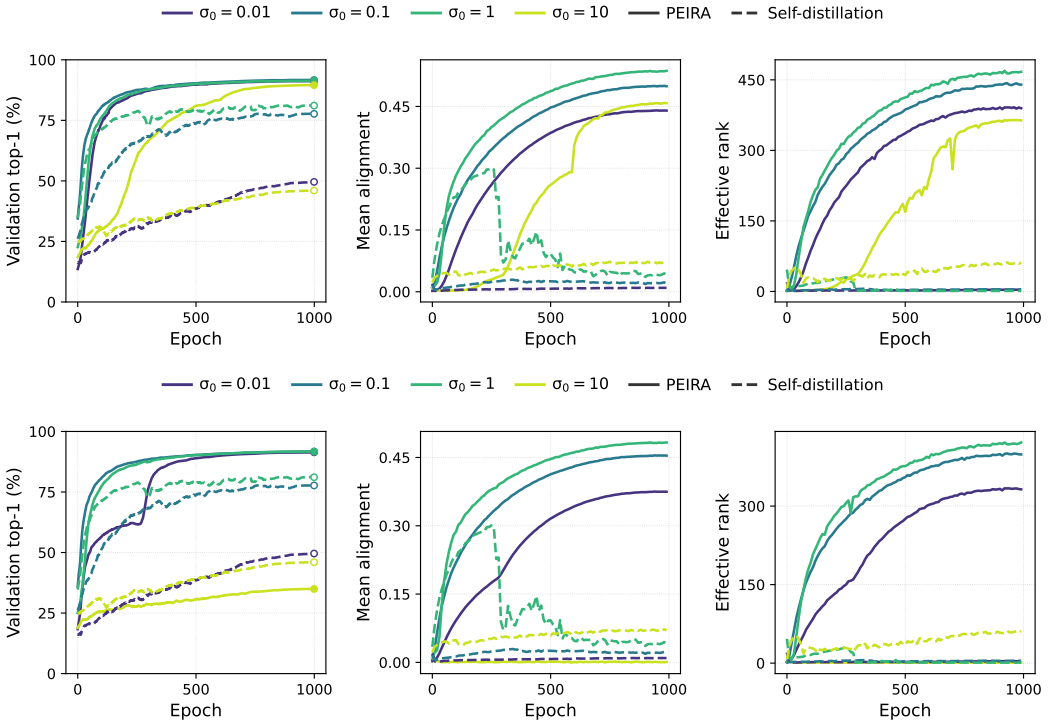


Figure G.7: **Initialization sensitivity and dimensional collapse on CIFAR-10.** PEIRA (solid) and self-distillation (dashed) use ResNet-18 and are compared at initialization scales $\sigma_0 \in \{10^{-2}, 10^{-1}, 1, 10\}$ (colors). *Left:* online linear-probe validation accuracy. *Middle:* mean inter-view alignment across projector directions. *Right:* entropy-based effective rank of the projector second-moment matrix. *Top:* PEIRA uses an EMA of the inter-view covariance and minibatch estimates of the within-view covariance, with the EMA coefficient scheduled from 0.80 to 0.50. *Bottom:* PEIRA uses EMAs of both covariances, with the coefficient scheduled from 0.99 to 0.95. The same self-distillation runs appear in both rows. Curves are averaged over three seeds. Accuracy curves use a centered 21-epoch moving average, and markers show the unsmoothed final means. For $\sigma_0 \leq 1$, PEIRA achieves higher final accuracy, alignment, and effective rank than self-distillation under both configurations. At $\sigma_0 = 10$, PEIRA recovers after an initial drop in effective rank in the top row, but remains near rank one in the bottom row.

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 abstract and introduction list three claims that match the paper's results: (i) a nonlinear-CCA characterization of the asymmetry-based JEPA dynamics with explicit stability and collapse conclusions (Propositions 4.1 and 4.2); (ii) the introduction of PEIRA as a non-contrastive objective defined through the trace of the optimal regularized regressor (Section 3, Section 5); and (iii) the recovery of canonical-correlation subspaces with collapse ruled out (Proposition 4.3), supported by experiments in Section 6.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The Limitations paragraph of the Conclusions section discusses both theoretical and empirical limitations: the fast-predictor / continuous-time idealization, the distinct-canonical-correlations and density-ratio assumptions, and the $\mathbb{P}_X = \mathbb{P}_Y$ requirement of Proposition 4.2 on the theoretical side; and the ImageNet-1K performance gap to VICReg/SIGReg (Table 1) together with the limited empirical scope (no ViT, no transfer benchmarks) on the empirical side.

   Guidelines:

   - The answer [N/A] 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: All formal results carry explicit assumptions (Assumption (A), distinct canonical correlations, $\lambda < 1$) stated either in the proposition or in the referenced appendix definition; full proofs of Propositions 4.1 to 4.4, and the gradient/Lipschitz proposition appear in Appendix A.1 and the Hilbert–Schmidt analysis in the general proofs appendix, with proof sketches included in Section 3.

   Guidelines:

   - The answer [N/A] 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: Appendix G.1 documents the datasets, backbones, projector geometry, optimizer, augmentation stack, schedule, and loss-specific coefficients for IN1K (Table G.1); the linear-probe protocol explicitly follows Appendix E of Bardes et al. [2022] and reported numbers are means $\pm$ one-sigma standard deviation over 3 seeds.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: Code, configuration files and reproduction commands for the main experiments are released at `https://github.com/Basile-Terv/PEIRA`, and pretrained checkpoints at `https://huggingface.co/peira-ssl/PEIRA`. All datasets are public (ImageNet-1K, CIFAR-10 and the transfer datasets of Appendix G.10).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and evaluation details are presented across Section 6 and the experimental-details appendix (Appendix G.1), including optimizer, batch size, schedule, augmentation pipeline, projector geometry, gradient clipping, and per-method loss coefficients in Table G.1; the IN1K linear-probe protocol follows Appendix E of Bardes et al. [2022].

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: All reported numbers in Table 1 are means $\pm$ one-sigma sample standard deviation over 3 random seeds, with the variability arising from random network initialization, data ordering, and augmentation sampling. The seed protocol is described in Section 6.

   Guidelines:

   - The answer [N/A] 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: Appendix G.1 reports that one 100-epoch ImageNet-1K ResNet-50 pretraining run takes 17 to 23 hours on a single node of $8\times$ NVIDIA H200 GPUs. The smaller CIFAR-10 diagnostic runs use a single H100 GPU. We report 3 seeds per configuration; preliminary hyperparameter exploration consumed roughly $20\times$ the compute budget of the final reported runs.

   Guidelines:

   - The answer [N/A] 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: The work is theoretical analysis of self-supervised learning dynamics plus controlled empirical evaluations on standard public benchmarks (CIFAR-10, ImageNet-1K); it involves no human subjects, no scraped data, and no high-risk model release, and the authors have reviewed and conform to the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: This is foundational self-supervised learning theory (a nonlinear-CCA characterization of asymmetry-based JEPA dynamics and a corresponding non-contrastive objective) evaluated on standard SSL benchmarks; there is no direct path from this work to a specific deployed application with identifiable positive or negative societal impacts beyond those generic to representation learning.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any pretrained model checkpoints, generative models, or scraped datasets that would carry a high risk of misuse; the only artifacts involved are linear-probe-evaluated SSL encoders trained on standard public benchmarks.

Guidelines:

- The answer [N/A] 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: All existing assets are credited to their creators via citation, and their licenses are stated where each asset is introduced: ImageNet-1K [Russakovsky et al., 2014] is documented in Appendix G.1 as the ILSVRC2012 split distributed under the ImageNet Terms of Access (a custom non-commercial research license); CIFAR-10 [Krizhevsky and Hinton, 2009], used only for the qualitative diagnostic, is noted in Appendix G.1 as released by the University of Toronto with no explicit license, freely available for research use. The VICReg/BYOL augmentation stack and linear-probe protocol [Bardes et al., 2022, Grill et al., 2020] and the LARS optimizer [You et al., 2017] are reimplemented from their cited descriptions.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not release any new datasets, models, or code artifacts at submission; PEIRA is described algorithmically (Section 5) and through its configuration table (Table G.1) but no asset is bundled with the submission.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects; all experiments are conducted on standard public image and navigation benchmarks.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve research with human subjects, so no IRB review is required.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology; PEIRA is a non-contrastive self-supervised objective for vision encoders analysed via nonlinear CCA, with no LLM involvement in the theory or in the IN1K, CIFAR-10.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.