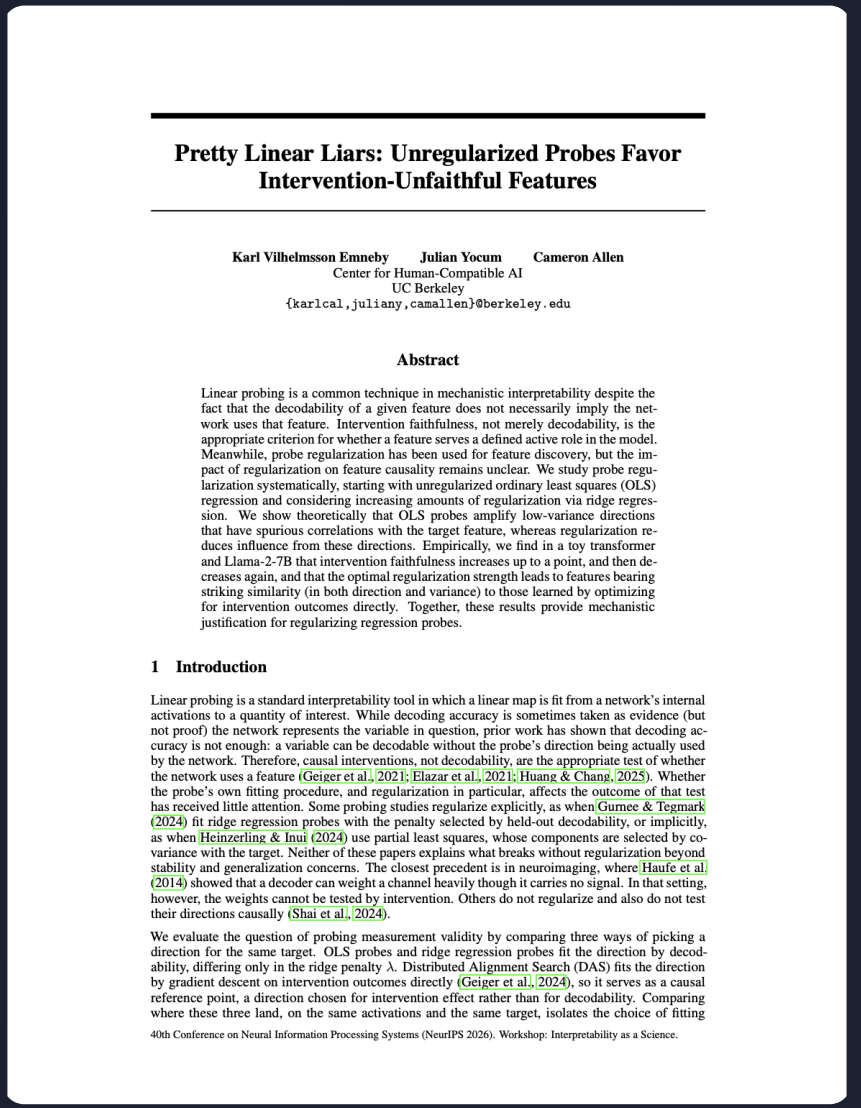
Pretty Linear Liars

Karl Vilhelmsson Emneby, Julian Yocum, Cameron Allen

Workshop paper at NeurIPS InterpScience, Dec 2026



Pretty Linear Liars: Unregularized Probes Favor Intervention-Unfaithful Features

Karl Vilhelmsson Emneby **Julian Yocum** **Cameron Allen**
Center for Human-Compatible AI
UC Berkeley
{karlcal,juliany,camallen}@berkeley.edu

Abstract

Linear probing is a common technique in mechanistic interpretability despite the fact that the decodability of a given feature does not necessarily imply the network uses that feature. Intervention faithfulness, not merely decodability, is the appropriate criterion for whether a feature serves a defined active role in the model. Meanwhile, probe regularization has been used for feature discovery, but the impact of regularization on feature causality remains unclear. We study probe regularization systematically, starting with unregularized ordinary least squares (OLS) regression and considering increasing amounts of regularization via ridge regression. We show theoretically that OLS probes amplify low-variance directions that have spurious correlations with the target feature, whereas regularization reduces influence from these directions. Empirically, we find in a toy transformer and Llama-2-7B that intervention faithfulness increases up to a point, and then decreases again, and that the optimal regularization strength leads to features bearing striking similarity (in both direction and variance) to those learned by optimizing for intervention outcomes directly. Together, these results provide mechanistic justification for regularizing regression probes.

1 Introduction

Linear probing is a standard interpretability tool in which a linear map is fit from a network's internal activations to a quantity of interest. While decoding accuracy is sometimes taken as evidence (but not proof) the network represents the variable in question, prior work has shown that decoding accuracy is not enough: a variable can be decodable without the probe's direction being actually used by the network. Therefore, causal interventions, not decodability, are the appropriate test of whether the network uses a feature (Geiger et al., 2021; Elazar et al., 2021; Huang & Chang, 2025). Whether the probe's own fitting procedure, and regularization in particular, affects the outcome of that test has received little attention. Some probing studies regularize explicitly, as when Gurnee & Tegmark (2024) fit ridge regression probes with the penalty selected by held-out decodability, or implicitly, as when Heinzerling & Inui (2024) use partial least squares, whose components are selected by covariance with the target. Neither of these papers explains what breaks without regularization beyond stability and generalization concerns. The closest precedent is in neuroimaging, where Haufe et al. (2014) showed that a decoder can weight a channel heavily though it carries no signal. In that setting, however, the weights cannot be tested by intervention. Others do not regularize and also do not test their directions causally (Shai et al., 2024).

We evaluate the question of probing measurement validity by comparing three ways of picking a direction for the same target. OLS probes and ridge regression probes fit the direction by decodability, differing only in the ridge penalty λ. Distributed Alignment Search (DAS) fits the direction by gradient descent on intervention outcomes directly (Geiger et al., 2024), so it serves as a causal reference point, a direction chosen for intervention effect rather than for decodability. Comparing where these three land, on the same activations and the same target, isolates the choice of fitting

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Interpretability as a Science.

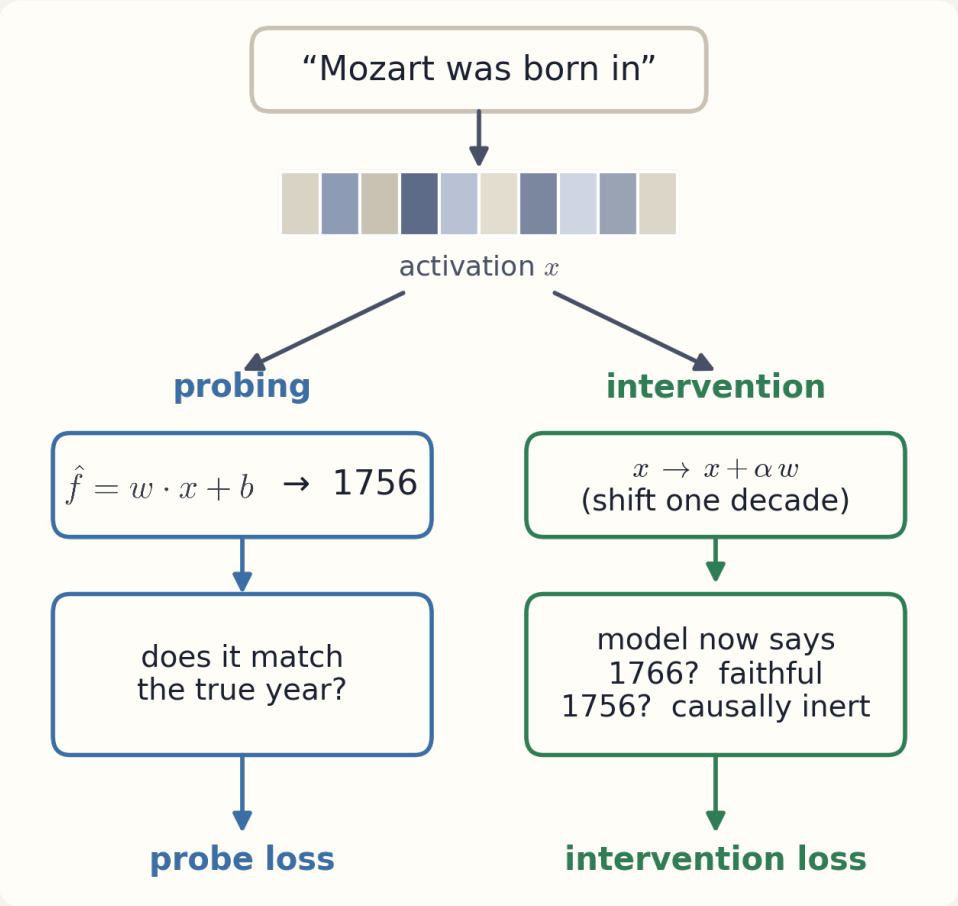
Probes and interventions

Linear probe — A regression from a layer's activations to a target, such as birth year. Its weight vector is the probe direction.

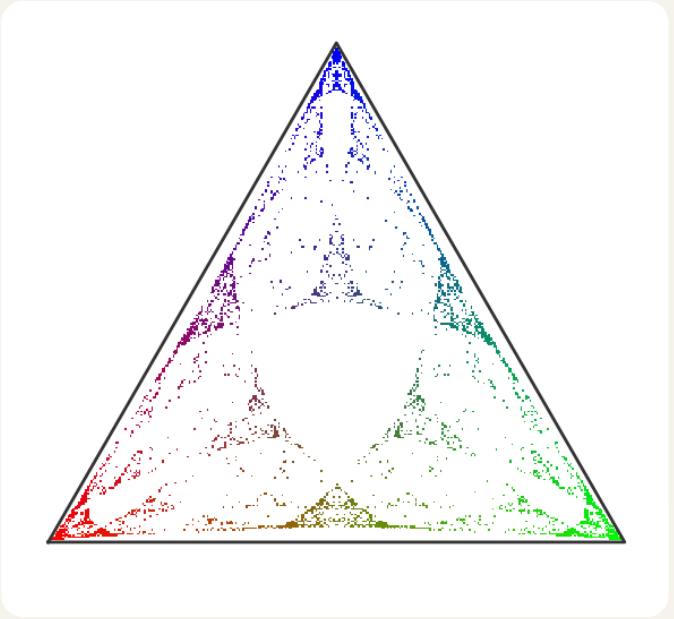
Decodability — How well the probe predicts the target, measured by probe loss.

Causal intervention — Edit the activations along the probe direction and check whether the output changes the way the target predicts.

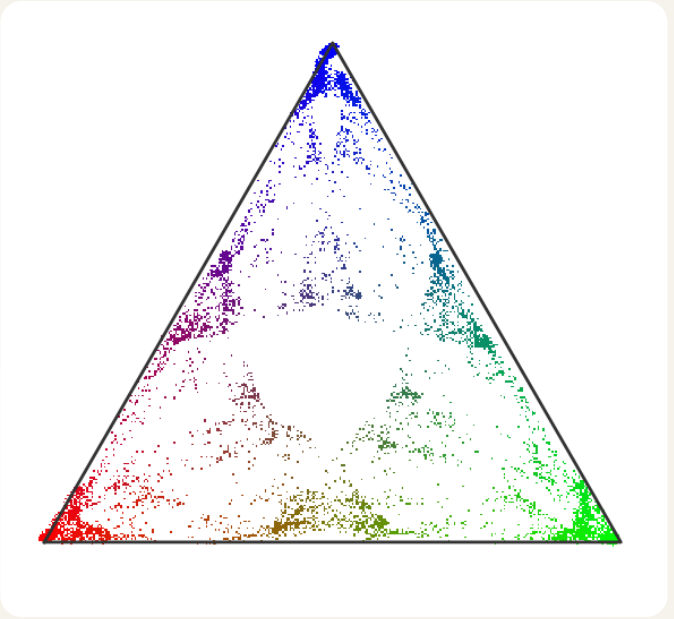
Intervention-faithful — The output moves as predicted. Intervention loss measures the mismatch.



Do these probes "find features" in a meaningful sense?



Ground-truth belief simplex



OLS probe reconstruction

It sure *looks* like it…

Belief simplex geometry over training

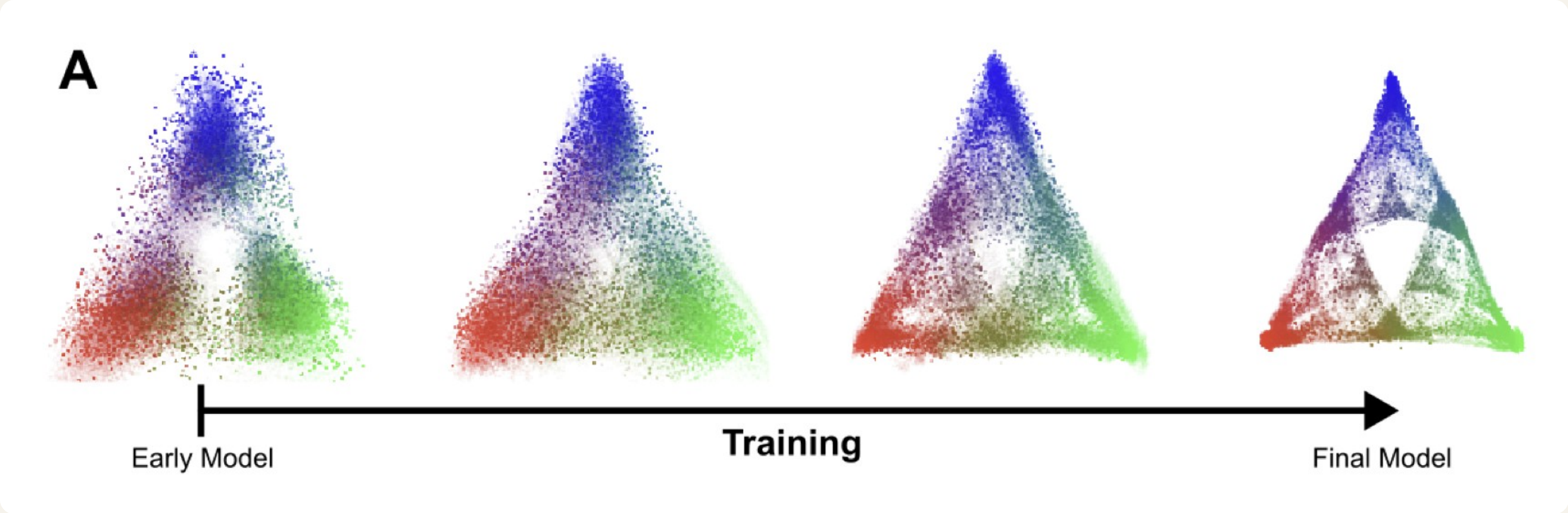
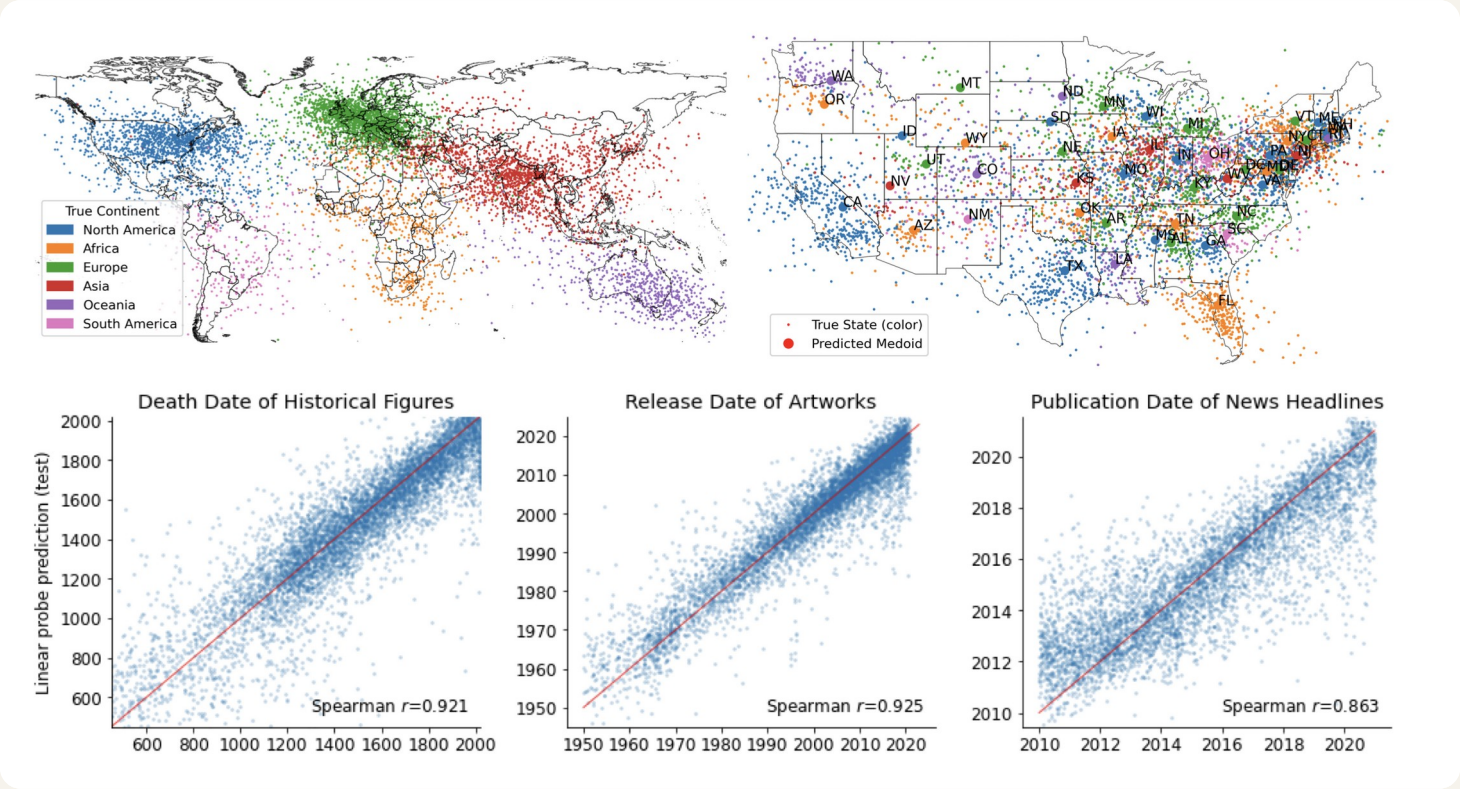


Figure from Shai et al. (2024), *Transformers Represent Belief State Geometry in their Residual Stream*.

Another well-known probing result (Gurnee & Tegmark, 2024)



From that paper: "*High predictive performance on out-of-sample data indicates that the base model has temporal and spatial information linearly decodable in its representations, although* ***this does not imply that the model actually uses these representations***"

Decodability is known to be insufficient

Causal interventions are important for claiming that a network uses a feature (correlation v. causation).

Geiger et al., 2021

Some probing studies regularize, some don’t, and none discuss explicitly whether it affects *intervention faithfulness*.

Gurnee & Tegmark, 2024 · Heinzerling & Inui, 2024 · Shai et al., 2024

Two testbeds and a causal reference

Testbed 1

Mess3 HMM

- Pre-trained transformer from Shai et al. (2024).
- Belief state and correct output are exactly computable, so intervention loss is exact.
- Intervention is steering along the plane formed by the probe vectors.

Testbed 2

Llama-2-7B, birth year

- Wikidata birth years from Heinzerling & Inui (2024), where a causally operative direction is known to exist.
- Intervention is an interchange of the direction’s coordinate from a source prompt into a base prompt.

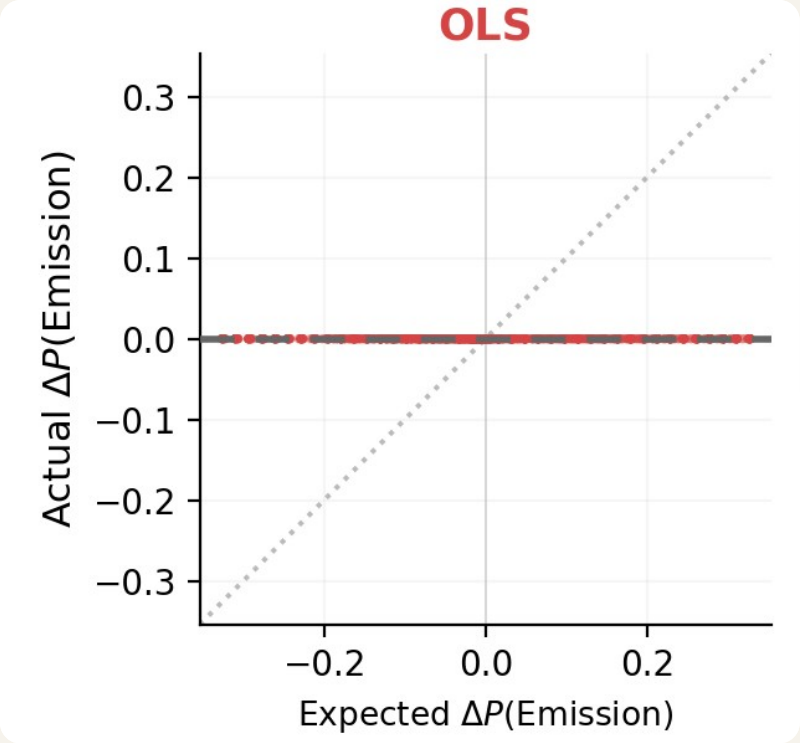
Causal "ground truth-ish" reference

DAS

- Distributed Alignment Search (Geiger et al., 2024) fits the direction by gradient descent on intervention outcomes rather than decodability.

We run OLS probes, ridge regression probes over a λ sweep, and DAS, on the same activations and target.

Steering along the plane given by the OLS probes



Causally inert.

- We **perturb activations** along the simplex plane to shift the represented belief
- Then we **compare** the neural network's change in **next-token distribution** to the change the **HMM would predict** for the shifted simplex belief.

An intervention-faithful direction would put the points on y = x.

It's known that OLS instability harms *generalization*. It hasn't been linked to *intervention-unfaithfulness*.

Probe weight on principal direction v

$$w_v = \frac{\mathrm{cov}(v, f)}{\sigma_v^2 + \lambda/n}$$

Sample covariance of a target-uncorrelated direction

$$|\mathrm{cov}(v, f)| \sim \frac{\sigma_v \sigma_f}{\sqrt{n}}$$

OLS (λ = 0)

$$|w_v| \sim \frac{\sigma_f}{\sigma_v \sqrt{n}} \to \infty$$

Grows without bound as the variance shrinks

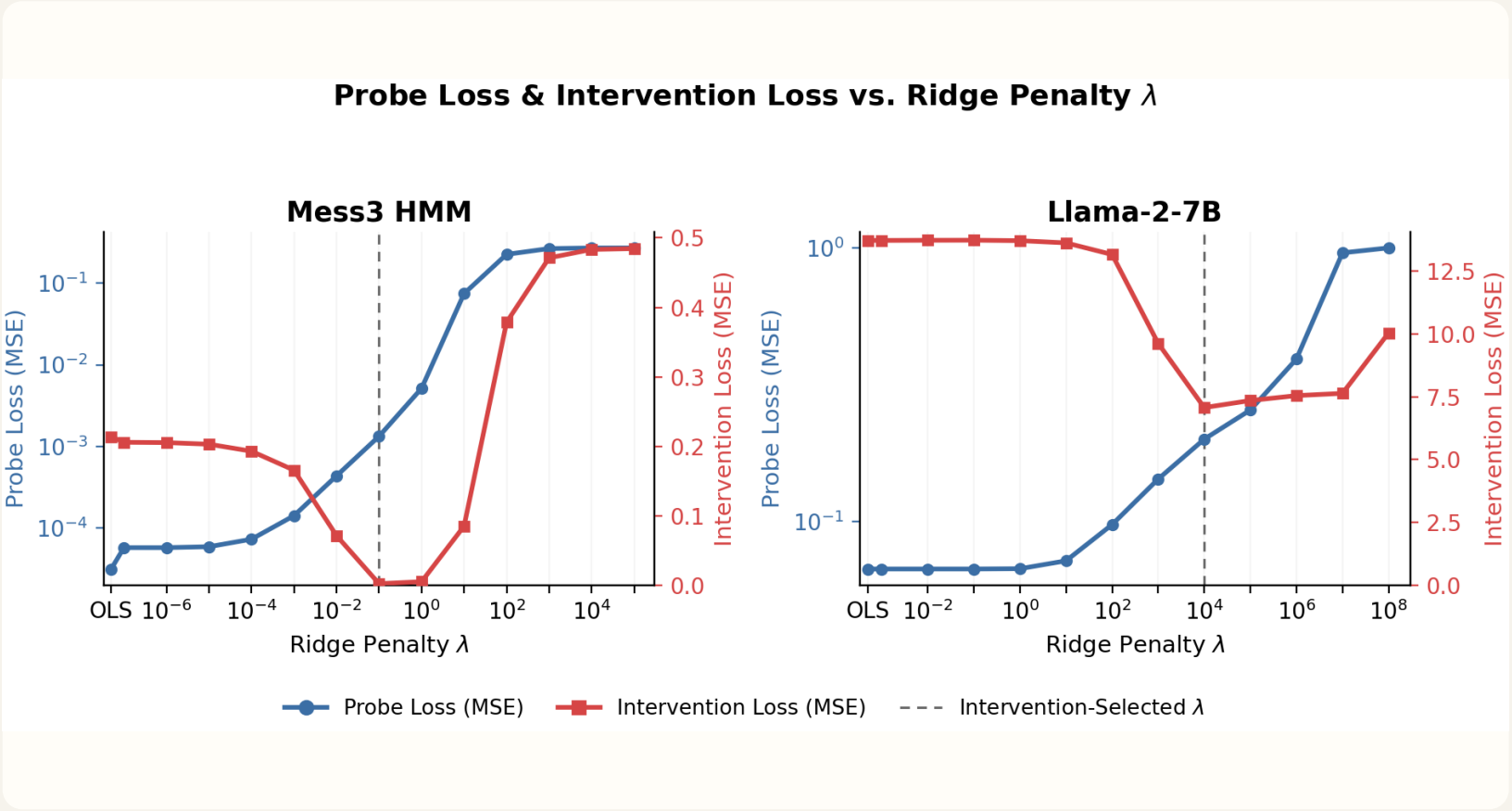
Ridge regression (λ > 0)

$$|w_v| \leq \frac{\sigma_v \sigma_f \sqrt{n}}{\lambda} \to 0$$

Vanishes as the variance shrinks

Derivation (which is classical) in the appendix (Hoerl & Kennard, 1970).

Probe loss and intervention loss disagree on λ



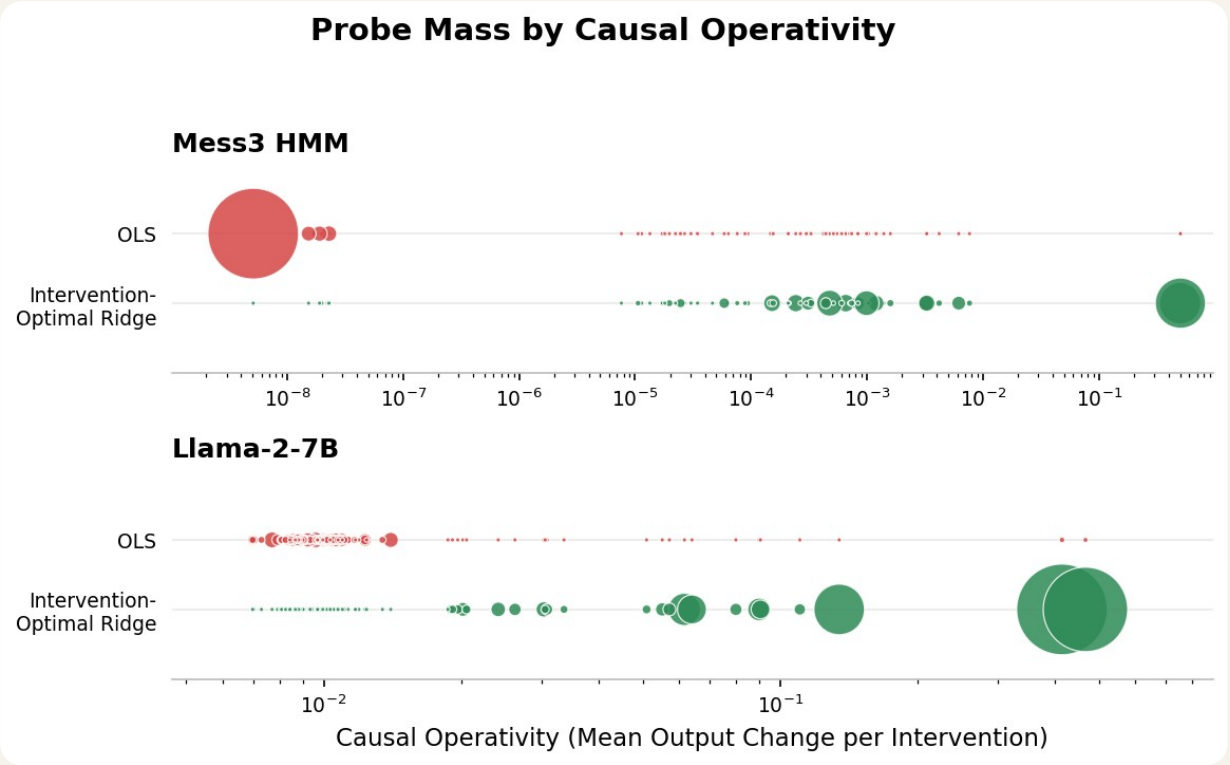
Probe loss (trivially) rises monotonically with λ.

Intervention loss has an interior minimum.

Regularization *buys intervention faithfulness at the expense of decodability.*

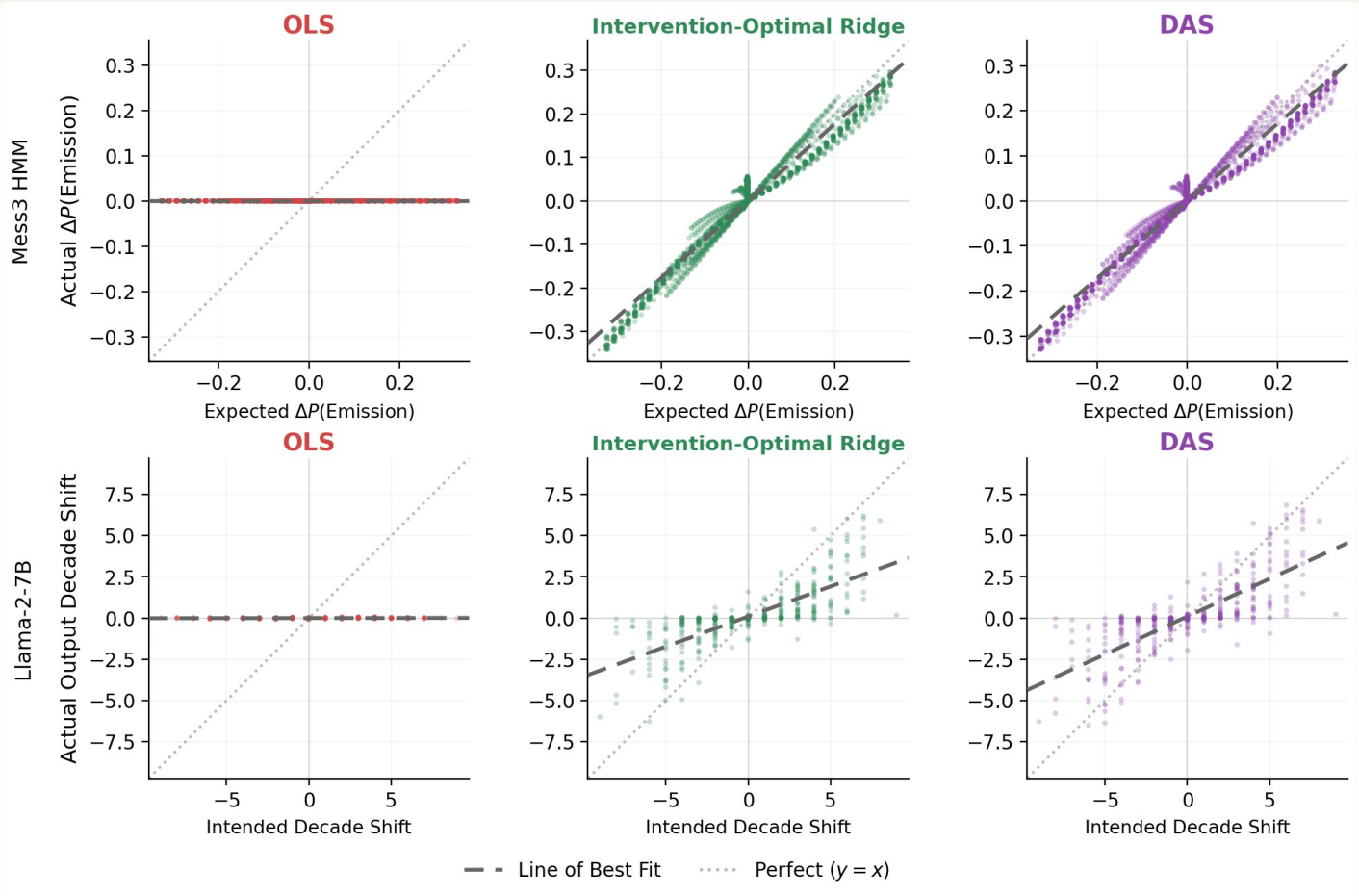
Note: Held out on 20% of the HMM data, OLS probes still decode 10× better than the intervention-optimal ridge regression probes

OLS probe mass sits on "dead weight"



Each bubble is a principal component. Area is its share of probe mass, position is its causal operativity.

Intervention-optimal ridge regression probes are quite intervention-faithful



Intervention loss across the belief simplex

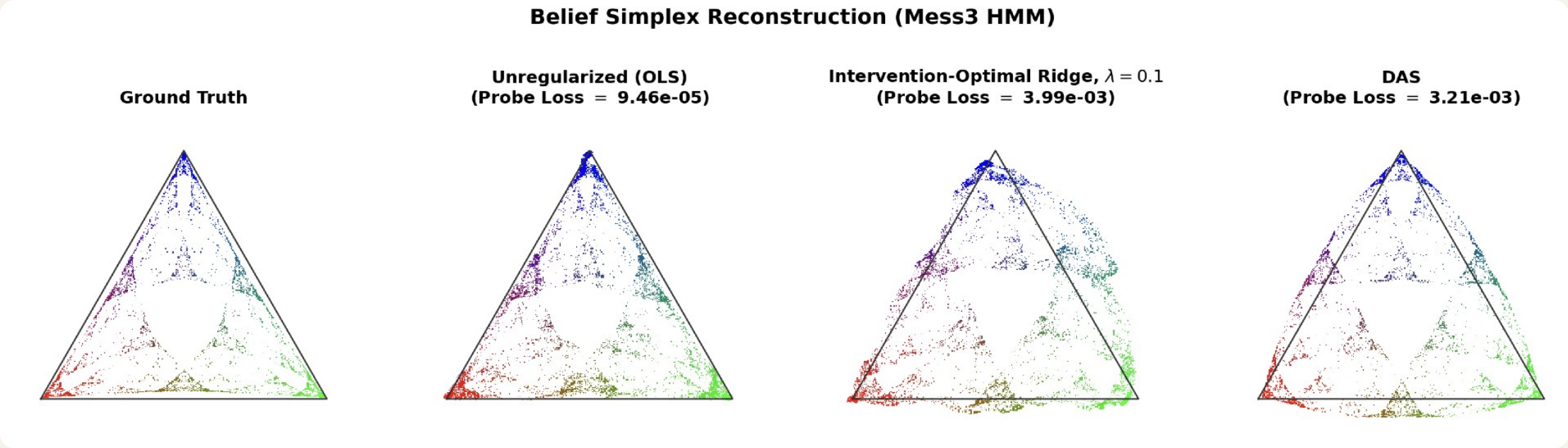


From each start state (star), we steer toward every target belief. Color is intervention loss.

Top, OLS probe. Causally inert.

Bottom, intervention-optimal ridge probe. Low almost everywhere.

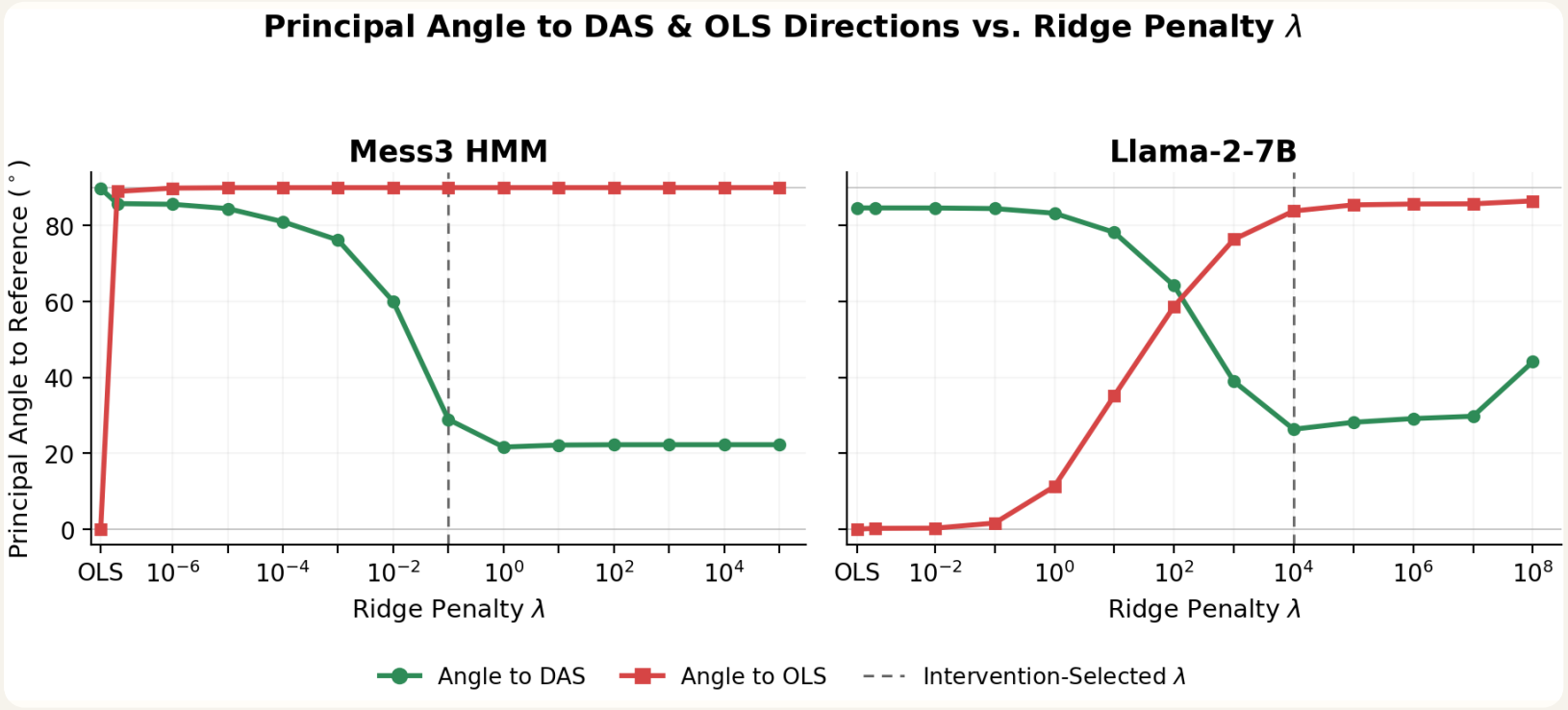
OLS has the lowest probe loss, yet it is causally inert



Intervention-optimal ridge probes align with DAS

OLS probes are nearly orthogonal to DAS

but properly regularized ridge regression probes are much closer to DAS



Note: Intervention-optimal ridge regression isn't just chasing the highest variance. On Llama-2-7B, the ridge direction (variance ~3.83) largely ignores the top PC (variance 25,570).

Select the ridge penalty by intervention loss

Recommendation

Fit ridge regression probes over a λ sweep and select λ by intervention loss, not decodability.

Limits

- Regression probes on continuous targets only.
- One LLM.
- Faithfulness is partial.
- While we expect OLS regression probes to fail, we don't guarantee that ridge regression always finds faithful features.

Next steps

A logistic counterpart to this paper?

Thank You to Julian, Cam, Shraya, Sarah and the rest of CHAI! I've had a blast.

Deriving the OLS blow-up

1. In the principal-component basis the coordinates are uncorrelated, so the objective decouples into one problem per direction (*f* and *z* are mean-centered here!).

$$\min_w \sum_j (w\, z_j - f_j)^2 + \lambda w^2$$

2. Set the derivative to zero and solve for the per-direction ridge regression weight.

$$\frac{d}{dw}\left[\sum_j (w\, z_j - f_j)^2 + \lambda w^2\right] = 2\sum_j z_j\,(w\, z_j - f_j) + 2\lambda w = 0 \quad \Rightarrow \quad w = \frac{\Sigma_j\, z_j f_j}{\Sigma_j\, z_j^2 + \lambda} = \frac{\mathrm{cov}_{vf}}{\sigma_v^2 + \lambda/n}$$

3. For a target-uncorrelated direction, the sample covariance is an average of n zero-mean terms, so its typical size is its standard error.

$$\mathrm{std}(\mathrm{cov}_{vf}) = \frac{\sigma_v\, \sigma_f}{\sqrt{n}}$$

4. OLS weight grows without bound as the variance along v goes to zero. Ridge weight vanishes.

$$|w_{\mathrm{OLS}}| = \frac{|\mathrm{cov}_{vf}|}{\sigma_v^2} \sim \frac{\sigma_f}{\sigma_v \sqrt{n}} \qquad |w_{\mathrm{ridge}}| \le \frac{n\,|\mathrm{cov}_{vf}|}{\lambda} \sim \frac{\sigma_v\, \sigma_f \sqrt{n}}{\lambda}$$

Textbook statistics (e.g. Hoerl & Kennard, 1970). **What's new is linking it to intervention-unfaithfulness!**

High level points

1

OLS probes are intervention-unfaithful on published testbeds

- Previously only on custom setups (Geiger et al., 2021)

We show it on:

- Mess3 HMM transformer (Shai et al., 2024)
- Llama-2-7B birth year (Heinzerling & Inui, 2024).

2

Classical theory suggests intervention-unfaithfulness

- It's classically known that OLS regression probes are unstable along low-variance directions
- This suggests OLS probes may be intervention-unfaithful, which we observe on both testbeds

3

What works instead!

- Ridge regression with λ selected by intervention loss suppresses low-variance spurious directions.
- The resulting probes are intervention-faithful and align with DAS (Geiger et al., 2024).