IDENTIFIABILITY • SELF-SUPERVISED LEARNING • WORLD MODELS

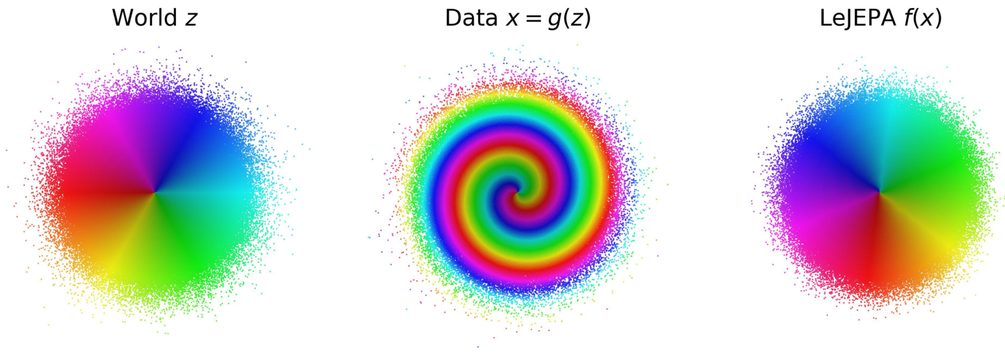
When Does LeJEPA Learn a World Model?

David Klindt[1] Yann LeCun[2] Randall Balestriero[3]

[1]Cold Spring Harbor Laboratory [2]New York University [3]Brown University

arXiv:2605.26379

PROOFS CHECKED IN LEAN 4



Latent variables (left), scrambled by an unknown nonlinear process into what we observe (centre), returned by LeJEPA as a rotation of the truth (right).

THE RESULT

Pull positive pairs together, hold the embedding Gaussian, and the encoder has no freedom left. It must invert the world up to a rotation.

$$\mathcal{L}(h) \;\geq\; 2(1-\rho)\,n$$
with equality $\iff h(z) = Qz,\;\; Q \in O(n)$

And the Gaussian is the *only* latent distribution in this class for which that holds.

THE QUESTION

When is a representation a world model?

When it **linearly recovers** the world's latent variables, which is what a linear probe assumes. No such result existed for JEPAs until SIGReg named the embedding distribution.

THE WORLD

A broad class of worlds

Latents z are never seen: an unknown g renders them into $x = g(z)$, and a positive pair is $(g(z), g(z^{'}))$.

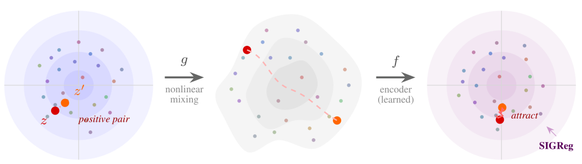
(i) Independence across coordinates, for latents and transitions alike.

(ii) Stationarity: $p(z) = p(z^{'})$.

(iii) Additive noise: $z_i^{'} = m_i(z_i) + \eta_i,\ \eta_i \perp z_i$.

Take $z \sim \mathcal{N}(0, I_n)$ and exactly one transition survives, the Ornstein–Uhlenbeck channel:

$$z^{'} = \rho\, z + \sqrt{1-\rho^2}\,\eta, \qquad \eta \sim \mathcal{N}(0, I_n)$$



LeJEPA pulls the pair together and holds the embedding Gaussian.

THE LEARNER

LeJEPA as a constraint

With $h = f \circ g$, the world's mixing composed with our encoder:

$$\min_h \underbrace{\mathbb{E}\big[\|h(z^{'}) - h(z)\|^2\big]}_{\text{alignment}} \quad \text{s.t.} \quad \underbrace{h(z) \sim \mathcal{N}(0, I_n)}_{\text{SIGReg}}$$

Nothing is assumed about h but measurability. Whitening turns alignment into correlation:

$$\mathcal{L}(h) = 2n - 2\sum\nolimits_i \mathbb{E}\big[h_i(z^{'})\, h_i(z)\big]$$

THE MECHANISM

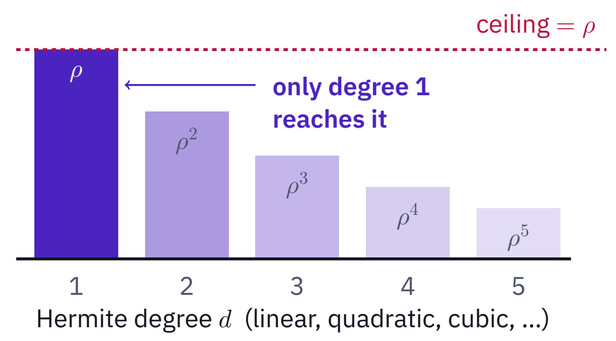
Every degree of nonlinearity is taxed

Hermite polynomials are the eigenfunctions of a Gaussian transition, with eigenvalue ρ^d at degree d. With w_d the variance share at degree d,

$$\mathbb{E}\big[h_i(z^{'})h_i(z)\big] = w_1\rho + w_2\rho^2 + w_3\rho^3 + \cdots \;\leq\; \rho$$

with equality if and only if $w_1 = 1$.

Correlation carried across a positive pair, at $\rho = 0.7$

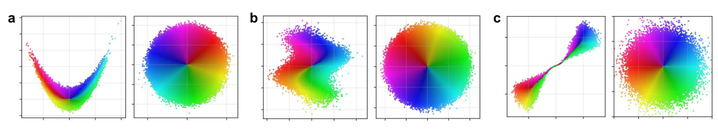


Theorem 1 Linear identifiability

Let h be *any* measurable map with $h(z) \sim \mathcal{N}(0, I_n)$. Then $\mathcal{L}(h) \geq 2(1-\rho)n$, with equality iff $h(z) = Qz$ for orthogonal Q, and at any such optimum

$$h(z^{'}) \mid h(z) \;\sim\; \mathcal{N}\big(\rho\, h(z),\, (1-\rho^2)I_n\big)$$

The latent variables *and* the dynamics come back, up to one rotation.



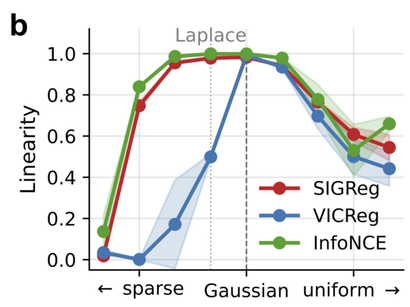
Three nonlinear mixings: parabolic shear, sinusoidal shear, RealNVP coupling. Observation left, learned embedding right.

At scale: SIGReg and VICReg hold $R^2 > 0.999$ out to $N = 1024$ latents; InfoNCE degrades at a fixed kernel width.

Theorem 2 The Gaussian is unique

Take any world satisfying (i) to (iii). If every minimiser with $\mathrm{Cov}(h(z)) = I_n$ is linear, then z is Gaussian.

Sturm–Liouville theory makes the slowest eigenfunction monotonic for any latent distribution. Forcing it *affine* forces a linear score, and only the Gaussian has one.



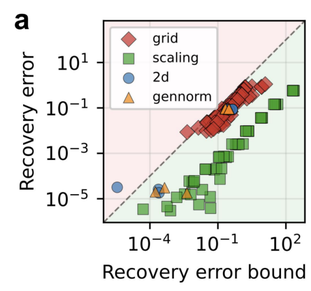
Across the generalised normal family, linear recovery peaks sharply at the Gaussian, $\alpha = 2$.

STABILITY

Theorem 3 Approximate recovery

With alignment gap δ, whitening error $\varepsilon = \|\mathrm{Cov}(h(z)) - I_n\|_F$ and $D = \delta/(2\rho(1-\rho))$, there is a $Q \in O(n)$ with

$$\mathbb{E}\big[\|h(z) - Qz\|^2\big] \;\leq\; D + (\varepsilon + D)^2$$



Recovery error against the bound, over every run. Alignment is the hard part; whitening is close to free, so training loss is a usable proxy for identifiability.

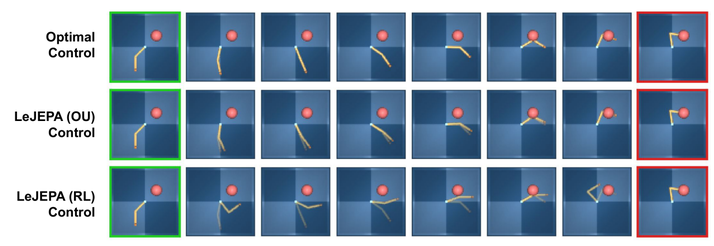
THE PAYOFF

A rotation is enough to plan

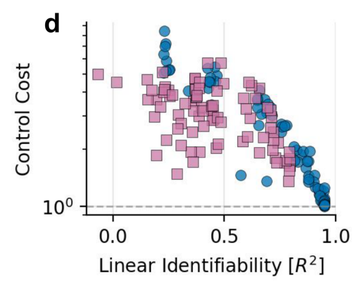
Theorem 4 Planning in the latent

Let $h(z) = Qz$ and let a finite-horizon control problem have costs that are $O(n)$-invariant in the state. Then $\hat{V}^*(h(z_0)) = V^*(z_0)$ and $\hat{a}^*_{1:T}(h(z_0)) = a^*_{1:T}(z_0)$.

Goal reaching, LQR, and any cost built from norms or inner products qualify.



Straight latent lines, decoded by nearest-neighbour retrieval.



Control cost falls monotonically with linear identifiability, across every model we trained.

ASSUMPTIONS AGAINST REALITY

ASSUMPTION	WHERE REALITY DEPARTS	STATUS
Gaussian latents $z \sim \mathcal{N}(0, I_n)$	Real latents are often heavy-tailed, bounded or multimodal, though task-relevant ones are aggregates, and aggregates drift Gaussian.	NECESSARY Theorem 2 rules out every alternative.
Isotropic transitions equal ρ across coordinates	The two Reacher joints decorrelate at different rates, $\rho_0 \neq \rho_1$, and per-dimension recovery goes anisotropic. Simultaneous extraction needs $\max_\alpha K_\alpha < 2 \min_\beta K_\beta$.	STRUCTURAL Escapes: sequential extraction, or a reweighted alignment.
Matched dimension $m = n$	$m < n$ leaves the subspace open, and says nothing about superposition. $m > n$ forces the spare dimensions to collapse or go redundant.	OPEN Not settled in either direction.
Exact Gaussianity, global optimum	Training stops short of the optimum, and SIGReg only approximately Gaussianises at finite batch size.	ABSORBED Theorem 3 bounds the error, continuously in δ and ε.

Scope. This is the state half of a world model: the action-conditioned transition $\hat{p}(\hat{z}^{'} \mid \hat{z}, a)$ still has to be learned, which points straight at interventional causal representation learning.

An empirically successful recipe turns out to be a mathematical guarantee, in exactly the worlds described above.

klindt@cshl.edu • arXiv:2605.26379v1 (25 May 2026) • All four theorems formalised in Lean 4 with Mathlib, zero `sorry` obligations.