Pretty Linear Liars: Unregularized Probes Learn Causally Unfaithful Features

Karl Vilhelmsson Emneby [1] **Julian Yocum** [1] **Cameron Allen** [1]

Abstract

Linear probing is a common technique in mechanistic interpretability, where the goodness of fit of a linear map from activations to a target variable is taken as evidence that the network "represents" that variable. To the extent that this inference is wrong, interpretability conclusions built on probing are unreliable. We show that unregularized probes can learn *causally unfaithful features*, sometimes with beautiful geometry, for which interventions produce unintended (or no) effects on outputs, even while achieving high decodability. We provide several arguments why this unfaithfulness should arise in certain training configurations, and find empirically that it occurs even in settings where those arguments do not apply, suggesting a more general phenomenon. We find that simply regularizing probes using ridge regression can mitigate this effect and encourage learning causally faithful features. One might conclude that the decodability of a feature thereby reveals the existence (for some appropriately chosen regularization strength) of a causally faithful version. However, we also find examples of decodable features that are nevertheless unfaithful for *any* amount of probe regularization, which suggests that regularization alone is not sufficient to compensate for a poorly selected target variable.

1. Introduction

The linear probe is one of the central tools in mechanistic interpretability to assess whether a neural network has learned to represent a particular concept (Conneau et al., 2018; Hewitt & Manning, 2019; Burns et al., 2022; Gurnee & Tegmark, 2023; Shai et al., 2024). At the same time, many authors (Geiger et al., 2021; Elazar et al., 2021; Huang et al., 2025) have argued that probing alone is insufficient for revealing the mechanisms a network uses, because high decodability (a linear map accurately recovering a variable from activations) does not imply the corresponding feature vector is causally faithful (that intervening on it predictably controls the network's output). This problem is compounded in high-dimensional activation spaces (Hewitt & Manning, 2019), which contain enough degrees of freedom to linearly decode many functions of the training data, whether or not the network uses those functions for downstream computation. However, when looking for features in practice, linear probes are still generally the first thing to try.

We examine this tension between the intuitive notion that linear probes will reveal the features we care about and the emerging consensus that they are not enough. In particular, we show that probes trained via ordinary least squares (OLS) regression, the default in most such work, are directly incentivized to learn features that lack the desired causal influence on outputs. OLS disproportionately weights low-variance directions in activation space (directions along which activations barely vary across inputs), producing what we call *causally unfaithful* features: ones for which intervening on the probe direction fails to predictably alter the model's output. We give reasons to expect these directions to be causally unfaithful when probing pre-LayerNorm activations or when weight decay is used, and we find empirically that OLS probes are causally unfaithful even when neither condition is met.

To ensure causal faithfulness, we introduce two modifications to standard linear probing. First, we swap out linear regression for ridge regression, which penalizes large weights with L_2 regularization. Second, we sweep the regularization coefficient to identify whether there exists a setting at which probe interventions produce the expected causal effect on outputs. Together, these modifications suppress the low-variance directions that OLS gravitates toward.

We demonstrate by training both regularized and unregularized probes on a simple hidden Markov model (Shai et al., 2024), where we can compute the underlying hidden state feature exactly. Despite the OLS objective achieving extremely low probe loss and the resulting features lining up beautifully with the ground-truth belief geometry, we find

Under review at the ICML 2026 Mechanistic Interpretability Workshop. [1]University of California, Berkeley, CA, USA. Correspondence to: Karl Vilhelmsson Emneby <karl-cal@berkeley.edu>.

Preprint. May 10, 2026.

that intervening on these features has nearly no effect on network output. By contrast, the features learned by the regularized probe are approximately causally faithful: intervening along them shifts the network's output roughly as predicted by Bayes' rule.

Our approach also enables precise mechanistic analysis, since we are able to reject plausible-seeming, OLS-decodable features if we do not find a causally faithful intervention. We demonstrate this for another statistical latent variable model: the Ising model from statistical physics, in which the hidden variables can be entirely marginalized out. We train OLS probes that predict spin configurations from lattice boundary conditions, and "recover" hidden-spin variables that pass every standard diagnostic (low probe loss, generalization to held-out data, gradual improvement during training). However, the ridge regression probes reveal that such features are causally unfaithful directions at every regularization strength, since they are not actually required for predicting outputs.

Our results demonstrate that decodability alone is not evidence of causal structure, and that sweeping the regularization strength against intervention loss provides a practical tool for distinguishing causally faithful features from causally unfaithful ones.

2. Background

Definition 2.1 (Linear probe)**.** Given activations $a \in \mathbb{R}^d$ at some layer of a neural network and a target variable $f \in \mathbb{R}^k$, a **linear probe** is a linear map that produces a prediction $\hat{f}$ of f from a:

$$\hat{f} = W^\top a + b, \tag{1}$$

where $W \in \mathbb{R}^{d\times k}$ is the probe weight matrix and $b \in \mathbb{R}^k$ is a bias vector.

Given n activation–target pairs (A, F) with $A \in \mathbb{R}^{n\times d}$ and $F \in \mathbb{R}^{n\times k}$, the **OLS probe** fits W and b by minimizing:

$$\mathcal{L}_{\text{OLS}} = \|AW + \mathbf{1}b^\top - F\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The **ridge probe** adds an L_2 penalty on W (but not b):

$$\mathcal{L}_{\text{ridge}} = \|AW + \mathbf{1}b^\top - F\|_F^2 + \alpha\|W\|_F^2, \tag{3}$$

for regularization strength $\alpha > 0$. The optimal bias is always $b = \bar{f} - W^\top\bar{a}$ (the arithmetic means), so fitting with a bias is equivalent to mean-centering A and F before fitting W. We say a target variable is **linearly decodable** at a layer if a linear probe achieves low mean squared error (MSE), and refer to the probe's MSE as the **probe loss**.

Definition 2.2 (Causal faithfulness)**.** Let M be a neural network, ℓ a layer, $f \in \mathbb{R}^k$ a latent variable, and g a known ground-truth function that specifies the correct network output for any value of f. Given a probe W for f at layer ℓ, a **causal intervention** perturbs activations a at ℓ by $\Delta a = W^+\delta f$, where W^+ is the pseudoinverse of W and δf is a desired shift in f (see section 4.3 for details).

A probe is **causally operative** to the extent that the intervention changes the network's output, and **causally inoperative** to the extent that it does not. A probe is **causally faithful** to the extent that the change in output when applying Δa to the activations is close to $g(f + \delta f) - g(f)$: the network's output shifts as the ground-truth function predicts, and **causally unfaithful** to the extent that it does not.

A causally faithful probe is necessarily causally operative, but a causally operative probe need not be causally faithful. We quantify both using the **intervention loss**: the MSE between the network's post-intervention output and the ground-truth prediction $g(f + \delta f)$.

3. Why OLS Probes Learn Unfaithful Directions

3.1. OLS upweights low-variance directions

The core of our argument is that OLS regression places disproportionate weight on low-variance directions in activation space, while ridge regression suppresses them. The intuition is straightforward: if a component of the activations varies by only $\mathcal{O}(\epsilon)$ across inputs, any probe that extracts an $\mathcal{O}(1)$ signal from it must assign a weight of $\mathcal{O}(1/\epsilon)$. Regularization penalizes exactly this kind of large weight, so ridge regression suppresses low-variance components. Since directions in activation space are linear combinations of components, this suppression extends from components to directions. A more precise version of this argument follows.

Since the bias absorbs the arithmetic means, W is determined by the mean-centered activations. Let $A = USV^\top$ be the SVD of the mean-centered activation matrix A, where v_i are the principal directions in activation space and s_i the singular values. The variance along v_i is s_i^2/n (Bishop, 2006, Ch. 12).

It is also a standard result that $v_i^\top W$ (the dot product of W with principal direction v_i) is proportional to $1/s_i$ for OLS and to $s_i/(s_i^2 + \alpha)$ for ridge regression (Hastie et al., 2009, Sec. 3.4.1). The remaining factor depends on the data and target but is identical for both methods, so the difference between OLS and ridge is entirely in how they weight each direction by its variance. In other words, the ridge weight vector cannot have as large a dot product with low-variance principal directions v_i (since $s_i/(s_i^2+\alpha) \to 0$ when α dominates over s_i^2), while the OLS weight vector is free to do so (since $1/s_i$ grows without bound). Since any direction in

activation space is a linear combination of principal directions, suppressing the low-variance components forces the ridge weight vector away from low-variance directions.

The $1/s_i$ weighting does not hurt prediction accuracy: the activation along v_i scales as s_i, so the product is $\mathcal{O}(1)$ regardless of variance. But W itself is dominated by its low-variance components, since the $1/s_i$ terms are largest where s_i is smallest.

3.2. Non-causal directions can be decodable

The above analysis assumes that low-variance directions have nonzero correlation with the target, so that OLS can extract signal from them. In high-dimensional networks, we believe this to be hard to avoid, for two reasons.

Leakage across directions. In a neural network, each component of a layer's output depends on all components of the previous layer (assuming the weight matrix has no exact zeros), so whatever signal the network routes through a few directions at one layer will influence all components at the next. LayerNorm, when used, adds further mixing, since every component's output depends on every other through the mean and standard deviation. It is therefore not surprising if a feature that is causally faithful along some direction in activation space is also decodable from causally unfaithful subspaces of the residual stream.

Decodability in high dimensions. The network's output, the residual stream, and each probe branching off from the residual stream, are functions of the input, by definition. So if the residual stream is high-dimensional, many dimensions may carry information about the input (all processed in different ways), and the probe has enough degrees of freedom that some linear combination of them approximates many given (sufficiently low-complexity) functions of the input. Regression can find such a combination whether or not the network causally uses the resulting feature direction for downstream computation.

3.3. Causally operative directions should have non-tiny variance in certain settings

The argument above shows that OLS places disproportionate weight on low-variance directions, but it matters for causal faithfulness only if the network avoids routing computation through those directions. We give two arguments for when trained networks should avoid low-variance computation. These arguments are heuristic and neither constitutes a proof that low-variance directions are never causally operative.

Argument 1: LayerNorm dilutes low-variance pre-LayerNorm signals. Even when a feature is linearly decodable from pre-LayerNorm activations, the model reads from post-LayerNorm activations, where low-variance signals may be drowned out. Consider a direction u (any unit vector in activation space) with low variance in pre-LayerNorm space; that is, the variance of $u^\top a$ across inputs is small (the signal is present but at a small scale). After LayerNorm, the readout along u becomes

$$u^\top \mathrm{LN}(a) = \sum_j u_j \frac{a_j - \mu}{\sigma} = \frac{1}{\sigma}\Bigg(\underbrace{u^\top a}_{\text{signal}} - \underbrace{\mu \sum_j u_j}_{\text{noise from } \mu}\Bigg), \quad (4)$$

where μ and σ are the component-wise mean and standard deviation. The original signal ($u^\top a$) hardly varies, but the new noise term ($\mu \sum_j u_j$) fluctuates across inputs because μ is driven by high-variance components. The post-LayerNorm readout is therefore dominated by μ-dependent noise rather than the direction's own content. Since the network reads from post-LayerNorm activations, it cannot reliably extract information from low-variance pre-LayerNorm directions, and has little incentive to route computation through them during training. If the network does not use these directions, then a probe that loads onto them will be causally unfaithful, regardless of how well it decodes.

Argument 2: Weight decay makes low-variance readouts expensive. The network faces the same problem as the probe: reading from a low-variance direction requires large weights. To extract an $\mathcal{O}(1)$ signal from a principal direction v_i with singular value s_i, the network's downstream weights must have a projection onto v_i of order $1/s_i$ to compensate for the small activation variance. Weight decay with coefficient λ penalizes the squared norm of the weight matrix, so the cost of maintaining this projection scales as λ/s_i^2. For sufficiently small s_i, this cost outweighs any loss reduction the direction could provide, and the network learns to ignore it. OLS, by contrast, pays no penalty for large weights, so it freely loads onto the same directions the network has abandoned. This applies whenever weight decay is used during training.

Neither argument applies to the HMM transformer of Shai et al. (2024): the weight decay argument relies on weight decay, and we probe post-LayerNorm, where the LayerNorm argument does not apply either. The empirical results (Section 5) nevertheless show that OLS finds causally unfaithful directions and that ridge regression with appropriate regularization strength finds approximately causally faithful ones. Identifying the precise mechanisms that discourage low-variance computation in this setting is an open question.

4. Experimental Setup

In both experiments, we take the latent-variable probabilities as the candidate feature to probe for and intervene on.

4.1. The Mess3 HMM

The Mess3 HMM (Shai et al., 2024) is an ideal testbed because the ground-truth latent variable (the belief state) and the correct output for any belief are both exactly computable, so we can measure intervention loss precisely. The HMM defines a 3-state hidden Markov model with transition matrix $T_{ij} = P(s_{t+1} = j \mid s_t = i)$ and emission distribution $E_{ik} = P(x_t = k \mid s_t = i)$, parametrized by a mixing parameter $x = 0.05$ and emission concentration $\alpha = 0.85$. The joint distribution over hidden states $s_{1:T}$ and observations $x_{1:T}$ is:

$$P(s_{1:T}, x_{1:T}) = P(s_1) \prod_{t=1}^{T-1} T_{s_t, s_{t+1}} \prod_{t=1}^{T} E_{s_t, x_t}. \quad (5)$$

The Bayes-optimal predictor for $P(x_{t+1} \mid x_{1:t})$ must maintain a *belief state* $b_t = P(s_t \mid x_{1:t})$, the posterior over hidden states given observations so far. This belief is the minimal sufficient statistic for next-token prediction:

$$P(x_{t+1} \mid x_{1:t}) = \sum_{s_t, s_{t+1}} b_t(s_t)\, T_{s_t, s_{t+1}}\, E_{s_{t+1}, x_{t+1}}. \quad (6)$$

With 3 hidden states, the belief lives on a 2-simplex.

Shai et al. (2024) trained a 4-layer transformer (hidden dimension 64, 4 attention heads, MLP width 256, ReLU activations, LayerNorm) on sequences of length 8 emitted by this HMM, and showed that linear probes on the concatenated residual stream activations recover the belief simplex with near-perfect decodability. This was a striking result that established the HMM transformer as a clean demonstration of latent structure emerging during training. We build directly on their work by asking whether these directions are also causally operative and causally faithful, using their pretrained network.

4.2. Ising model

To test whether decodability implies a causally faithful direction exists, we need a model where the latent variables are provably unnecessary for the task.

The Ising model defines a probability distribution over binary spins $\sigma_i \in \{-1, +1\}$ on a lattice with Hamiltonian:

$$H(\sigma) = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j, \quad (7)$$

where the sum runs over neighboring pairs. The Boltzmann distribution is:

$$P(\sigma) \propto \exp\bigl(-H(\sigma)/T\bigr). \quad (8)$$

We use a 4×5 lattice with periodic boundary conditions, coupling $J = 1$, and temperature $T = 1.0$.

We partition lattice sites into three groups (Figure 5): **input spins** (12 boundary sites), **target spins** (2 interior sites), and **hidden spins** (6 remaining sites). Crucially for our purposes, the conditional distribution of targets given inputs can be computed exactly by marginalizing out the hidden spins via transfer matrices:

$$P(\text{targets} \mid \text{inputs}) = \sum_{\text{hidden}} P(\text{targets}, \text{hidden} \mid \text{inputs}). \quad (9)$$

This marginalization yields a closed-form function $P(\text{targets} \mid \text{inputs})$ that depends only on the input spins; the hidden spins are integrated out algebraically. The network can therefore achieve optimal prediction accuracy by learning this direct function from inputs to targets, without computing $P(\text{hidden} \mid \text{inputs})$ as an intermediate step.

By contrast, in the Mess3 HMM, predicting $P(x_{t+1} \mid x_{1:t})$ requires the network to compute a function of the entire preceding sequence. The belief state is the *efficient compression* of that history, and our causal interventions (Section 5) confirm that the trained network uses it. While a lookup table over all possible HMM histories would also work in principle (there are only 6,561 unique prefixes of length ≤ 8), the belief state is a 2-dimensional summary that works just as well and generalizes to arbitrary sequence lengths. In the Ising model, no such intermediate is needed: the sum over hidden spins collapses to a direct function of the 12 input spins.

For our Ising model experiments, we train a 4-layer transformer (hidden dimension 512, 4 attention heads, MLP width 128, ReLU, pre-norm, no dropout) to predict $P(\text{targets} \mid \text{inputs})$ from the 12 input spins, processed as a single token. We widen the residual stream from 64 to 512 dimensions: at 64, hidden-spin probes achieved poor decodability (consistent with the point about decodability in high dimensions above). The network did not need the extra capacity to solve the task; the wider residual stream gave OLS more room to find spurious correlations.

4.3. Probing and intervention methodology

We probe activations at the output of the final LayerNorm, which is also the site at which we perform causal interventions.

Probe fitting and regularization strength sweep. For the HMM, we follow Shai et al. (2024)'s methodology: exhaustive enumeration of all $3^1 + 3^2 + \cdots + 3^8 = 9{,}840$ possible sequences, deduplication to 6,561 unique prefixes and probability weighting. We add a ridge penalty and sweep the regularization strength α over several orders of

magnitude (from 10^{-2} to 10^{4}): for each α, we fit

$$W_\alpha = \arg\min_W \sum_i p_i \|a_i W - b_i\|^2 + \alpha \|W\|^2, \quad (10)$$

where p_i is the sequence probability, a_i is the activation, and b_i is the ground-truth belief. At $\alpha = 0$, this reduces to unregularized OLS. The regularization strength sweep is a core part of our methodology: it traces out the tradeoff between probe loss and intervention quality, and identifies the α that minimizes intervention error (Figure 1).

For the Ising model, we fit standard (unweighted) OLS and ridge probes on a 70/30 train/test split of all $2^{12} = 4{,}096$ input configurations, sweeping α over the same range.

Intervention procedure. Given a probe $W \in \mathbb{R}^{d\times k}$ and a desired shift $\delta f \in \mathbb{R}^k$ in the target variable, we compute the activation perturbation:

$$\Delta a = W^+ \delta f, \quad (11)$$

where $W^+ = W(W^\top W)^{-1}$ is the right pseudoinverse; this is the minimum-norm solution to the underdetermined constraint $W^\top \Delta a = \delta f$, ensuring the edit lies in the column span of W and introduces no change in directions the probe cannot see. We add Δa to the activations after the final LayerNorm (last token position for the HMM, single token for the Ising model) and record the network's post-intervention output probabilities. We compare these to the ground-truth prediction: for the HMM, this is the Bayes-optimal emission probability given the shifted belief; for the Ising model, this is the exact conditional $P(\text{targets} \mid \text{inputs}, \text{hidden} = h_{\text{target}})$ computed via transfer matrices.

Computing intervention loss (Figure 1, Appendix B). For the HMM, we select three starting sequences whose beliefs lie near each simplex corner and, for each, intervene toward every point on a uniform grid of 496 beliefs spanning the simplex. For each intervention, we compute the MSE between the network's post-intervention output probabilities and the Bayes-optimal prediction at the target belief. The reported MSE is the average over all starting points and target beliefs. For the Ising model, we compute intervention loss analogously, averaging over all input configurations and target hidden-spin states.

Computing intervention slopes (Figures 3 and 7). For the HMM, we select 20 random sequences and, for each, push the belief toward each of the three simplex corners at 21 linearly spaced strengths from 0 to 0.5. At each strength, we record the change in all three output token probabilities. Each trajectory is one (starting belief, target corner, output token) triple. The slope is the OLS fit of actual ΔP against expected ΔP (from Bayes' rule) across all trajectories; a slope of 1.0 would indicate the intervention shifts the network's output by the predicted amount on average. For the Ising model, we again follow an analogous procedure, intervening toward each possible hidden-spin configuration and comparing the network's output to the exact Ising conditional.

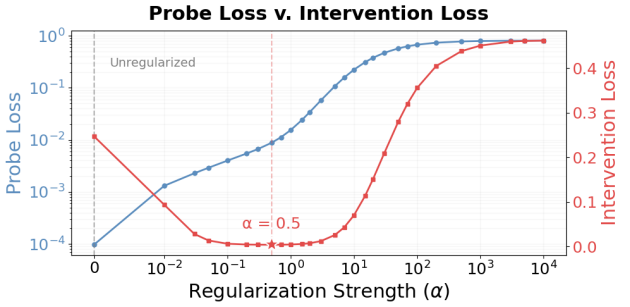

Figure 1. Mean probe loss (blue) and intervention loss (red) as a function of ridge penalty α on the HMM transformer (probability-weighted). Intervention loss is the mean MSE between post-intervention output and Bayes-optimal prediction (Section 4.3). Intervention error is minimized at $\alpha \approx 0.5$, where probe loss has increased by nearly two orders of magnitude.

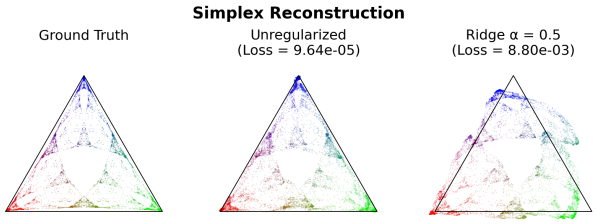

Figure 2. Belief simplex reconstruction. Left: ground truth. Center: unregularized probe (MSE $= 9.64 \times 10^{-5}$). Right: ridge regression, $\alpha = 0.5$ (MSE $= 8.80 \times 10^{-3}$). The intervention-optimal simplex is visibly distorted relative to ground truth, but interventions along these directions produce substantially lower intervention loss (Figure 1).

Intervention slopes measure causal operativity (whether the network responds at all), while intervention loss measures causal faithfulness (whether the response matches the ground-truth prediction).

5. Results

5.1. HMM

In our HMM experiments, we find that the most decodable directions are causally unfaithful. Figure 1 shows probe loss and intervention loss as a function of the ridge penalty α. Since the probes are fit on the full set of 6,561 unique prefixes (no train/test split), the probe loss shown is the training loss; intervention loss is computed as described in Section 4.3. Intervention error is minimized around $\alpha = 0.5$ (probability-weighted), where probe loss has increased

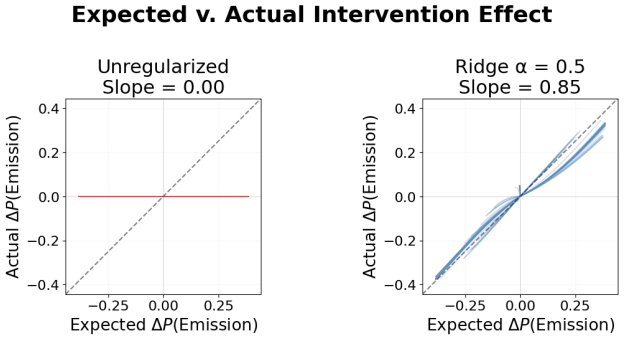

Figure 3. Causal interventions on the HMM transformer. Left: unregularized probe (slope $= 0.00$). Right: ridge regression, $\alpha = 0.5$ (slope $= 0.85$). Each line traces one trajectory across 21 intervention strengths. The x-axis is the Bayes-predicted change in emission probability; the y-axis is the network's actual change.

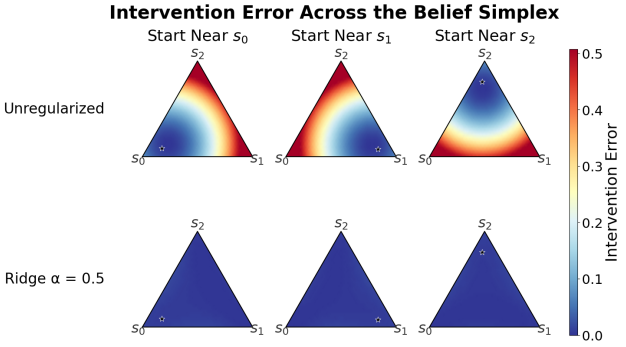

Figure 4. Intervention error on the HMM. Top row: unregularized. Bottom row: ridge regression, $\alpha = 0.5$. For each of three starting beliefs (one near each simplex corner), we intervene toward every other point on the simplex and color by intervention loss. Error grows sharply with intervention distance for the unregularized probe but remains low for ridge regression, confirming that unregularized probe directions are causally unfaithful across the entire belief simplex.

by nearly 2 orders of magnitude, from 0.0001 to 0.0088. The simplex recovered by the intervention-optimal probe is visibly distorted relative to ground truth (Figure 2), but the mean intervention loss is roughly $65\times$ lower than at $\alpha = 0$.

Figure 3 shows intervention slopes pooled into a single panel per method. The x-axis is the Bayes-predicted change in output probability; the y-axis is the network's actual change. For the unregularized probe, all trajectories are flat: the probe is causally inoperative (slope $= 0.00$). For the ridge probe, most trajectories approximately follow the diagonal (slope $= 0.85$), indicating that the probe is approximately causally faithful: the network changes its emission probabilities roughly as predicted by Bayes' rule. The slope is less than 1.0, suggesting that the belief representation is not perfectly linear. Figure 4 confirms that this pattern holds across the entire belief simplex.

Without probability weighting, the intervention-optimal α shifts to 5×10^3 and the simplex has somewhat lower MSE (1.67×10^{-3}), but the core result is identical: unregularized probes are causally unfaithful, ridge probes are approximately causally faithful (see Appendix A).

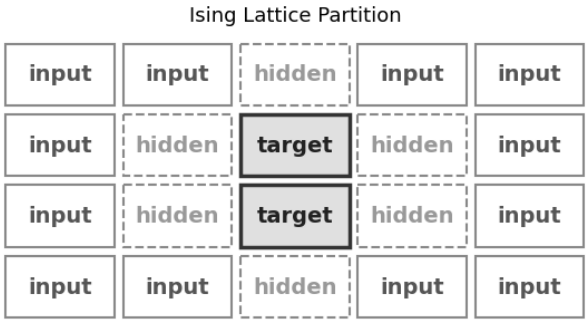

Figure 5. Ising lattice partition. The network receives the input spins and predicts the target spins; hidden spins (dashed borders) can be marginalized out exactly. Adjacent cells interact; the lattice has periodic boundary conditions (top and bottom rows wrap around).

5.2. Decodability does not imply a causally faithful direction exists

One might be tempted to conclude the following from the HMM result: "high decodability implies the belief state structure is present; the features are represented, but the OLS direction just needs a small correction." Prior work has already established that *OLS decodability* does not imply OLS interventions work as expected (Geiger et al., 2021; Elazar et al., 2021; Huang et al., 2025). Our claim is stronger:

> **Known:** Decodability does not imply the OLS directions are causally operative. This leaves open the possibility that the feature is causally operative and OLS just found the wrong directions.
>
> **Our claim:** Decodability does not imply a causally faithful direction exists for linear probes at any regularization strength.

To show this, we consider a model where the latent variables are plausibly useful, but ultimately unnecessary. If even this setting produces high decodability with all the usual probe-level hallmarks, then no amount of probe-level evidence can rescue the inference from decodability to causal faithfulness.

5.3. Ising model: decodability without a causal direction

The default linear-probe diagnostic suggests the hidden-spin features are represented: probe accuracy improves gradually during training, going from MSE $= 0.0245$ at initialization to MSE $= 0.0002$ after convergence (Figure 6).

But how impressive is this? To check whether the network

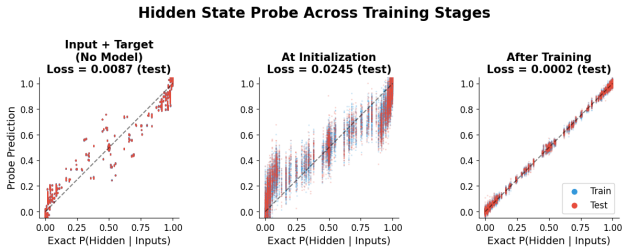

Figure 6. Hidden-spin probe prediction quality on the Ising model (all hidden nodes summed). Left: Baseline linear fit from inputs *and known target spins* directly to hidden spins, without the network (MSE $= 0.0087$). Center: probing the untrained network (MSE $= 0.0245$). Right: probing the trained network (MSE $= 0.0002$). Probe accuracy improves gradually during training and surpasses all baselines.

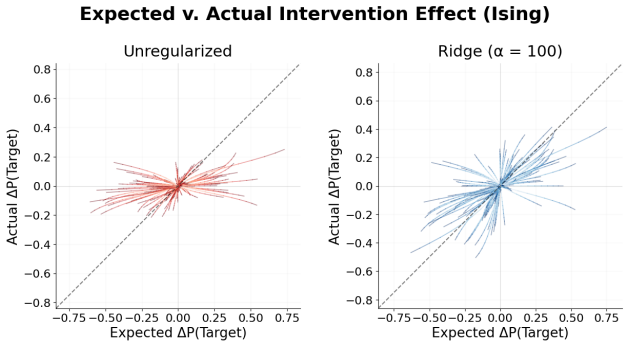

Figure 7. Causal interventions on the Ising model. Left: OLS. Right: ridge regression ($\alpha = 100$, the α with lowest intervention loss). Trajectories do not follow the diagonal at any regularization strength. No probe is causally faithful.

is performing useful computation, we consider a baseline fit that predicts the hidden-spin probabilities from regressing on the training inputs and targets alone, without ever looking at the network's activations. If this baseline performs well, the network might just be retaining input-output information that happens to correlate with the hidden spins, rather than explicitly representing them.

In the Ising case, this baseline achieves MSE $= 0.0087$ on validation data: higher loss than the trained network achieves, so the activations do seem to contain more information about the hidden spins than the raw inputs and targets. By contrast, in the HMM, the same baseline achieves near-perfect decodability, meaning the belief state probes do not reveal anything beyond what is already in the input-sequence-plus-network-output. The Ising probes also generalize to held-out data, which suggests that the model has learned the right function.

By every probe-level diagnostic (probe loss, generalization to held-out data, improvement over baselines, probe accuracy improving gradually during training) the latent Ising features perform at least as well as the HMM belief states. However, ridge probes reveal that Ising features are *not causally faithful*.

Interventions on ridge probes find no evidence of causality across all regularization strengths for the Ising model (Figure 7). The network's output does not change in the way the exact Ising conditional predicts: probes are not causally faithful at any α. The Ising model thus produces the same probe-level evidence as the HMM, or better, for features that have no reason to exist as causal structure. Decodability is not evidence of anything causal.

6. Discussion

As the Ising example shows, regularized regression will not always find causally faithful directions, because sometimes latent variables genuinely are not used by the network. But even when latent variables are causally operative, OLS has a tendency to make use of low-variance directions, and our experiments show that the network does not use these directions for downstream computation. Furthermore, absent explicit causal analysis, no amount of probe-level evidence (probe loss, generalization, probe accuracy improving gradually during training, geometric structure) can distinguish causally faithful directions from causally unfaithful ones. Decodability alone remains inconclusive.

Practical recommendation. When ground-truth interventions are available, we recommend the following protocol: sweep the regularization strength α over several orders of magnitude, run causal interventions at each α, and select the value that minimizes intervention loss. This is what produced our HMM results (Figure 1). To our knowledge, selecting regularization strength by minimizing intervention loss has not been previously proposed in the probing literature. Whether a practical heuristic exists for choosing α without access to ground-truth interventions is an important open question.

Limitations. Our experiments use small transformers in toy settings. However, the core argument, that OLS weights directions by $1/s_i$, is pure linear algebra and applies to any activation space.

Scope of applicability. The $1/s_i$ amplification of low-variance directions explains why OLS points toward directions the network does not use, even when it achieves near-perfect decodability. Most probing in mechanistic interpretability uses classification (logistic regression), where the effect does not have a closed-form SVD characterization. However, the general principle that unpenalized probes can capitalize on low-variance directions applies to any fitting procedure, and the regularization strength sweep methodology should generalize to classification probes, though we

have not tested this.

Connection to prior work. Our work is most directly indebted to Shai et al. (2024), whose HMM transformer and probing methodology we adopt; our contribution is to show that probe directions in this setting are mostly causally faithful—under regularization. Roger (2023) showed that heavy L_2 penalty in logistic regression on binary targets recovers the mean-difference direction, which is more likely to be causal, though through an entirely different mechanism than what this paper proposes.

Implications. If interpretability is to make AI systems safe, it must find the directions the network actually uses. Regularized regression with a regularization strength sweep (when ground-truth interventions are available) is a better starting point than OLS.

Impact Statement

This paper presents work whose goal is to advance the field of mechanistic interpretability. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2126–2136, 2018.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral properties preserved and disrupted. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4459–4480, 2021.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4129–4138, 2019.

Huang, J. et al. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025.

Roger, F. Linear probing for causal features. *LessWrong / Alignment Forum*, 2023. L2 logistic regression on binary targets recovers causal directions.

Shai, A., Teixeira, L., Oldenziel, A., Marzen, S., and Riechers, P. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024.

A. Unweighted Ridge Results

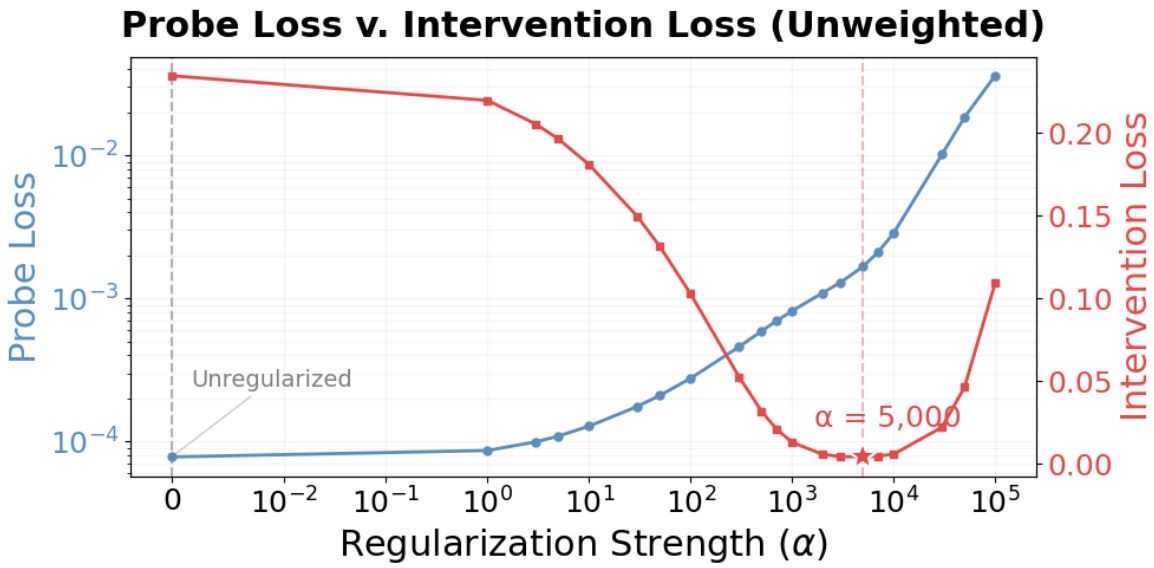

Figure 8. Probe loss and intervention loss vs ridge regularization strength α (unweighted). Intervention-optimal $\alpha = 5{,}000$.

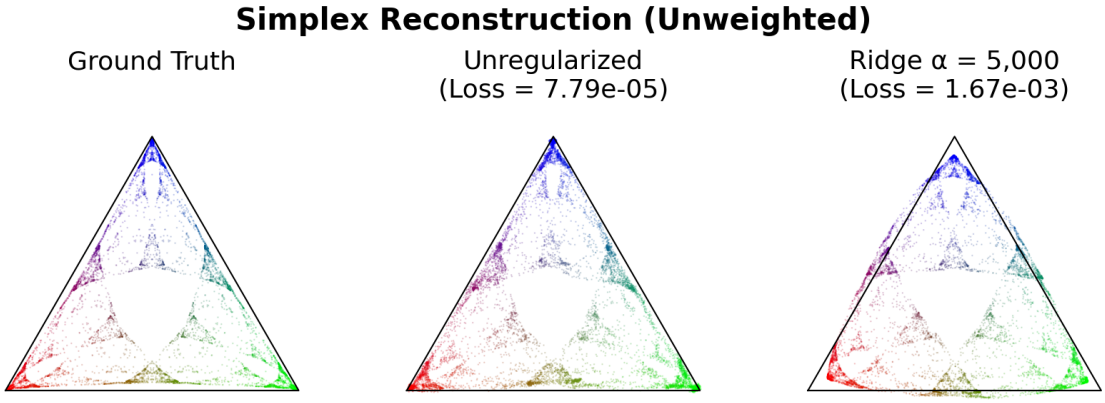

Figure 9. Belief simplex reconstruction (unweighted). Left: ground truth. Center: unregularized (MSE $= 7.79 \times 10^{-5}$). Right: ridge regression, $\alpha = 5{,}000$ (MSE $= 1.67 \times 10^{-3}$).

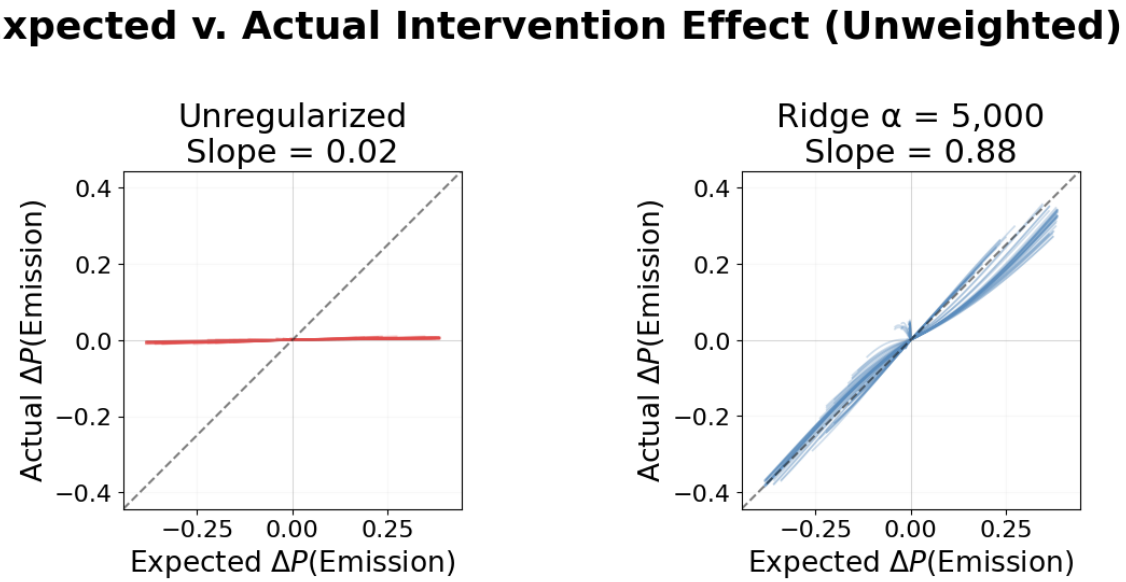

Figure 10. Causal interventions (unweighted). Left: unregularized (slope $= 0.02$). Right: ridge regression, $\alpha = 5{,}000$ (slope $= 0.88$). The core result is identical to the weighted case: unregularized probes are causally unfaithful, ridge probes are approximately causally faithful.

B. Ising Ridge Sweep

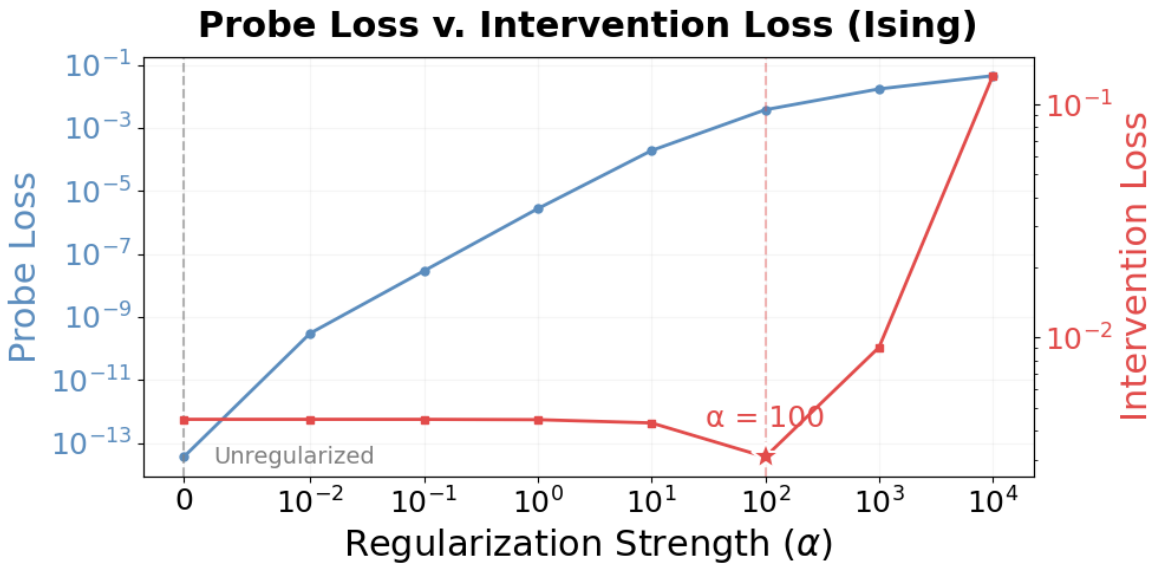

Figure 11. Ising model: probe loss and intervention loss as a function of ridge regularization strength α. Intervention error is minimized around $\alpha = 100$ but remains high at all regularization strengths, confirming that no causally faithful direction exists.