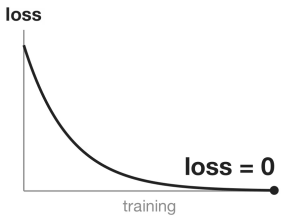
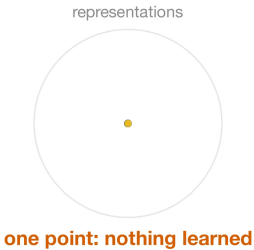
Drive vs. Decay

On the Training Dynamics of Joint-Embedding Predictive Architectures





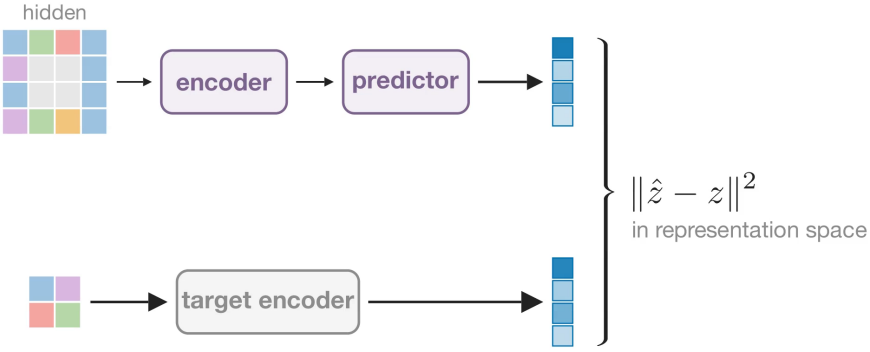
José Lucas De Melo Costa · Seong Woo Ahn · Fabrice Popineau · Arpad Rimmel · Bich-Liên Doan

Université Paris-Saclay, CNRS, CentraleSupélec, LISN **NeurIPS 2026**

CentraleSupélec | université PARIS-SACLAY

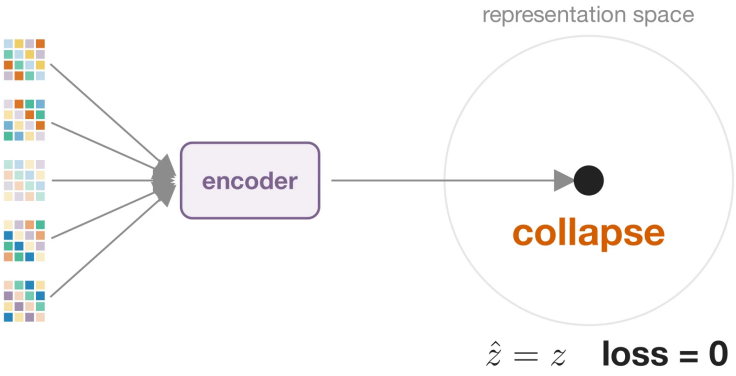
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

A JEPA predicts in representation space



A JEPA predicts the representation of a hidden view from a visible one, in latent space.

Collapse



Collapse: every input maps to the same point, the prediction is exact, the loss is zero, nothing is learned.

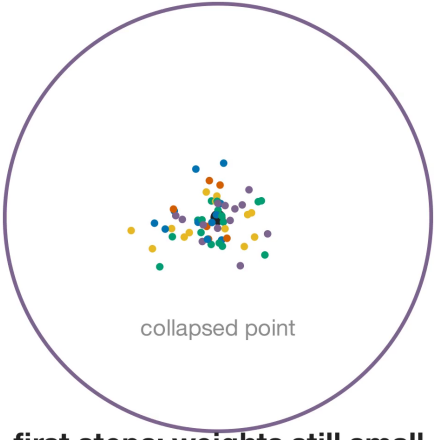
The usual tricks

stop-gradient | EMA target | predictor | masking

when? **why?**

The usual anti-collapse tricks work, but mostly as heuristics.

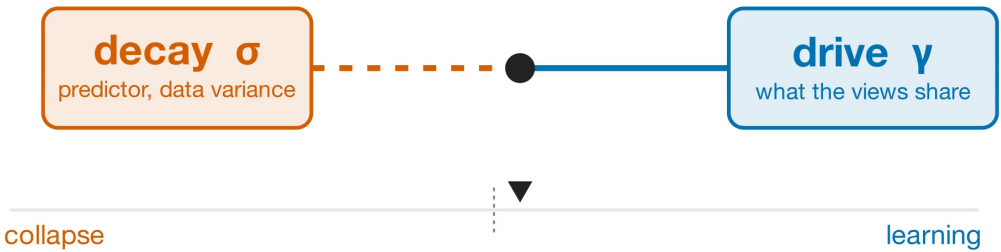
Zoom in: the first moments of training



first steps: weights still small

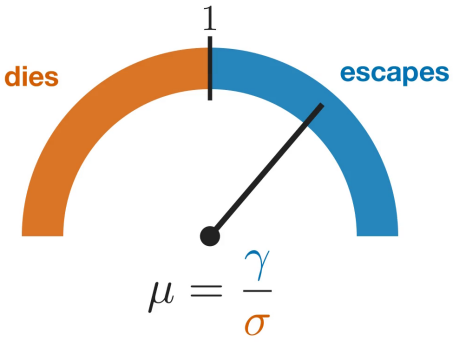
We linearise the coupled gradient flow (encoder, EMA target, predictor) around the collapsed point.

A tug of war



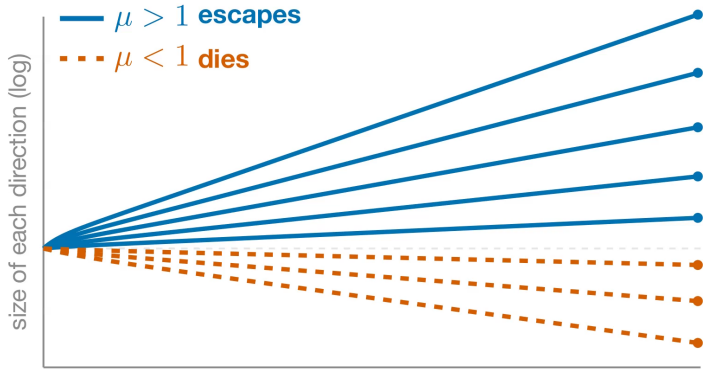
Drive γ: cross-covariance between views. Decay σ: predictor capacity and context covariance.

One ratio decides



Per mode, $\mu_i = \gamma_i / \sigma_i$: if $\mu_i > \mathbf{1}$ the collapsed point is unstable and the mode grows; if $\mu_i < \mathbf{1}$ it decays.

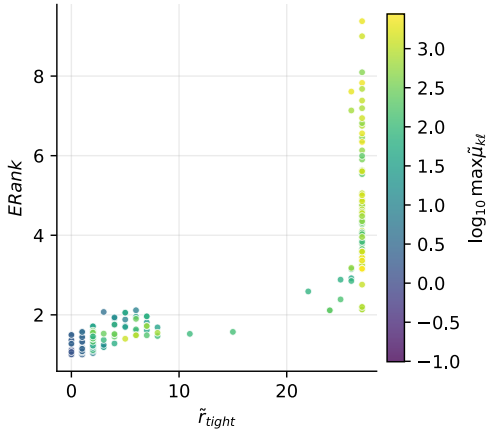
The linear model, direction by direction



exact solution of the linear model, one line per direction

Linear model: directions with $\mu > 1$ escape, the others fade. Collapse is decided early, mode by mode.

A rank budget



count of directions with $\mu > 1$

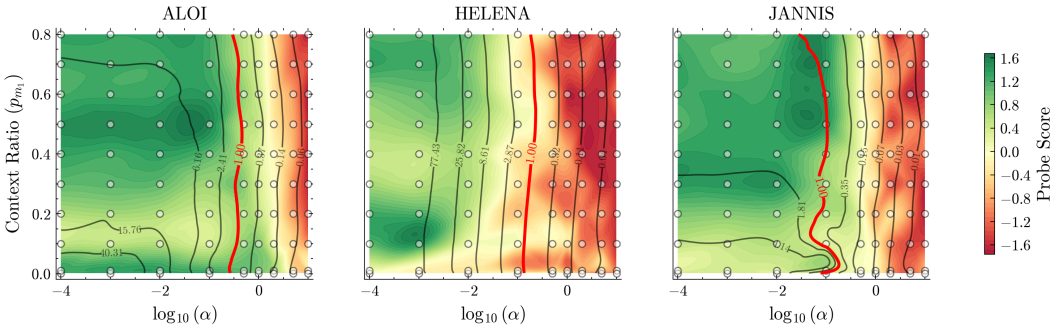
$\rho = 0.91$

on `helena`

810 runs

The count of modes with $\mu > 1$ is a rank budget; it tracks the effective rank ($\rho = +0.91$, helena, 810 runs).

The boundary $\mu = 1$ shows up in the data

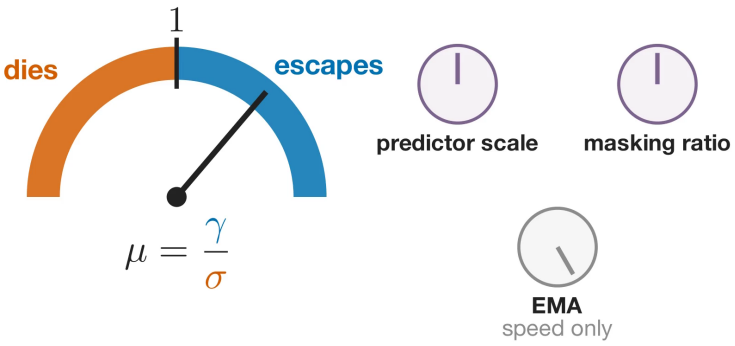


← **learning** **red line:** $\mu = 1$ **collapse** →

more than **800** Tabular-JEPA configurations

The predicted boundary $\mu = 1$ (red) matches the data across more than 800 Tabular-JEPA configurations.

The tricks are knobs on one dial

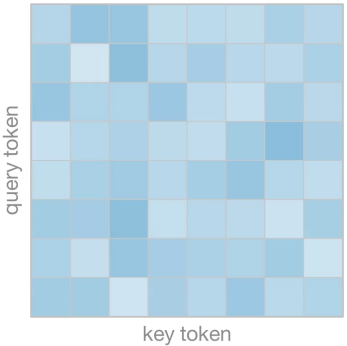


in the linear model: seven mechanisms, seven doses of μ

Predictor scale and masking ratio move μ; EMA sets only the timescale.

Can we turn the dial on purpose?

predictor attention at initialisation



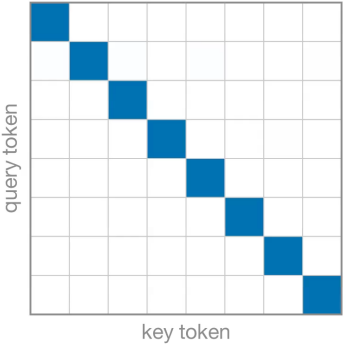
spread out

prediction far from target

A standard transformer predictor starts with diffuse attention, far from the identity.

ResidualPred

ResidualPred at initialisation



$$\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \beta_{\text{init}}\, I\right) V$$

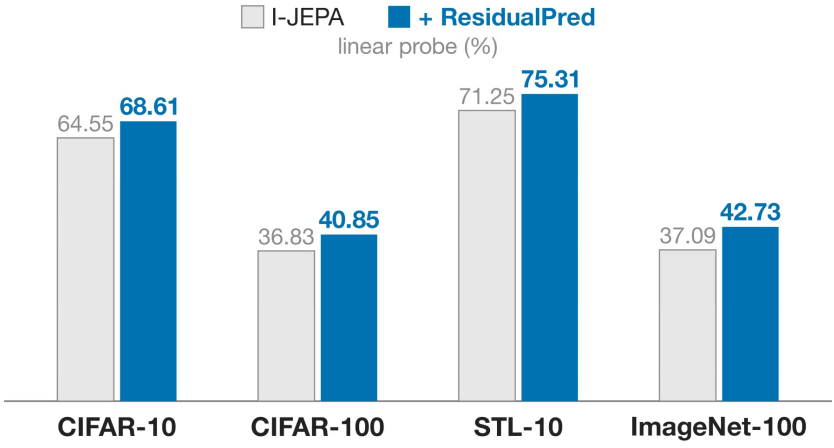
each token looks at itself

one hyperparameter

no new loss, no extra cost

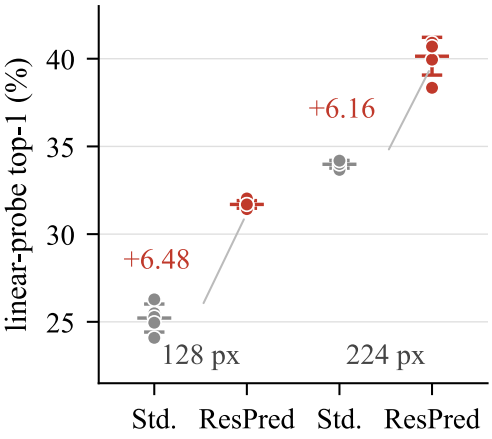
ResidualPred: a diagonal bias $\beta_{\text{init}} I$ on the attention logits, so each token attends to itself at initialisation.

I-JEPA pretraining



I-JEPA linear probe (%), standard predictor vs. ResidualPred (Table 4 of the paper).

ImageNet-1k pilot (matched budget)



25.22 → 31.69

linear probe, top-1 (%)

every ResidualPred seed
above every standard seed

ViT-S/16 · 30 epochs · five seeds

ImageNet-1k pilot, matched budget: every ResidualPred seed is above every standard seed.

Scope

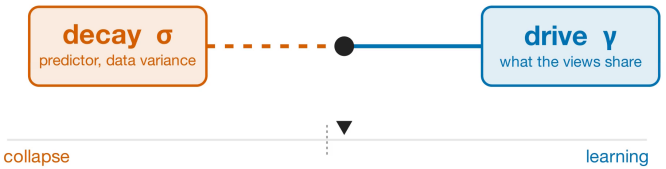
linear model

at initialisation

images: a prediction, not a proof

Scope: linear and local theory at initialisation; on images the criterion is an empirical prediction.

Drive vs. decay: one ratio tells who wins



$$\mu = \gamma \,/\, \sigma$$

paper arXiv:2610.02344

code
jose-melo/drive-vs-decay

poster in Sydney

Paper: arXiv:2610.02344. Code: github.com/jose-melo/drive-vs-decay.