Learning Predictive Embeddings through Inter-View Regressor Alignment

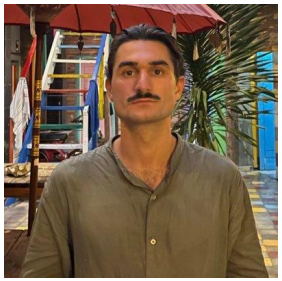
CILVR seminar

Center for Data Science, NYU

Michael Arbel

April 21, 2026

Joint work with



Basile Terver

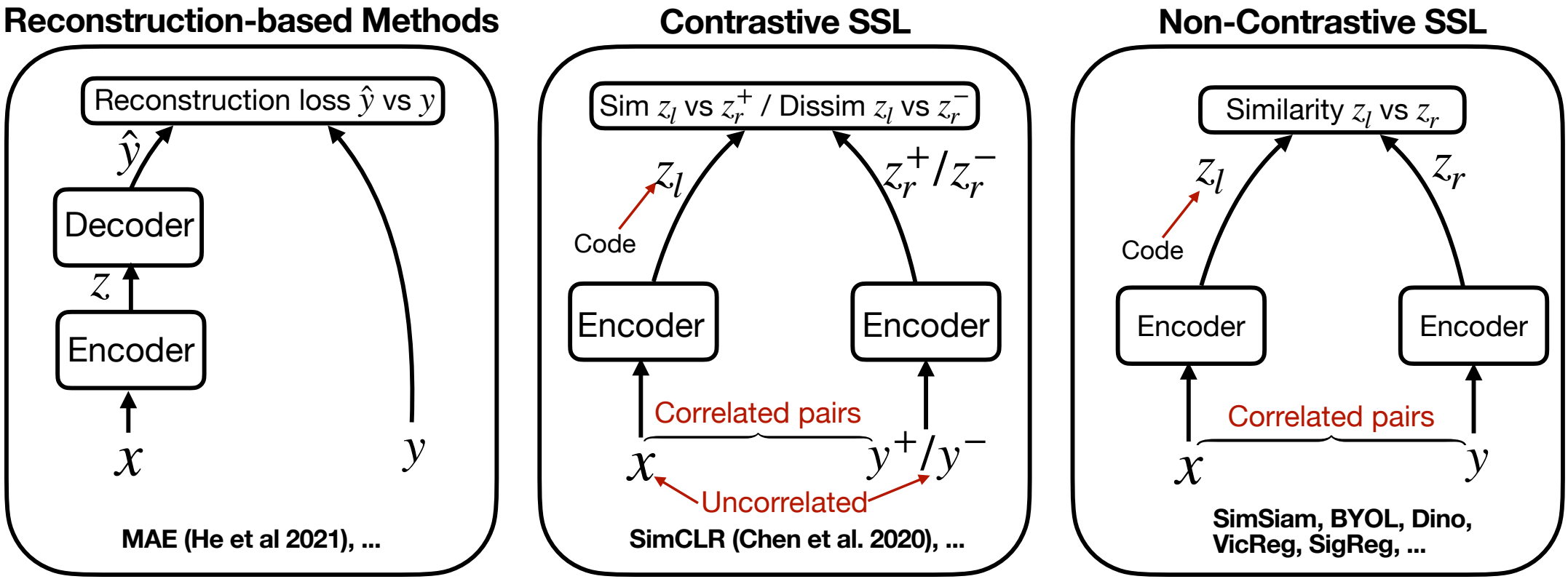
AMILabs/ENS Paris

Jean Ponce

NYU/ENS Paris

Self-supervised learning

Learn data representation without supervision by comparing altered/partial views of the same input

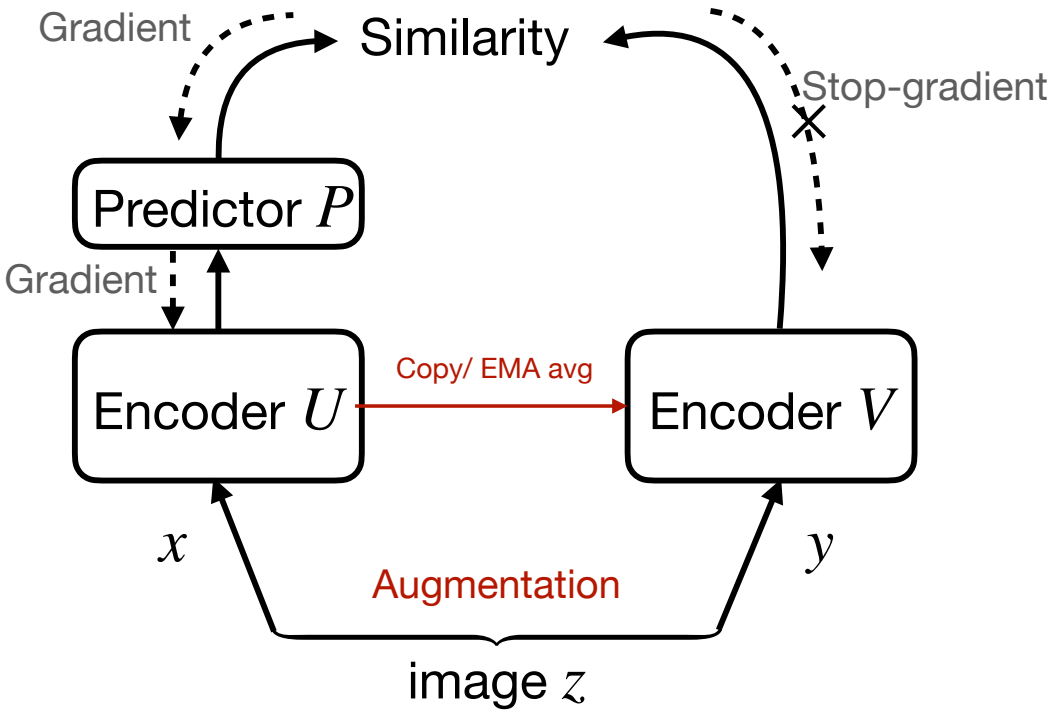


SimSiam-like dynamics for SSL

What is known

SimSiam-like SSL methods: A puzzle

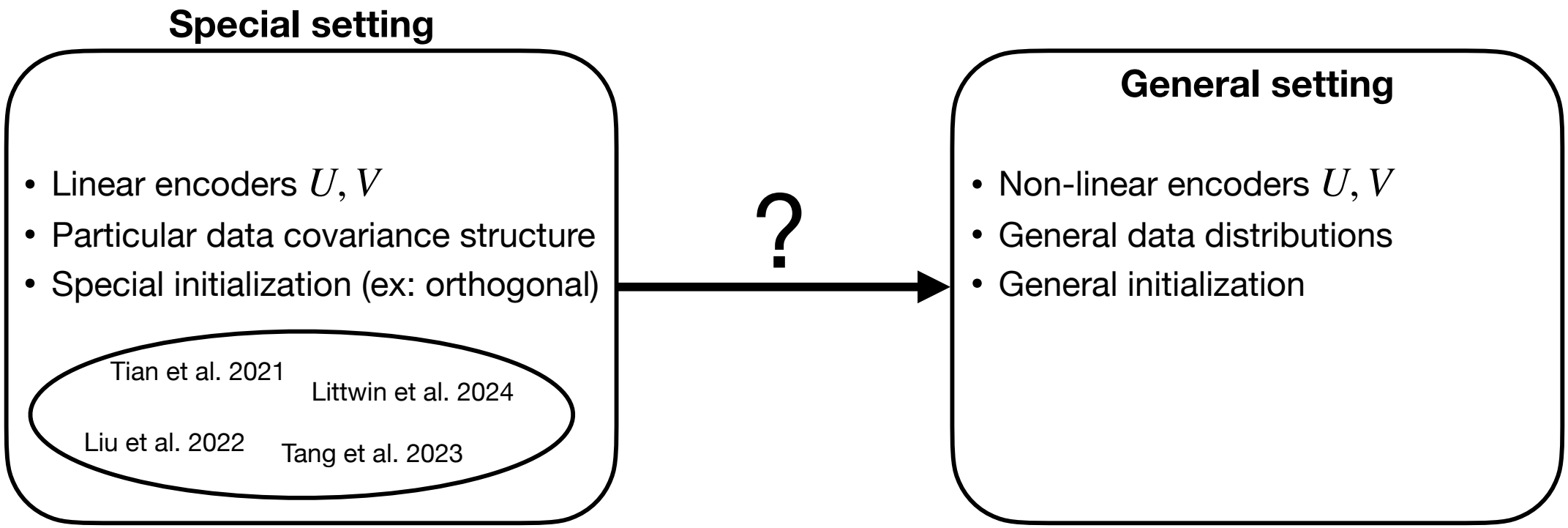
- Strong empirical performance
- Heuristic asymmetry (Stop-gradient, EMA)
- Avoid collapse in some cases



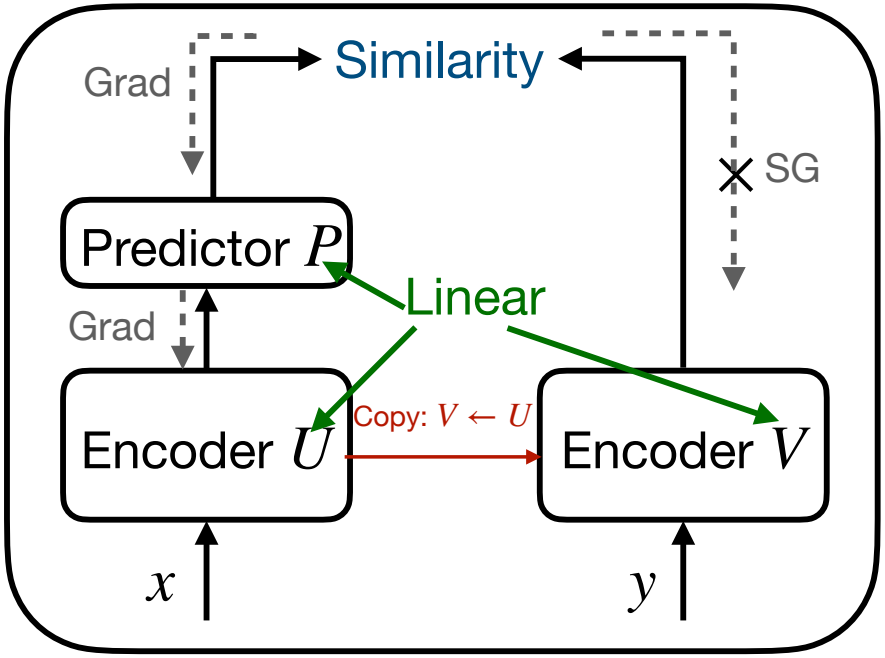
Goal: Understanding:

- What is learned?
- When does collapse occur?
- How to reliably prevent collapse?

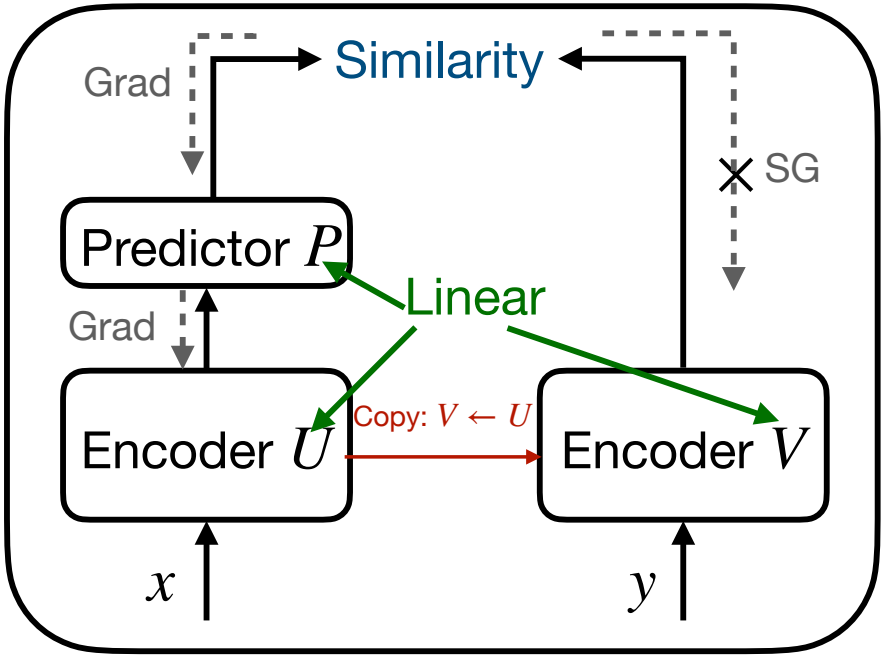
SimSiam-like SSL: Existing theory is informative but restricted



SimSiam-like SSL: Linear encoder and predictor



SimSiam-like SSL: Linear encoder and predictor



Setting considered in Tian et al. 2021, etc

Similarity

$$\mathscr{L}(P,U,V) = \frac{1}{2}\mathbb{E}[\|PUx - Vy\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\|U\|^2$$

Continuous-time dynamics

$$\begin{cases} \dot{P} = -\partial_P \mathscr{L}(P,U,V) \\ \dot{U} = -\partial_U \mathscr{L}(P,U,V) \\ V = U \end{cases}$$

Not the gradient flow of any objective
(Ponce et al. 2025)

Linear encoder and optimal predictor (Tang et al. 2023)

$$\mathscr{L}(P,U,V) = \frac{1}{2}\mathbb{E}[\|PUx - Vy\|^2] + \cancel{\frac{\lambda}{2}\|P\|^2} + \cancel{\frac{\lambda}{2}\|U\|^2}$$

$$\begin{cases} \dot{P} = -\partial_P \mathscr{L}(P,U,V) \\ \dot{U} = -\partial_U \mathscr{L}(P,U,V) \\ V = U \end{cases}$$

Optimal predictor

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P,U,V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star,U,V) \\ V = U \end{cases}$$

Linear encoder and optimal predictor (Tang et al. 2023)

$$\mathscr{L}(P,U,V)=\frac{1}{2}\mathbb{E}[\|PUx-Vy\|^2]+\frac{\lambda}{2}\|P\|^2+\frac{\lambda}{2}\|U\|^2$$

$$\begin{cases}\dot{P}=-\partial_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P,U,V)\\ V=U\end{cases}$$

Optimal predictor

$$\begin{cases}P^\star=\arg\min_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P^\star,U,V)\\ V=U\end{cases}$$

Theorem (Adapted)

Assume:

- Orthogonal initialization $U_0U_0^\top=I\in\mathbb{R}^{k\times k}$
- Matrix $\Sigma_{xx}=\mathbb{E}[xx^\top]=\alpha I$
- Matrix $\Sigma_{yx}=\mathbb{E}[yx^\top]$ is symmetric

Then:

- U stays orthogonal over time
- $\mathscr{F}(U)=-\frac{1}{2}Tr((U\Sigma_{yx}U^\top)^2)$ decreases over time
- Rows of an equilibrium $U^\star$ span an eigen-subspace of Σ_{yx} of at most k dimensions

Linear encoder and optimal predictor (Tang et al. 2023)

$$\mathscr{L}(P,U,V)=\frac{1}{2}\mathbb{E}[\|PUx-Vy\|^2]+\frac{\lambda}{2}\|P\|^2+\frac{\lambda}{2}\|U\|^2$$

$$\begin{cases}\dot{P}=-\partial_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P,U,V)\\ V=U\end{cases}$$

Optimal predictor

$$\begin{cases}P^\star=\arg\min_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P^\star,U,V)\\ V=U\end{cases}$$

Theorem (Adapted)

Assume:

- Orthogonal initialization $U_0U_0^\top=I\in\mathbb{R}^{k\times k}$
- Matrix $\Sigma_{xx}=\mathbb{E}[xx^\top]=\alpha I$
- Matrix $\Sigma_{yx}=\mathbb{E}[yx^\top]$ is symmetric

Then:

- U stays orthogonal over time
- $\mathscr{F}(U)=-\frac{1}{2}Tr((U\Sigma_{yx}U^\top)^2)$ decreases over time
- Rows of an equilibrium $U^\star$ span an eigen-subspace of Σ_{yx} of at most k dimensions

Q: Convergence to an equilibrium $U^\star$? Which eigen-subspace of Σ_{yx} is selected?

A: Unknown

Linear encoder and optimal predictor (Tang et al. 2023)

$$\mathscr{L}(P,U,V)=\frac{1}{2}\mathbb{E}[\|PUx-Vy\|^2]+\frac{\lambda}{2}\|P\|^2+\frac{\lambda}{2}\|U\|^2$$

$$\begin{cases}\dot{P}=-\partial_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P,U,V)\\ V=U\end{cases}$$

Optimal predictor

$$\begin{cases}P^\star=\arg\min_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P^\star,U,V)\\ V=U\end{cases}$$

Theorem (Adapted)

Assume:

- Orthogonal initialization $U_0U_0^\top=I\in\mathbb{R}^{k\times k}$
- Matrix $\Sigma_{xx}=\mathbb{E}[xx^\top]=\alpha I$
- Matrix $\Sigma_{yx}=\mathbb{E}[yx^\top]$ is symmetric

Then:

- U stays orthogonal over time
- $\mathscr{F}(U)=-\frac{1}{2}Tr((U\Sigma_{yx}U^\top)^2)$ decreases over time
- Rows of an equilibrium $U^\star$ span an eigen-subspace of Σ_{yx} of at most k dimensions

Q: General initialization? General data? (Still with linear encoders)

A: Complex dynamical system, hard to analyze (Ponce et al. 2025)

Linear encoder and optimal predictor (Tang et al. 2023)

$$\mathscr{L}(P,U,V)=\frac{1}{2}\mathbb{E}[\|PUx-Vy\|^2]+\frac{\lambda}{2}\|P\|^2+\frac{\lambda}{2}\|U\|^2$$

$$\begin{cases}\dot{P}=-\partial_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P,U,V)\\ V=U\end{cases}$$

Optimal predictor

$$\begin{cases}P^\star=\arg\min_P\mathscr{L}(P,U,V)\\ \dot{U}=-\partial_U\mathscr{L}(P^\star,U,V)\\ V=U\end{cases}$$

Theorem (Adapted)

Assume:

- Orthogonal initialization $U_0U_0^\top=I\in\mathbb{R}^{k\times k}$
- Matrix $\Sigma_{xx}=\mathbb{E}[xx^\top]=\alpha I$
- Matrix $\Sigma_{yx}=\mathbb{E}[yx^\top]$ is symmetric

Then:

- U stays orthogonal over time
- $\mathscr{F}(U)=-\frac{1}{2}Tr((U\Sigma_{yx}U^\top)^2)$ decreases over time
- Rows of an equilibrium $U^\star$ span an eigen-subspace of Σ_{yx} of at most k dimensions

Q: What about non-linear encoders?

A: Seems even harder, but .... not so if done in function spaces

Non-linear SimSiam-like dynamics:

New convergence results

A SimSiam-like dynamics in function space

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

A SimSiam-like dynamics in function space

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

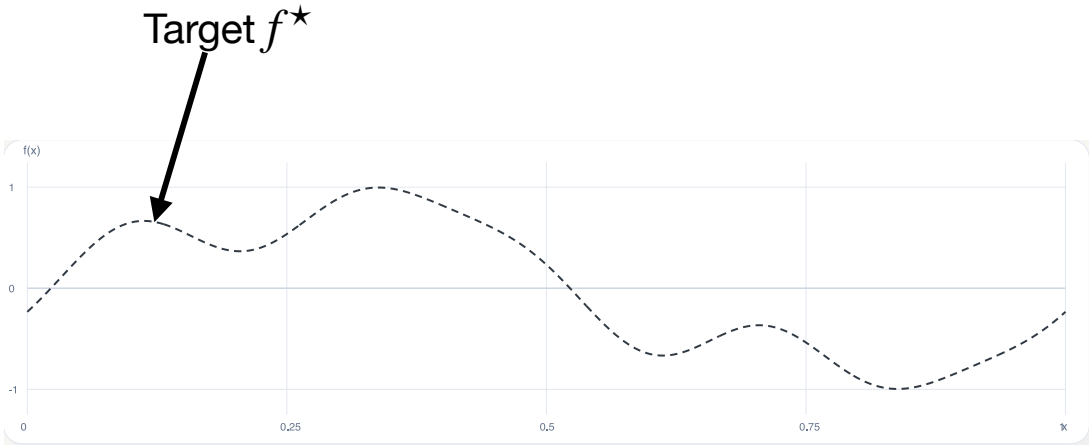
- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

Why care about function spaces?

Aren't we training a neural network after all?

Why care about function spaces?

Neural Networks are function approximators

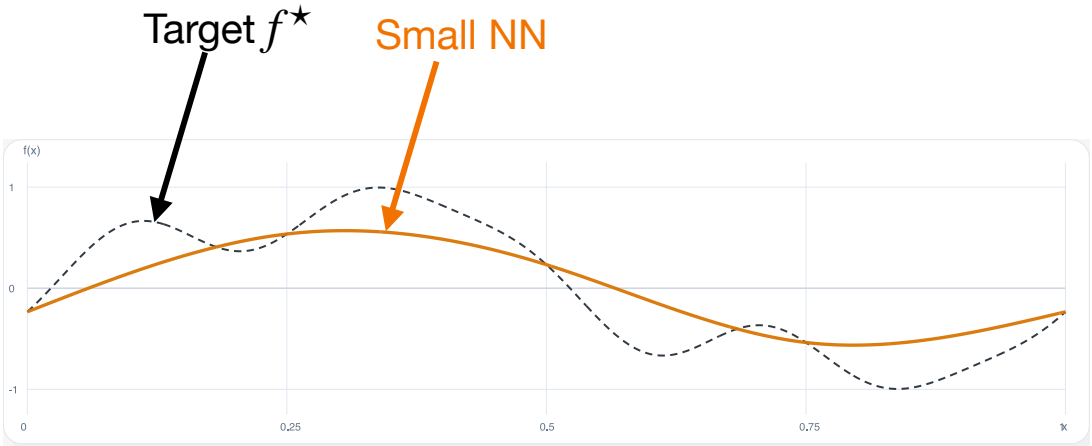


Regression: $f^\star = \arg\min_f \mathbb{E}\left[\|y - f(x)\|^2\right]$

Target: $f^\star(x) = \mathbb{E}\left[y \,|\, x\right]$

Why care about function spaces?

Neural Networks are function approximators

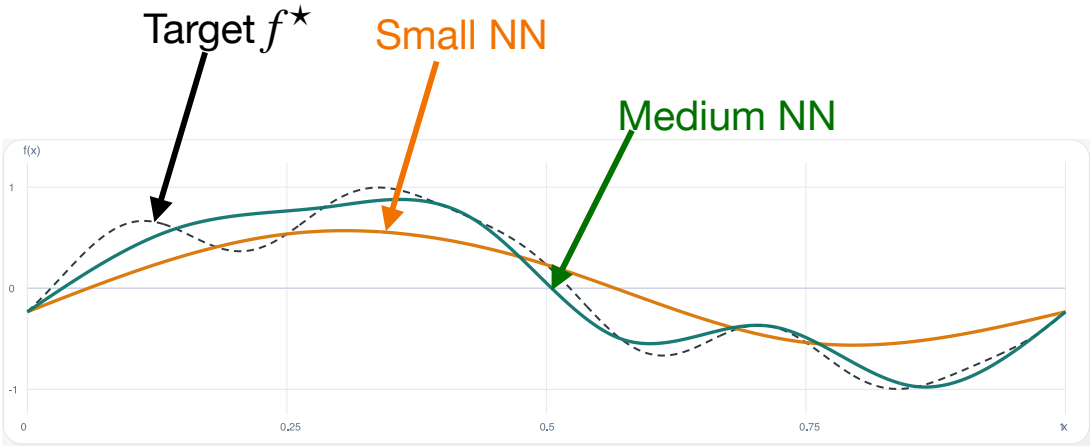


Regression: $f^\star = \arg\min_f \mathbb{E}\left[\|y - f(x)\|^2\right]$

Target: $f^\star(x) = \mathbb{E}\left[y \,|\, x\right]$

Why care about function spaces?

Neural Networks are function approximators

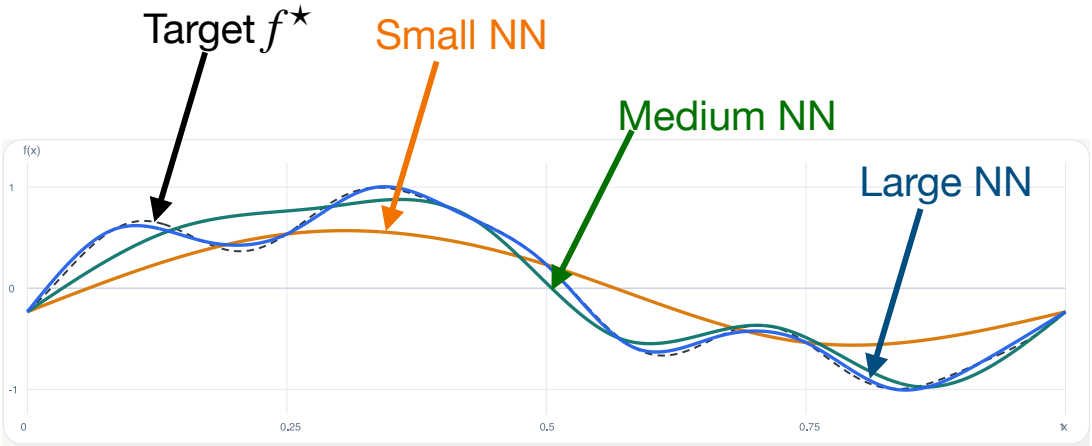


Regression: $f^\star = \arg\min_f \mathbb{E}\left[\|y - f(x)\|^2\right]$

Target: $f^\star(x) = \mathbb{E}\left[y \,|\, x\right]$

Why care about function spaces?

Neural Networks are function approximators

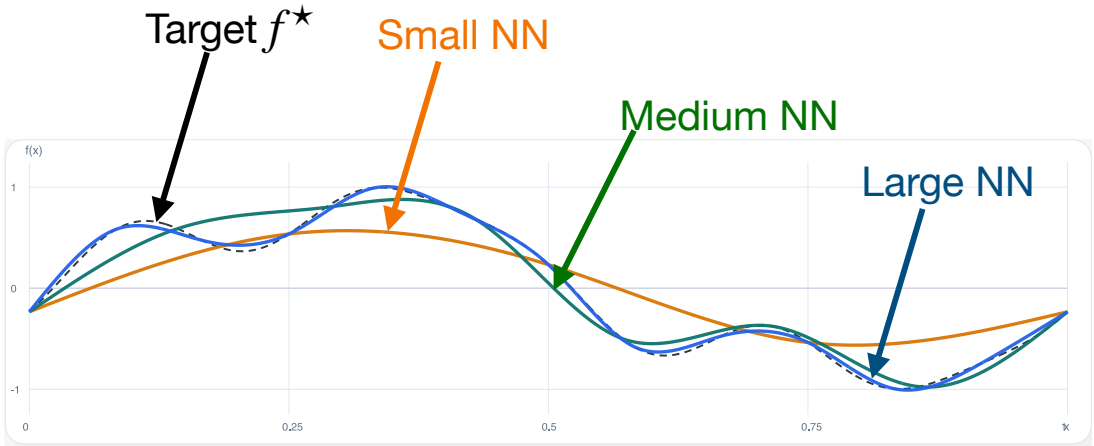


Regression: $f^\star = \arg\min_f \mathbb{E}\left[\|y - f(x)\|^2\right]$

Target: $f^\star(x) = \mathbb{E}\left[y \,|\, x\right]$

Why care about function spaces?

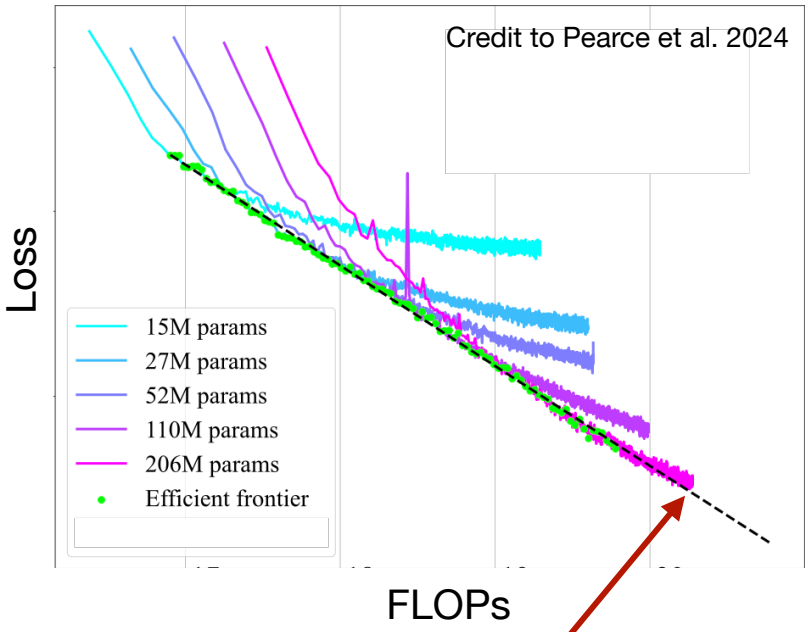
Neural Networks are function approximators



Regression: $f^\star = \arg\min_f \mathbb{E}\left[\|y - f(x)\|^2\right]$

Target: $f^\star(x) = \mathbb{E}\left[y \,|\, x\right]$

Scale expressivity, data and compute



Approach a functional target

Function space analysis reveals the target (properties, stability, etc)

A SimSiam-like dynamics in function space

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

Why care about function spaces?

Function space analysis reveals the target (properties, stability, etc)

A SimSiam-like dynamics in function space: Convergence

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

Theorem (Informal): Assume x and y play symmetric roles and data regular enough.

Then: 1- For any initialization U_0, the dynamics converges to an equilibrium $U^\star$.

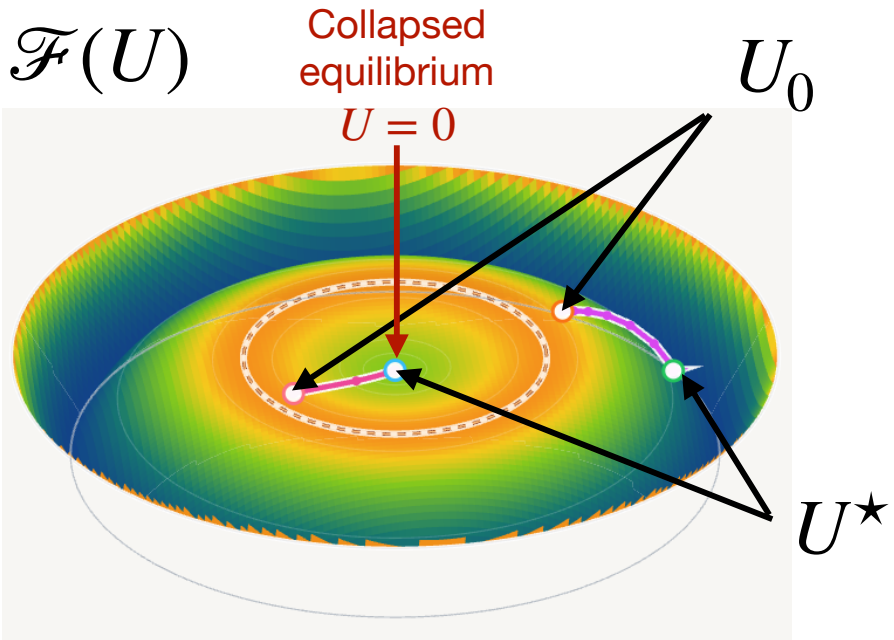
2- The following objective decreases in time (Lyapunov function):

$$\mathscr{F}(U) = -\frac{1}{2}Tr(\mathbb{E}[U(x)U(y)^\top]^2) + \frac{1}{3}Tr((\mathbb{E}[U(x)U(x)^\top] + \lambda I)^3)$$

Recovers the objective from Tang et al. 2023 under their assumptions

No assumption on the data covariances, or orthogonal initialization, as in Tang et al. 2023

A SimSiam-like dynamics in function space: Convergence



- No limit cycles, no divergence.
- Convergence to a critical point of the Lyapunov function.

Key technical ingredients

- **Gradient-like system:** The vector field aligns well with the gradient of the Lyapunov function
- **Lojasiewicz-Simon inequality:** Small gradients $\Longrightarrow$ Small gap to a critical value.
- **Trapping argument:** The trajectory is trapped near an accumulation point (Bolte et al. 2014)
- **Existence of accumulation points:** Non-trivial in infinite dimensions (Palais & Smale 1964)

A SimSiam-like dynamics in function space: Equilibria

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

What are the equilibria? Which ones are stable?

CCA to the rescue

Canonical Correlation Analysis

Linear Canonical Correlation Analysis

Correlation maximization

$$\max_{U,V \in \mathbb{R}^{k \times d}} \mathbb{E}[(Ux)^\top (Vy)]$$

s.t.

$$\mathbb{E}[(Ux)(Ux)^\top] = I$$

$$\mathbb{E}[(Vy)(Vy)^\top] = I$$

whitening constraints

Linear Canonical Correlation Analysis

Correlation maximization

$$\max_{U,V \in \mathbb{R}^{k \times d}} \mathbb{E}[(Ux)^\top (Vy)]$$

s.t. $\mathbb{E}[(Ux)(Ux)^\top] = I$

$\mathbb{E}[(Vy)(Vy)^\top] = I$

whitening constraints

Whitened cross-covariance

$$M = \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}}$$

$\mathbb{E}[yy^\top]$ $\mathbb{E}[yx^\top]$ $\mathbb{E}[xx^\top]$

Singular value decomposition

$$M = (v_1, \ldots, v_d) \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_d \end{pmatrix} \begin{pmatrix} u_1^\top \\ \vdots \\ u_d^\top \end{pmatrix}$$

CCA solution: top-k singular vector

$$U = \Sigma_{xx}^{-\frac{1}{2}} \begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix}, \quad V = \Sigma_{yy}^{-\frac{1}{2}} \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}$$

Linear Canonical Correlation Analysis

Correlation maximization

$$\max_{U,V \in \mathbb{R}^{k \times d}} \mathbb{E}[(Ux)^\top (Vy)]$$

s.t. $\mathbb{E}[(Ux)(Ux)^\top] = I$

$\mathbb{E}[(Vy)(Vy)^\top] = I$

whitening constraints

Recovers similar eigen-subspace as Tang et al. 2023 under their assumptions

(Σ_{yx} symmetric and $\Sigma_{xx} = \Sigma_{yy} = \alpha I$)

Whitened cross-covariance

$$M = \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}}$$

$\mathbb{E}[yy^\top]$ $\mathbb{E}[yx^\top]$ $\mathbb{E}[xx^\top]$

Singular value decomposition

$$M = (v_1, \ldots, v_d) \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_d \end{pmatrix} \begin{pmatrix} u_1^\top \\ \vdots \\ u_d^\top \end{pmatrix}$$

CCA solution: top-k singular vector

$$U = \Sigma_{xx}^{-\frac{1}{2}} \begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix}, \quad V = \Sigma_{yy}^{-\frac{1}{2}} \begin{pmatrix} v_1 \\ \vdots \\ v_k \end{pmatrix}$$

Non-Linear Canonical Correlation Analysis (Michaeli et al. 2015)

Correlation maximization

$$\max_{(U,V)\in\mathcal{U}\times\mathcal{V}} \mathbb{E}[U(x)^\top V(y)]$$

$L_2(\mathbb{P}_x)^k$ $L_2(\mathbb{P}_y)^k$

s.t. $\mathbb{E}[U(x)U(x)^\top] = I$

$\mathbb{E}[V(y)V(y)^\top] = I$

whitening constraints

Non-Linear Canonical Correlation Analysis (Michaeli et al. 2015)

Correlation maximization

$$\max_{(U,V)\in\mathcal{U}\times\mathcal{V}} \mathbb{E}[U(x)^\top V(y)]$$

$L_2(\mathbb{P}_x)^k$ $L_2(\mathbb{P}_y)^k$

s.t. $\mathbb{E}[U(x)U(x)^\top] = I$

$\mathbb{E}[V(y)V(y)^\top] = I$

whitening constraints

Non-linear CCA solution: top-k canonical functions

$$U(x) = \begin{pmatrix} u_1(x) \\ \vdots \\ u_k(x) \end{pmatrix}, \quad V(y) = \begin{pmatrix} v_1(y) \\ \vdots \\ v_k(y) \end{pmatrix}$$

Canonical correlation $1 = s_1 \geq s_2 \geq \dots$

$\mathbb{E}[u_i(x)v_i(y)] = s_i$ $\mathbb{E}[u_i(x)v_j(y)] = 0$

De-correlation (when $i \neq j$)

Orthonormal basis of $L_2(\mathbb{P}_y)$

Orthonormal basis of $L_2(\mathbb{P}_x)$

Non-Linear Canonical Correlation Analysis (Michaeli et al. 2015)

Correlation maximization

$$\max_{(U,V)\in\mathcal{U}\times\mathcal{V}} \mathbb{E}[U(x)^\top V(y)]$$

$L_2(\mathbb{P}_x)^k$ $L_2(\mathbb{P}_y)^k$

s.t. $\mathbb{E}[U(x)U(x)^\top] = I$

$\mathbb{E}[V(y)V(y)^\top] = I$

whitening constraints

Connection with VICReg, BarlowTwins (Balestriero & LeCun 2022)

What about SimSiam with nonlinear encoder ?

Non-linear CCA solution: top-k canonical functions

$$U(x) = \begin{pmatrix} u_1(x) \\ \vdots \\ u_k(x) \end{pmatrix}, \quad V(y) = \begin{pmatrix} v_1(y) \\ \vdots \\ v_k(y) \end{pmatrix}$$

Canonical correlation $1 = s_1 \geq s_2 \geq \dots$

$\mathbb{E}[u_i(x)v_i(y)] = s_i$ $\mathbb{E}[u_i(x)v_j(y)] = 0$

De-correlation (when $i \neq j$)

Orthonormal basis of $L_2(\mathbb{P}_y)$

Orthonormal basis of $L_2(\mathbb{P}_x)$

Non-linear SimSiam-like dynamics:

Equilibria and their stability

A SimSiam-like dynamics in function space: Equilibria

$$\mathscr{L}(P, U, V) = \frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{\lambda}{2}\|P\|^2 + \frac{\lambda}{2}\mathbb{E}[\|U(x)\|^2]$$

$$\begin{cases} P^\star = \arg\min_P \mathscr{L}(P, U, V) \\ \dot{U} = -\partial_U \mathscr{L}(P^\star, U, V) \\ V = U \end{cases}$$

- Optimal linear predictor (as in Tang et al. 2023)
- No parametric form for the encoder U
- Dynamics defined in function space $L_2(\mathbb{P}_x, \mathbb{R}^k)$

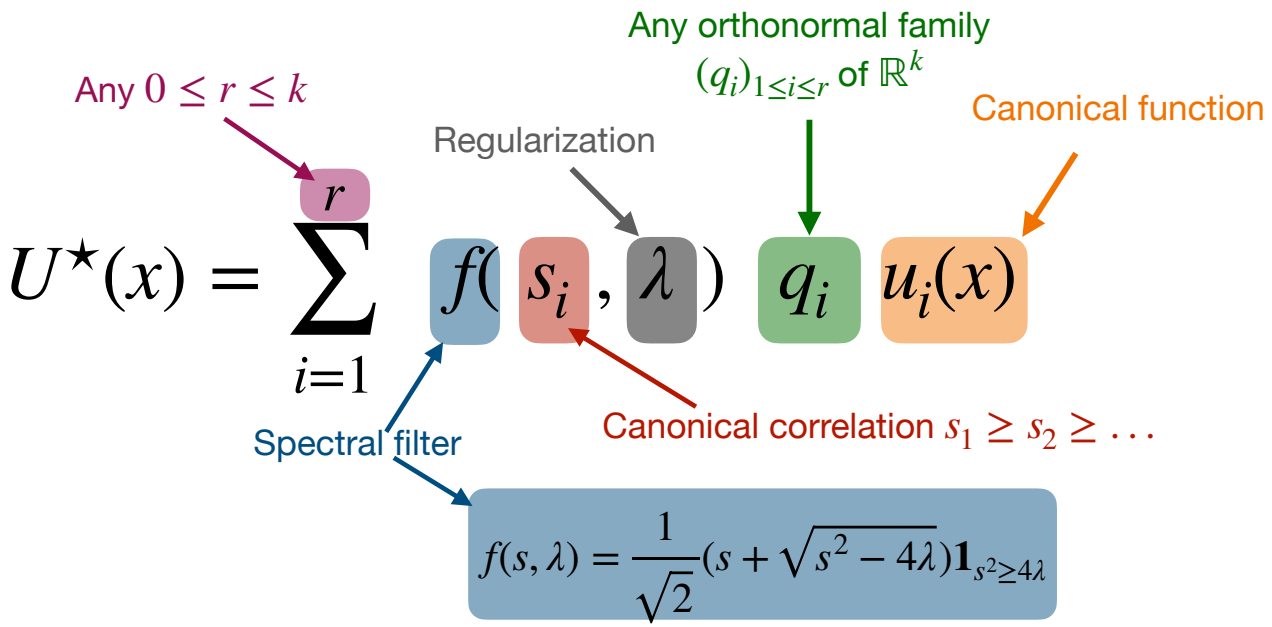
What are the equilibria? Which ones are stable?

Equilibria span CCA subspaces

SimSiam-like dynamics in function space: Stability of equilibria

Theorem (Informal):

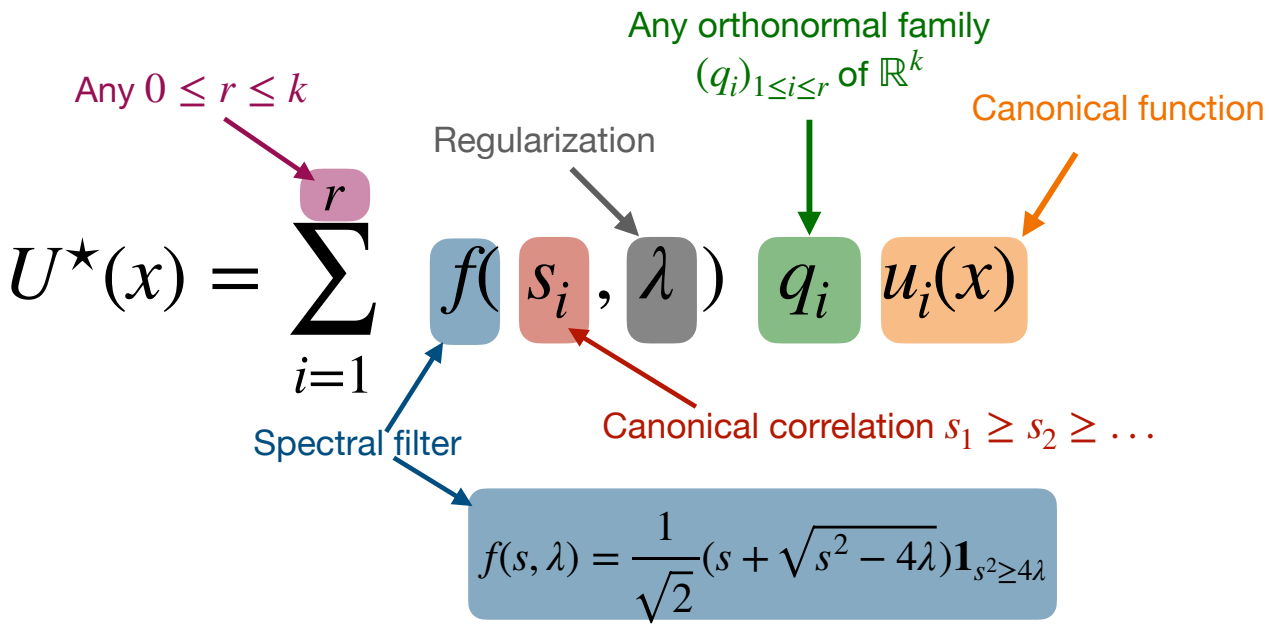
Equilibria span CCA subspaces. The only stable ones are of the form:



SimSiam-like dynamics in function space: Stability of equilibria

Theorem (Informal):

Equilibria span CCA subspaces. The only stable ones are of the form:

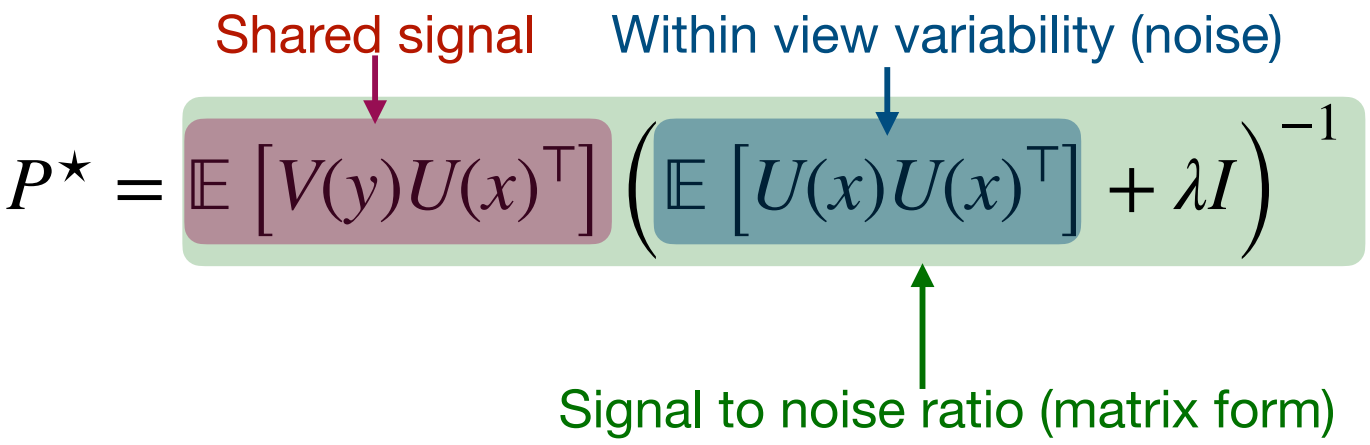


Collapse occurs when an admissible CCA direction ($f(s_i, \lambda) > 0$ for $i \le k$) is not selected ($r < i$)

Predictive Embeddings through Inter-View Regressor Alignment

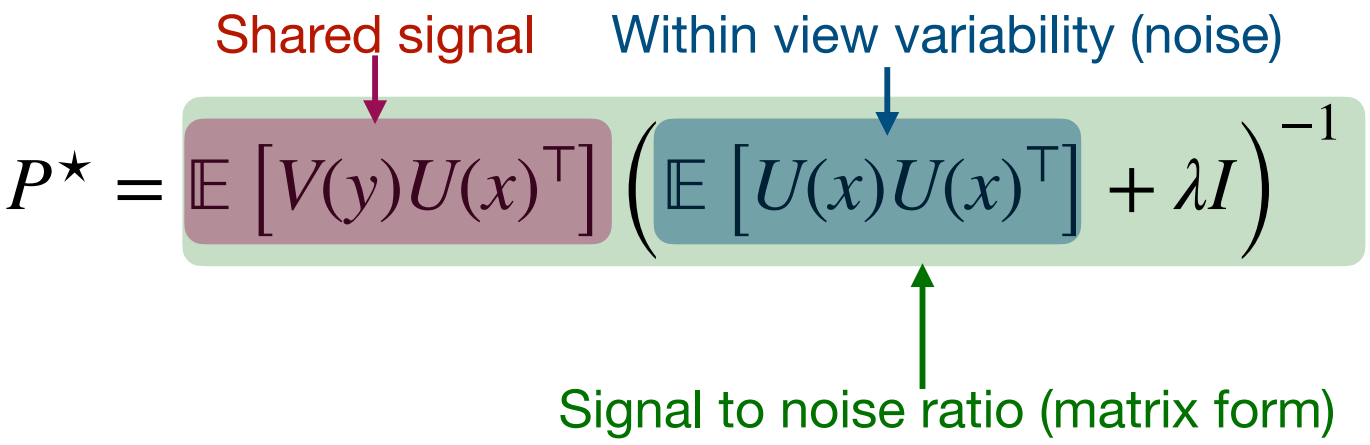
Intuition and Motivation: Trace maximization of a predictor

- Optimal predictor between two views:



Intuition and Motivation: Trace maximization of a predictor

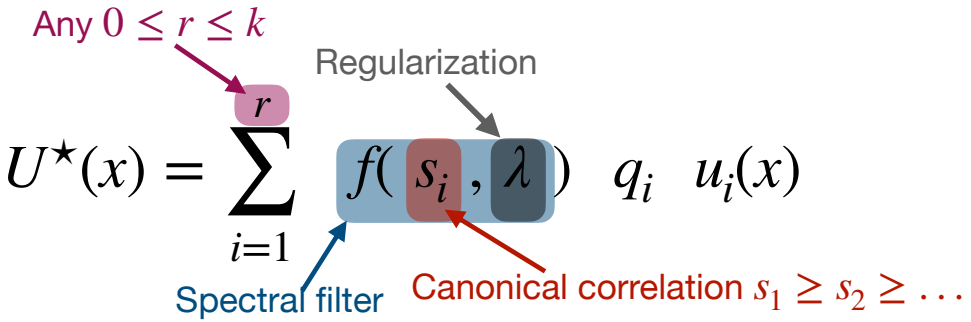
- Optimal predictor between two views:



$Tr(P^\star)$ acts as a scalar summary of the matrix signal to noise ratio

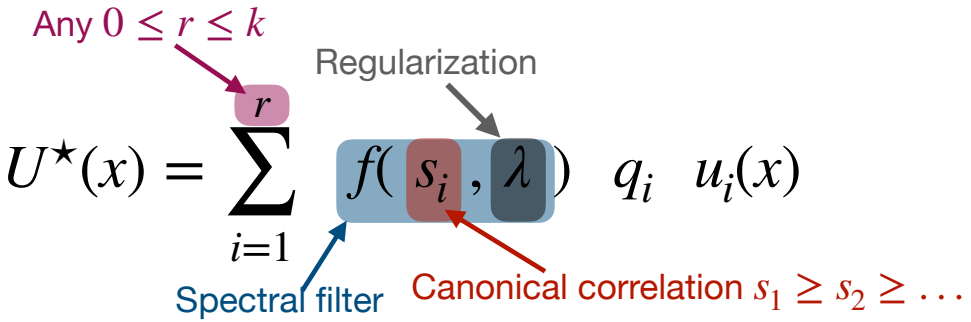
Intuition and Motivation: Trace maximization of a predictor

- Stable equilibria of the SimSiam-like dynamics in function space



Intuition and Motivation: Trace maximization of a predictor

- Stable equilibria of the SimSiam-like dynamics in function space



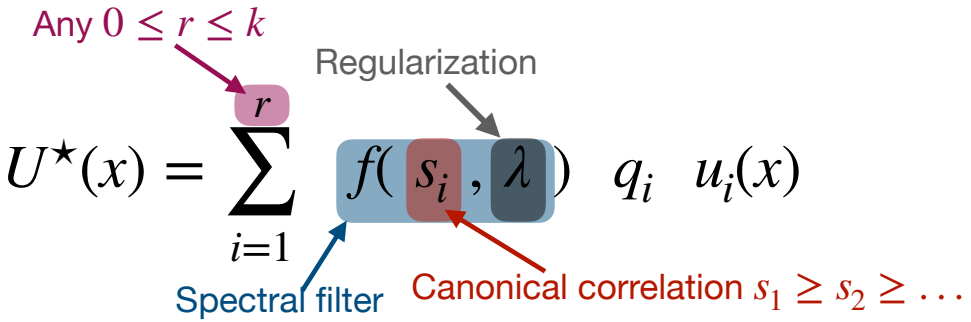
- Optimal predictor carries spectral information about an equilibrium:

maximized by non-collapsed equilibria

$$Tr(P^\star) = \sum_{i=1}^{r} s_i \left(1 - \frac{\lambda}{f(s_i, \lambda)^2 + \lambda}\right)$$

Intuition and Motivation: Trace maximization of a predictor

- Stable equilibria of the SimSiam-like dynamics in function space



- Optimal predictor carries spectral information about an equilibrium:

maximized by non-collapsed equilibria

$$Tr(P^\star) = \sum_{i=1}^{r} s_i \left(1 - \frac{\lambda}{f(s_i, \lambda)^2 + \lambda}\right)$$

- Trace maximization of the optimal predictor as a representation learning mechanism ?

Predictive Embeddings through Inter-View Regressor Alignment

- Find optimal symmetrized regressor for each encoders pairs (U, V)

$$\frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{1}{2}\mathbb{E}[\|PV(y) - U(x)\|^2] + \frac{\lambda}{2}\|P\|^2$$

$$P^{\star}_{U,V} = \arg\min_{P \in \mathbb{R}^{k \times k}} \mathscr{L}_{reg}(P; U, V)$$

Predictive Embeddings through Inter-View Regressor Alignment

- Find optimal symmetrized regressor for each encoders pairs (U, V)

$$\frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{1}{2}\mathbb{E}[\|PV(y) - U(x)\|^2] + \frac{\lambda}{2}\|P\|^2$$

$$P^{\star}_{U,V} = \arg\min_{P \in \mathbb{R}^{k \times k}} \mathscr{L}_{reg}(P; U, V)$$

- Learn encoders by minimizing regularized negative trace of optimal regressor:

Negative trace minimization to align representation

Regularization to fix the scale

$$\mathscr{L}_{PEIRA}(U, V) = -\frac{1}{2}Tr(P^{\star}_{U,V}) + \frac{\lambda}{2}\mathbb{E}[\|V(x)\|^2 + \|U(y)\|^2]$$

Predictive Embeddings through Inter-View Regressor Alignment

- Find optimal symmetrized regressor for each encoders pairs (U, V)

$$\frac{1}{2}\mathbb{E}[\|PU(x) - V(y)\|^2] + \frac{1}{2}\mathbb{E}[\|PV(y) - U(x)\|^2] + \frac{\lambda}{2}\|P\|^2$$

$$P^{\star}_{U,V} = \arg\min_{P \in \mathbb{R}^{k \times k}} \mathscr{L}_{reg}(P; U, V)$$

- Learn encoders by minimizing regularized negative trace of optimal regressor:

Negative trace minimization to align representation

Regularization to fix the scale

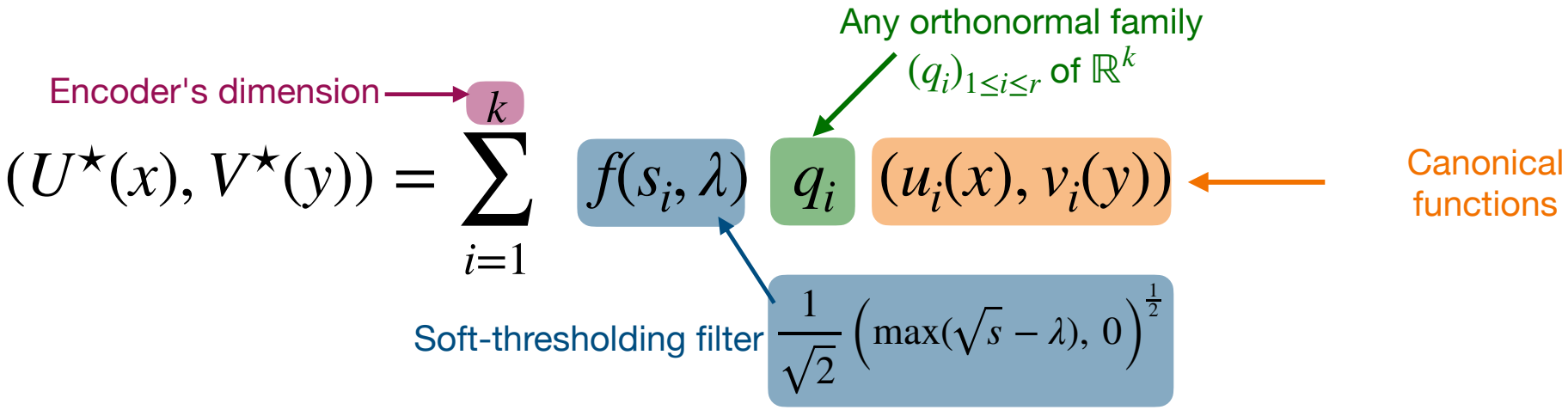
$$\mathscr{L}_{PEIRA}(U, V) = -\frac{1}{2}Tr(P^{\star}_{U,V}) + \frac{\lambda}{2}\mathbb{E}[\|V(x)\|^2 + \|U(y)\|^2]$$

What does it learn? Does it avoid collapse? Practical algorithms? Does it work?

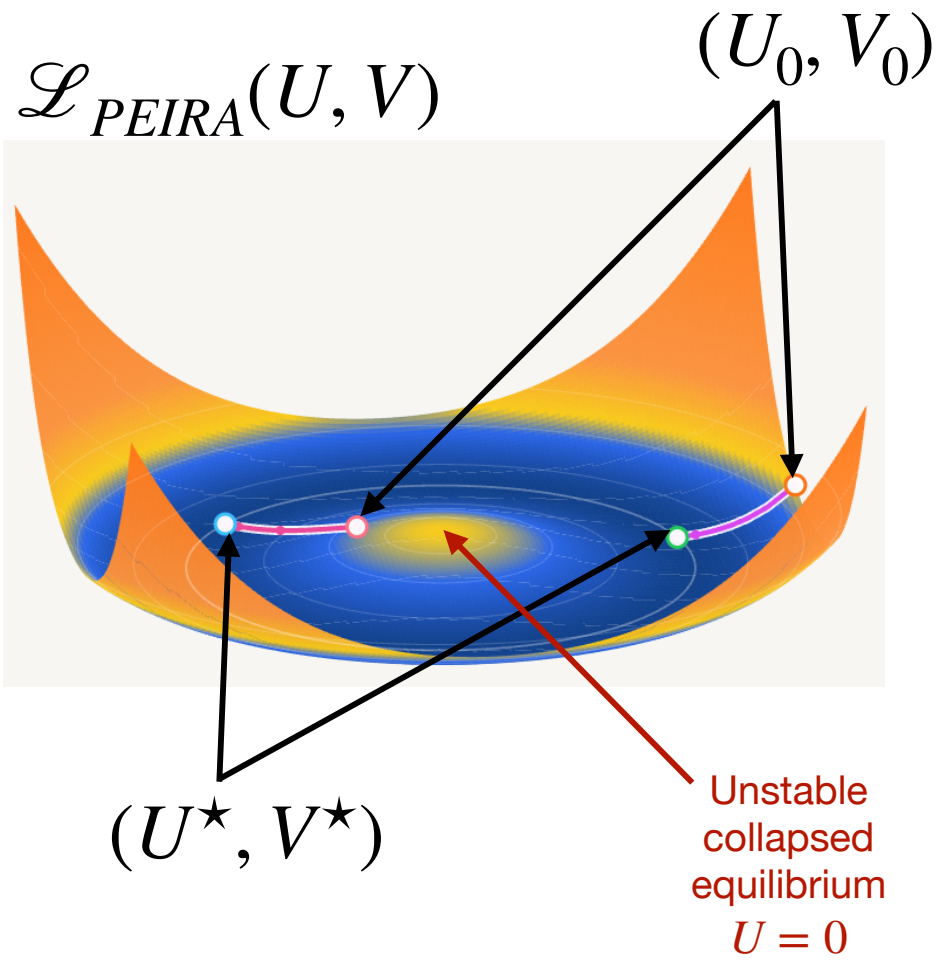
PEIRA objective: Global maximizers

Theorem (Informal): Assume the data distribution is regular enough (existence of non-linear CCA).

Then: 1- Global minimizers of $\mathscr{L}_{PEIRA}$ are exactly:



PEIRA objective: Stability and non-collapse

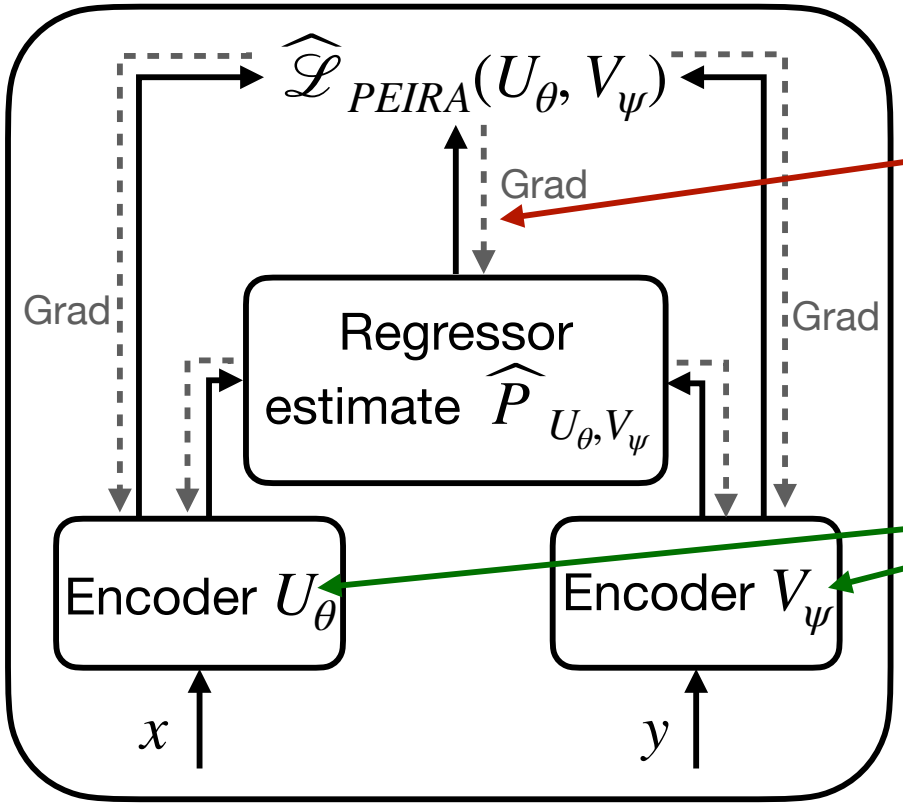


Theorem (Informal): Under the same assumptions as before:

1- Only global minimizers are stable.

2- All other critical points are unstable.

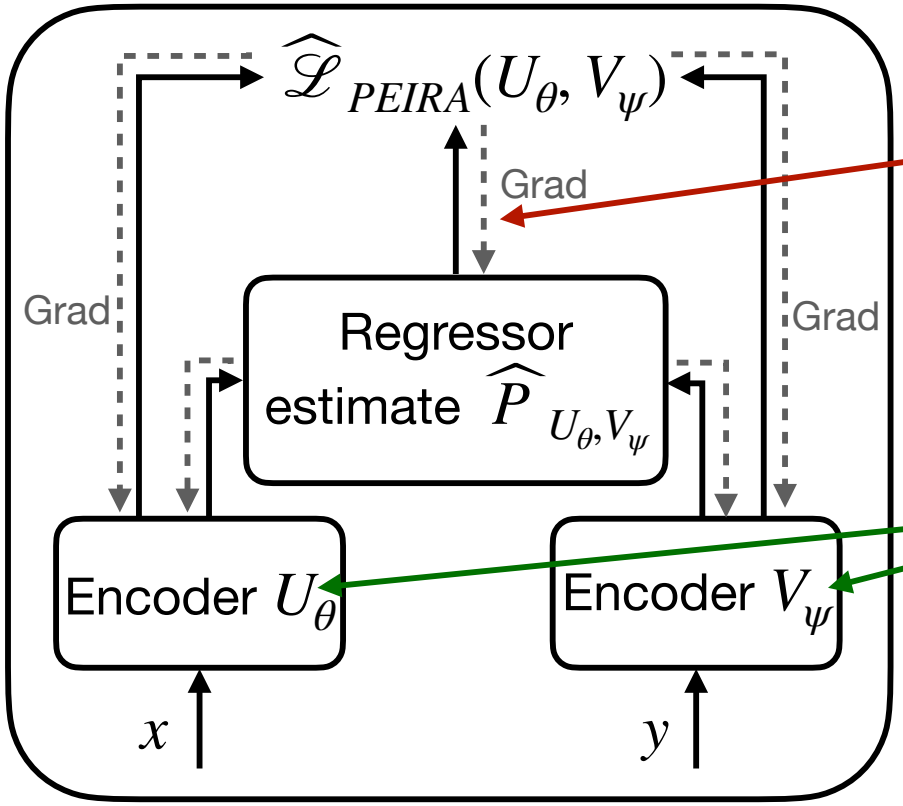
Optimizing PEIRA objective: A "bad" Algorithm



Automatic differentiation
through the regressor

Parametrization (ex: Neural Networks)
with optional **parameter sharing**

Optimizing PEIRA objective: A "bad" Algorithm

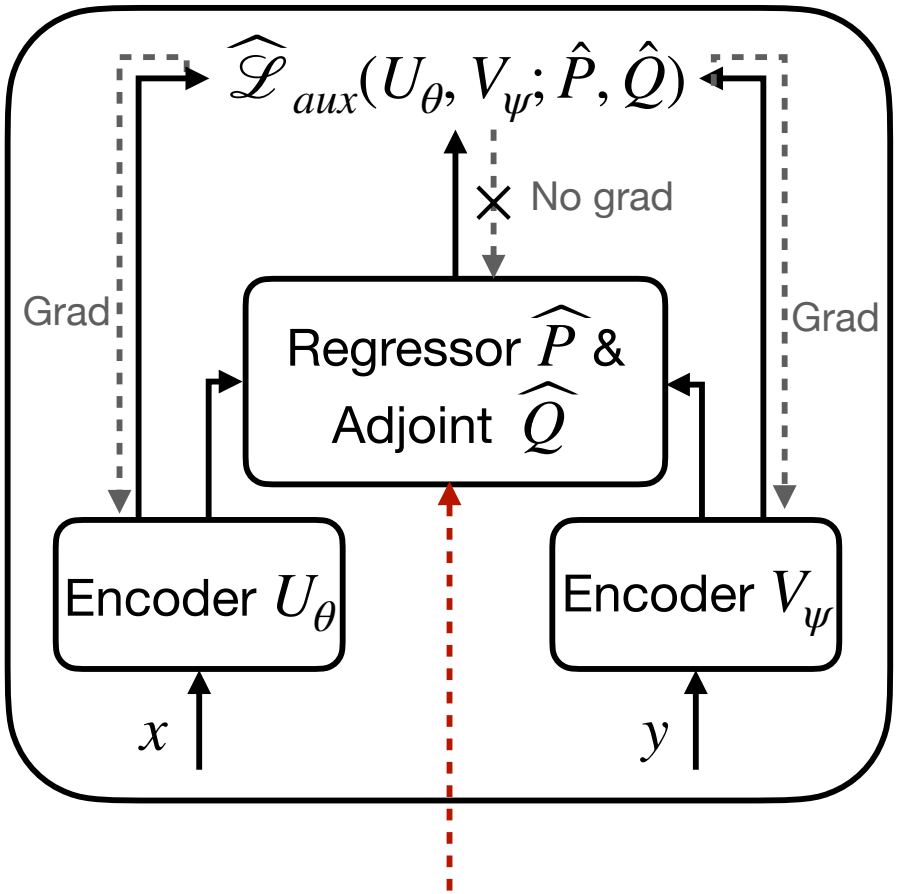


Automatic differentiation
through the regressor

Bad idea!
Noisy gradient

Parametrization (ex: Neural Networks)
with optional **parameter sharing**

Optimizing PEIRA objective: A better Algorithm



EMA for updating covariances + Exact solve

Proposition:

$\mathscr{L}_{PEIRA}(U, V)$ has the same gradient as that of an auxiliary loss $\mathscr{L}_{aux}(U, V; P, Q)$ if:

- P is the optimal regressor
- Q satisfies

$$Q = \left(\mathbb{E}\left[U(x)U(x)^\top + V(y)V(y)^\top\right] + \lambda I\right)^{-1}$$

The auxiliary loss:

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

How does PEIRA do in practice?

Tasks

SSL for natural images

Pre-training:

- Cifar10
- Imagenet

Downstream task:

- Classification Cifar10
- Classification Imagenet1k

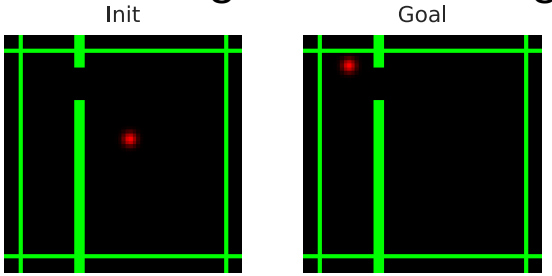
SSL for world modeling

Pre-training: Action conditioned video

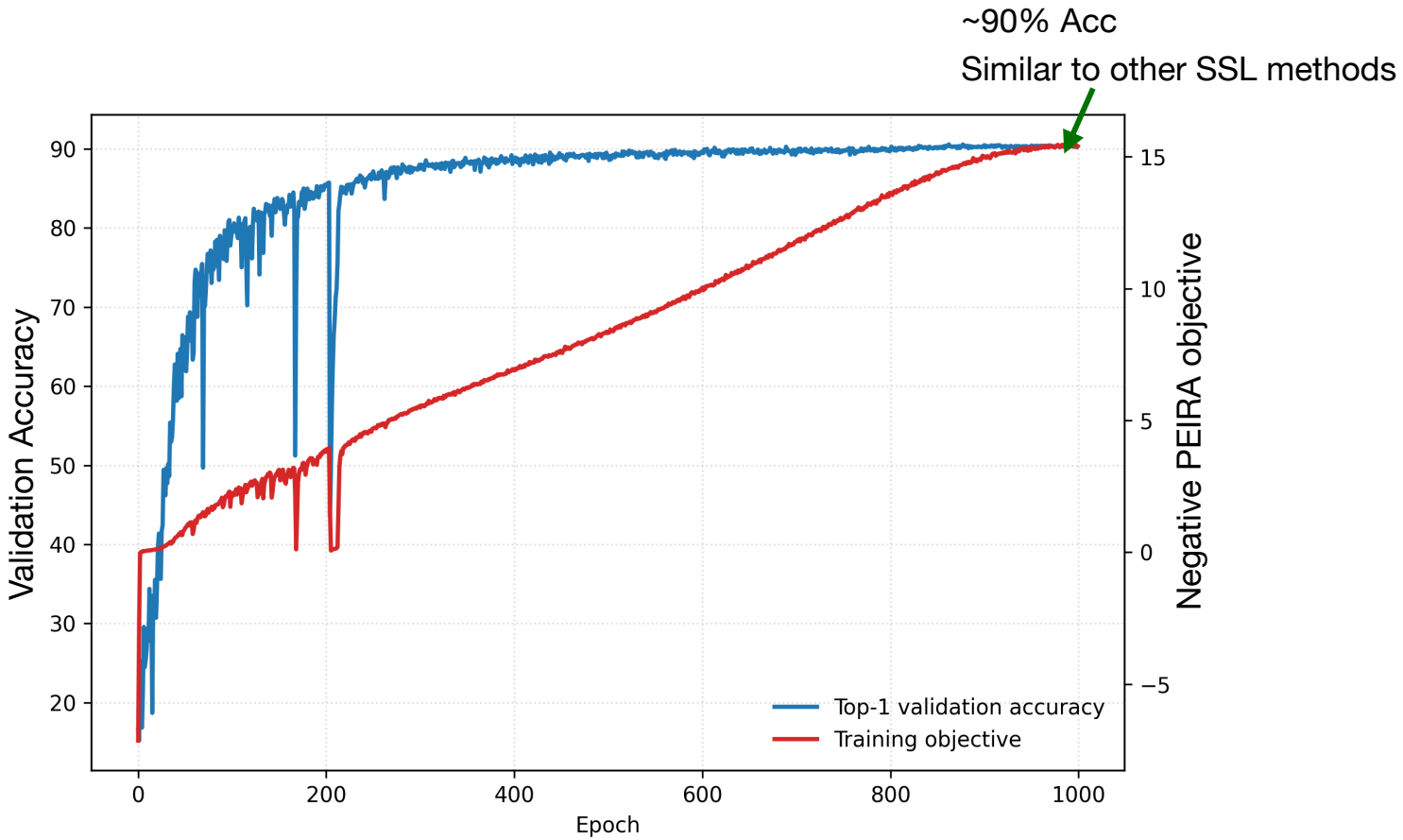
- Two rooms environment

Downstream task:

- Planning from initial to goal state

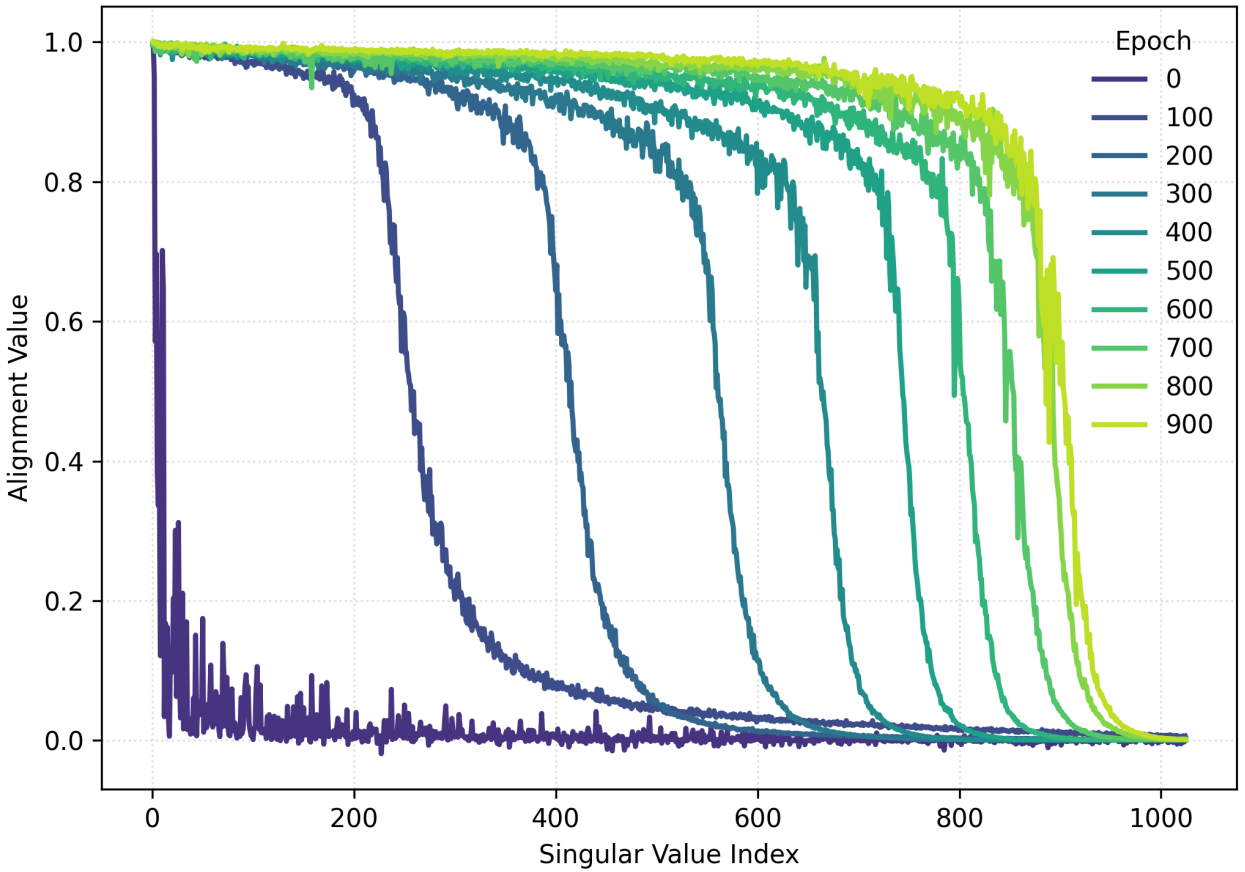


SSL for images: Cifar10



Strong correlation between Val acc and PEIRA objective

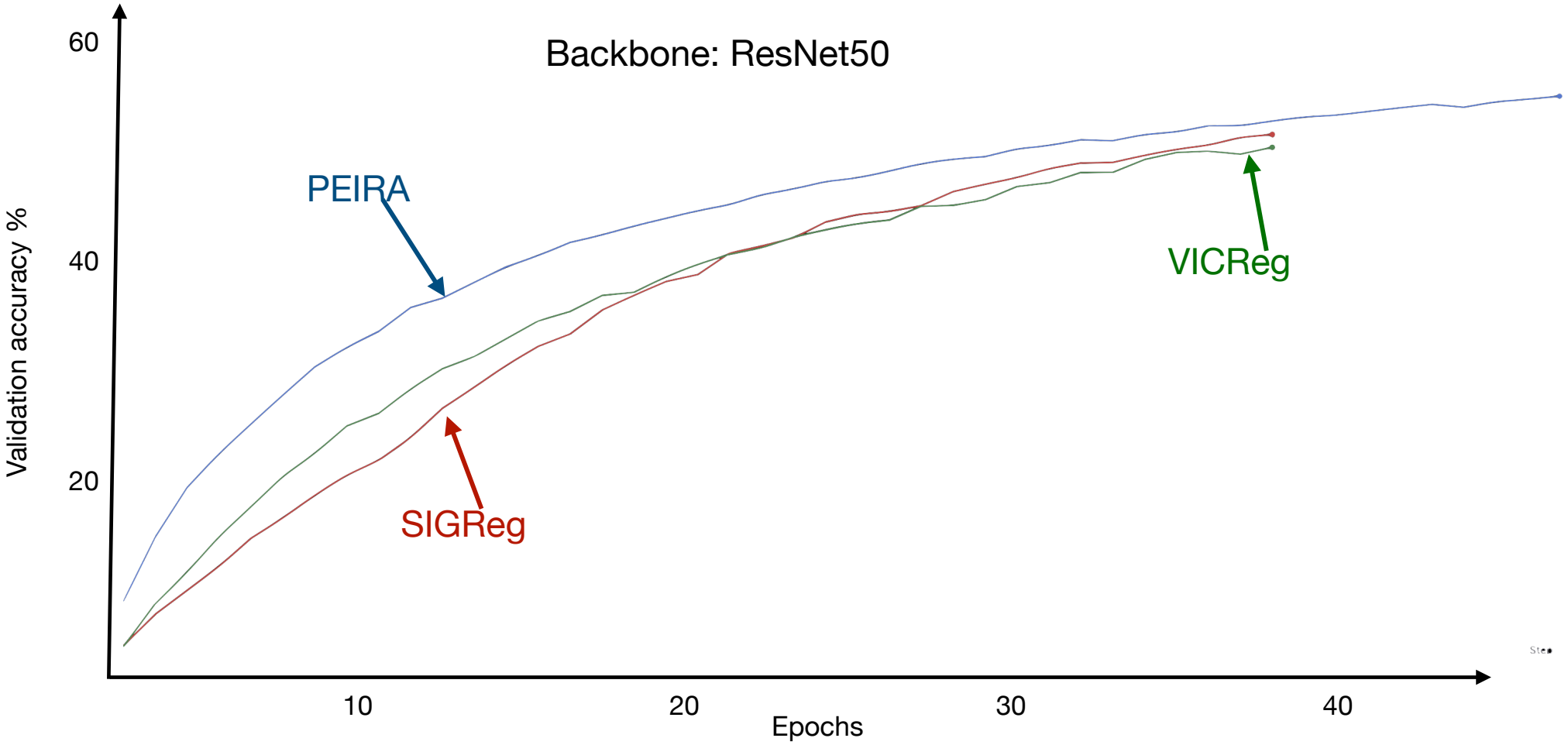
SSL for images: Cifar10



Gradual alignment between regressor and cross-covariance

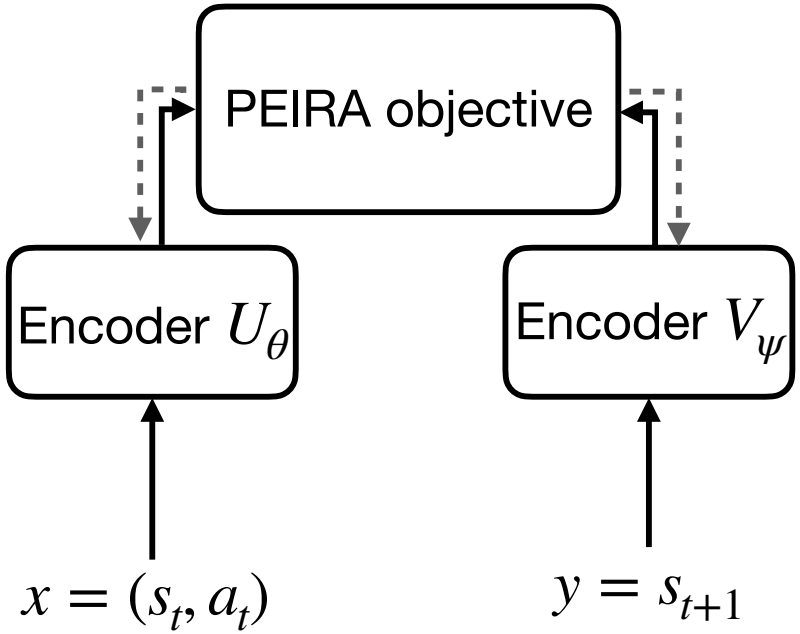
Similar to Tian et al. 2022

SSL for images: Preliminary results on ImageNet1k

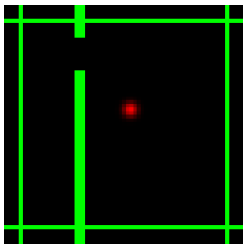


Main hyper-parameter: regularization λ

Self-supervised World models: Two rooms



- State s_t: Image

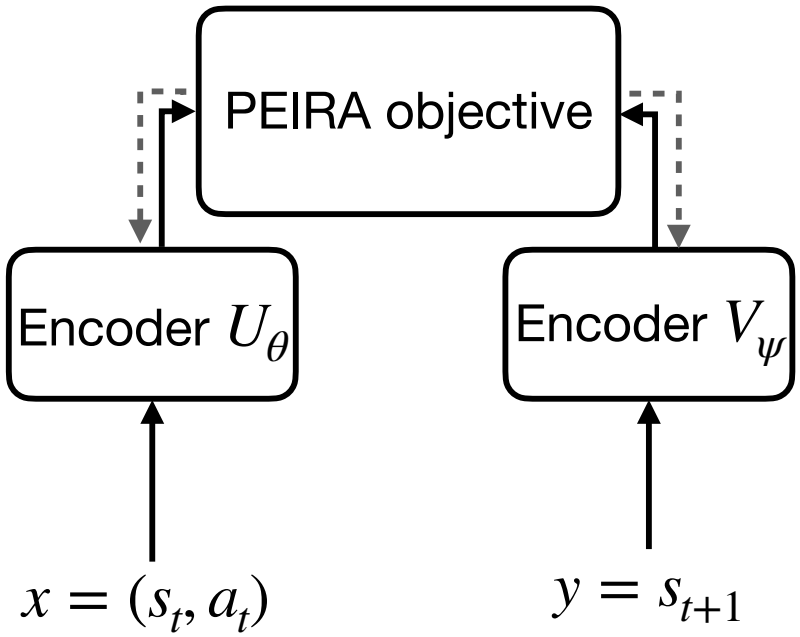


- Action a_t: Up, down, left, right
- State encoder V_ψ: Impala
- State action encoder $U_\theta(s, a) = RNN_\phi(V_\psi(s), a)$

Parameter sharing

$$\theta = (\phi, \psi)$$

Self-supervised World models: Two rooms



Setup: Joint pre-training of predictor and encoder

Comparison:

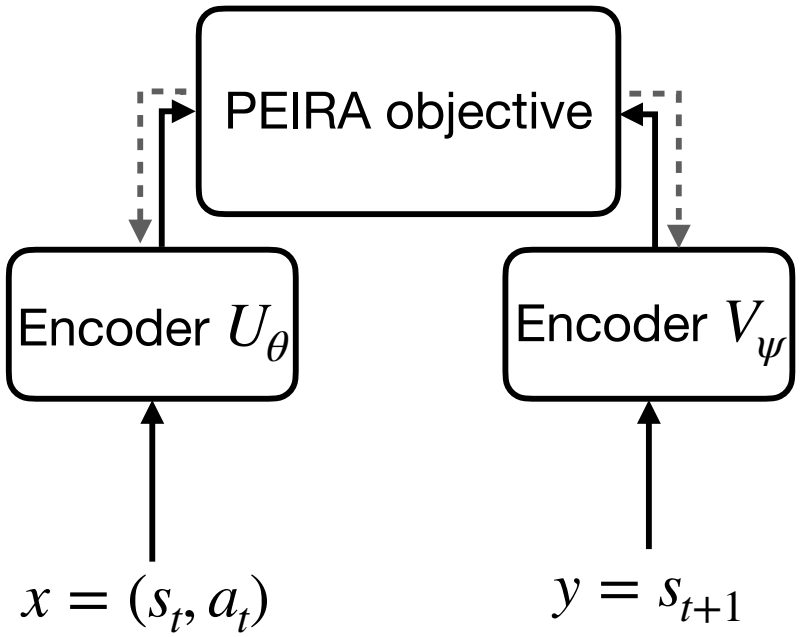
- The objective from PLDM (Sobal et al. 2025)

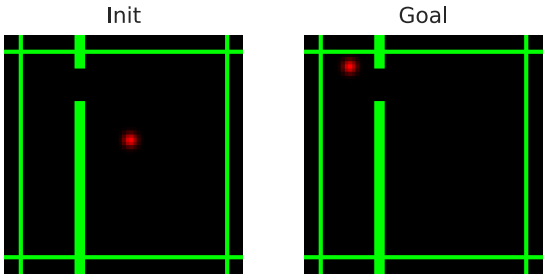
$$\mathcal{L}_{PLDM} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{var} + \beta\mathcal{L}_{cov} + \delta\mathcal{L}_{sim} + \omega\mathcal{L}_{IDM}$$

- Le WorldModel (Maes et al. 2026)

$$\mathcal{L}_{LeWM} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{Sigreg}$$

Self-supervised World models: Two rooms





Method	Success Rate (%)
$\mathcal{L}_{PEIRA}$	80%
$\mathcal{L}_{LeWM}$	87%
$\mathcal{L}_{PLDM}$	**97%**
$\mathcal{L}_{PLDM}$ w/o $\mathcal{L}_{var}$	47%
$\mathcal{L}_{PLDM}$ w/o $\mathcal{L}_{cov}$	46%
$\mathcal{L}_{PLDM}$ w/o $\mathcal{L}_{sim}$	61%
$\mathcal{L}_{PLDM}$ w/o $\mathcal{L}_{IDM}$	1%

Preliminary but encouraging

Conclusion

Summary

- There is a formal connexion between SimSiam-like SSL methods and non-linear CCA
- Collapse is still an issue for these methods
- PEIRA leverages this connection with CCA while provably avoiding collapse
- Promising results on image representation and world modeling

Future directions/open questions

- Similar connexion for Dino? SIGReg? etc
- More applications? Video, time series
- Smarter algorithms? Avoiding full matrix inversion? Variance reduction, etc

Thank you!