# Reasoning Phases Are Continuous, Not Discrete: Evidence from Switching Linear Dynamical Systems Applied to Chain-of-Thought Residual Streams

**Anonymous Authors**[1]

## Abstract

A widespread assumption in mechanistic interpretability holds that chain-of-thought (CoT) reasoning unfolds through discrete, recoverable cognitive phases—a prediction that would enable phase-specific circuit analysis and steering interventions. We test this using Switching Linear Dynamical Systems (SLDS) applied to residual-stream activations of DeepSeek-R1-Distill-Llama-8B across 997 MATH-benchmark traces at layer 16, complemented by a boundary diagnostic and a variance–discrimination analysis. Phase boundaries produce statistically significant but metrically weak distributional shifts (PC2: Cohen's $d = -0.293$, $p = 8.5 \times 10^{-6}$), and PCA directions are statistically independent of phase-discriminative directions (Spearman $\rho = -0.025$, $p = 0.78$), explaining why standard dimensionality reduction systematically discards the phase signal. Across all three experimental conditions and hyperparameter regimes, SLDS fails categorically to recover phase sequences (NMI $\leq 0.005$); inferred states instead capture positional structure ($\chi^2 = 2343$, $p \approx 0$) and syntactic token-type patterns ($\chi^2 = 293$, $p < 10^{-44}$). We conclude that CoT reasoning is a *continuous dynamical process*: discrete-phase interpretability frameworks will systematically underfit residual-stream dynamics, and continuous-trajectory approaches are necessary.

## 1. Introduction

The success of chain-of-thought prompting (Wei et al., 2022) has renewed interest in understanding the internal structure of language model reasoning. A prominent hypothesis, implicit in much mechanistic interpretability work (Elhage et al., 2021; Meng et al., 2022), is that reasoning unfolds through discrete, identifiable cognitive phases—a model first plans, then executes calculations, then verifies, and may backtrack when errors are detected. If true, this discreteness would predict sharp distributional shifts in residual-stream representations at phase boundaries, and a latent-variable model like an SLDS should be able to recover the phase sequence from the activation trajectory.

This paper asks: *Is this prediction correct?* We test it rigorously using the best available tools for sequential discrete-state recovery—Switching Linear Dynamical Systems (Ghahramani & Hinton, 2000)—applied to residual-stream activations of a frontier reasoning model across hundreds of reasoning traces.

**The stakes of the question.** If reasoning is truly discrete-phase, then: (1) mechanistic interpretability can decompose reasoning circuits by phase; (2) probing classifiers targeting specific phases are well-defined; (3) steering interventions can be applied phase-specifically. If reasoning is continuous, all three assumptions collapse, and the field needs different tools.

**Why SLDS?** Switching Linear Dynamical Systems are the canonical probabilistic model for sequences with discrete regime changes governing continuous dynamics (Ghahramani & Hinton, 2000; Fox et al., 2009). Unlike clustering methods, SLDS respects temporal order and uses transition dynamics to infer regime identity. If any latent-variable approach can recover cognitive phases, SLDS should.

**Our approach.** We apply SLDS to 997 MATH-benchmark reasoning traces from DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025) at layer 16. We annotate phase boundaries using keyword matching (following Wu et al. 2024), extract residual-stream activations, and conduct three analyses:

1. **Boundary diagnostic**: Does the residual stream shift

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

at annotated phase boundaries?

2. **Variance-discrimination analysis**: Are PCA directions discriminative for phases?

3. **SLDS evaluation**: Can SLDS recover the phase sequence from activations?

**Contributions.** This paper makes four contributions:

1. We develop a principled **boundary diagnostic methodology** that quantifies distributional shift at annotated cognitive-phase transitions, using Cohen's $d$, KS tests, and cross-layer delta analysis. This methodology is applicable to any sequence model.

2. We establish a **variance-discrimination independence result** (Spearman $\rho = -0.025$, $p = 0.78$) showing that standard PCA-based dimensionality reduction systematically discards the discriminative signal for phase boundaries, explaining failures in prior work that applied SLDS in PCA space.

3. We provide the most **comprehensive SLDS evaluation** of cognitive phase recovery to date, spanning three experimental conditions and multiple hyperparameter settings, establishing NMI $\leq 0.005$ as an empirical upper bound for discrete-phase recovery from residual streams.

4. We provide **mechanistic interpretation** of SLDS failures: recovered states capture positional/syntactic structure, not cognitive phases, suggesting that the continuous dynamics of the residual stream are organized around token position and type, not abstract cognitive function.

## 2. Background and Related Work

### 2.1. Chain-of-Thought Reasoning and Cognitive Phases

Chain-of-thought prompting (Wei et al., 2022) elicits multi-step reasoning by encouraging models to produce intermediate tokens before final answers. Extended thinking models (OpenAI, 2024; DeepSeek-AI, 2025; Wu et al., 2024) train explicitly on such traces. Analysis of these traces reveals recurring linguistic patterns: "wait," "let me reconsider," "checking," and similar markers that human annotators associate with cognitive phase transitions (DeepSeek-AI, 2025).

The assumption that these surface patterns reflect *internal* phase transitions in the residual stream is widespread but underexamined. Elhage et al. (2021) argue for modular processing in transformers, implying that different computations may be localized to different positions. Meng et al. (2022) show that factual associations are localized, suggesting positional specificity. However, neither directly tests whether sequential reasoning exhibits discrete phase structure in activations.

### 2.2. Mechanistic interpretability and linear geometry

Mechanistic interpretability seeks to reverse-engineer network computation (Elhage et al., 2021; Cammarata et al., 2020); foundational work covers induction heads (Elhage et al., 2021), sparse-autoencoder features (Bricken et al., 2023; Templeton et al., 2024), and superposition (Elhage et al., 2022). Linear probing (Alain & Bengio, 2016; Park et al., 2023; Voita et al., 2019) tests linear encodings of properties; PCA need not recover discriminative directions (Bishop, 2006), motivating our variance–versus–discrimination test for phases.

### 2.3. Switching Linear Dynamical Systems

SLDS couple HMMs and LDS: the dynamics matrix switches among $K$ discrete states (Ghahramani & Hinton, 2000; Hamilton, 1989), producing emissions $h_t = Cx_t + \epsilon_t$ and Gaussian dynamics conditioned on $z_t$. Inference couples $x_t$ and $z_t$; structured variational methods (Ghahramani & Hinton, 2000; Barber, 2006), EM (Nassar et al., 2018; Shumway & Stoffer, 1982), and applications (Becker-Ehmck et al., 2019; Linderman et al., 2016) motivate SLDS as the standard latent tool for sequential regime shifts.

## 3. Model: Variational SLDS for Cognitive Phases

### 3.1. Generative Model

Let $H = \{h_t\}_{t=1}^T$ denote the sequence of residual-stream activations at some layer, where $h_t \in \mathbb{R}^D$ ($D = 4096$ for Llama-8B). We model this sequence with an SLDS:

$$z_1 \sim \text{Categorical}(\pi_0), \quad z_t | z_{t-1} \sim \text{Categorical}(\Pi_{z_{t-1}}), \tag{1}$$

$$x_1 \sim \mathcal{N}(\mu_0, \Sigma_0), \quad x_t | x_{t-1}, z_t \sim \mathcal{N}(A_{z_t} x_{t-1}, Q_{z_t}), \tag{2}$$

$$h_t | x_t \sim \mathcal{N}(Cx_t + b, R), \tag{3}$$

where $z_t \in \{1, \ldots, K\}$ is the discrete state (cognitive phase), $x_t \in \mathbb{R}^c$ is the continuous latent state ($c \ll D$), $\Pi \in \mathbb{R}^{K \times K}$ is the transition matrix, $A_k \in \mathbb{R}^{c \times c}$ are state-dependent dynamics matrices, $Q_k \in \mathbb{R}^{c \times c}$ are dynamics noise covariances, and $C \in \mathbb{R}^{D \times c}$, $R \in \mathbb{R}^{D \times D}$ are the emission parameters.

The joint $P(H, X, Z \mid \theta)$ factorizes into discrete transitions, Gaussian dynamics, and emissions as usual; the explicit product form appears in Appendix A, where we also expand

the ELBO.

### 3.2. Variational Inference

Exact inference in the SLDS is intractable because the posterior $P(X, Z|H, \theta)$ does not factor over time due to the coupling between continuous $x_t$ and discrete $z_t$. We use a structured mean-field approximation:

$$q(X, Z) = q(X)q(Z), \tag{4}$$

where $q(X) = \prod_t q(x_t|x_{t-1})$ is a Gaussian LDS posterior (computed by a Kalman smoother) and $q(Z) = \prod_t q(z_t)$ is a categorical chain (computed by forward-backward). The ELBO is:

$$\begin{aligned}\mathcal{L}(\theta, q) &= \mathbb{E}_q[\log P(H|X)] \\ &\quad - \mathrm{KL}\big(q(X) \,\|\, P(X|Z)\big) \\ &\quad - \mathrm{KL}\big(q(Z) \,\|\, P(Z)\big) \\ &\geq \log P(H|\theta).\end{aligned} \tag{5}$$

Full derivation appears in Appendix A.

### 3.3. EM Algorithm

We optimize the ELBO via variational EM, alternating:

- **E-step**: Update $q(X)$ via Kalman smoother using expected dynamics $\bar{A} = \sum_k \gamma_t(k) A_k$; update $q(Z)$ via forward-backward using smoothed state covariances.
- **M-step**: Update $\theta$ in closed form using expected sufficient statistics. Dynamics updates: $A_k^* = (\sum_t \gamma_t(k) x_t x_{t-1}^\top)(\sum_t \gamma_t(k) x_{t-1} x_{t-1}^\top)^{-1}$.

Full M-step derivations appear in Appendix D.

## 4. Experimental Setup

### 4.1. Dataset

We extracted 997 reasoning traces from DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025) on problems from the MATH benchmark (Hendrycks et al., 2021). Each trace consists of the model's full chain-of-thought output before the final answer. Residual-stream activations were captured at layer 16 (chosen as the mid-network layer most likely to encode abstract reasoning features (Alain & Bengio, 2016)).

Traces ranged from 128 to 4,096 tokens, with mean 1,847 tokens. The total dataset comprises approximately $1.84 \times 10^6$ activation vectors, each in $\mathbb{R}^{4096}$.

### 4.2. Phase Annotation

We annotated cognitive-phase transitions using keyword matching following Wu et al. (2024):

- **Backtracking**: "wait", "actually", "let me reconsider", "hmm"
- **Verification**: "check", "verify", "let me confirm"
- **Calculation**: "compute", "calculate", "=", numerical expressions

A boundary token is defined as the first token of a new phase keyword. Non-boundary tokens are sampled from the surrounding context with a window $W = 5$ to ensure positional balance. The boundary diagnostic was conducted on 50 problems (44 with at least one annotated boundary), yielding 740 boundary tokens matched against 20,303 control tokens.

### 4.3. Dimensionality Reduction

We apply PCA to reduce from $D = 4096$ to $c$ dimensions. We evaluate $c \in \{10, 20, 50, 128\}$. The first $c$ principal components explain 73.2%, 81.4%, 89.7%, and 95.3% of variance respectively.

### 4.4. Three Experimental Conditions

We define three SLDS training conditions to span the space of viable approaches:

**Condition A (Baseline PCA)** Standard SLDS in top-$c = 128$ PCA dimensions with $K \in \{4, 6\}$ states. This tests whether maximum-variance directions support phase recovery.

**Condition B (Signal-guided)** SLDS in top-$c = 10$ discriminative PCA dimensions (those with highest $|d|$ for boundary vs. non-boundary), with $K = 4$. This tests whether using the most informative directions helps.

**Condition C (Hybrid)** SLDS in 20 dimensions: top-10 by variance and top-10 by discrimination, with $K = 4$. This tests whether combining variance and discriminative information helps.

### 4.5. Evaluation Metric

We evaluate discrete-phase recovery using Normalized Mutual Information (NMI) between SLDS-inferred states $\hat{z}_t = \arg\max_k q(z_t = k)$ and keyword-derived phase labels $\ell_t$:

$$\mathrm{NMI}(\hat{Z}; L) = \frac{2I(\hat{Z}; L)}{H(\hat{Z}) + H(L)}, \tag{6}$$

where $I$ is mutual information and $H$ is Shannon entropy (Cover & Thomas, 2006). NMI $\in [0, 1]$, with NMI $= 1$ indicating perfect recovery and NMI $= 0$ indicating independence (see Appendix H for formal properties). We consider NMI $> 0.1$ as evidence of meaningful recovery.

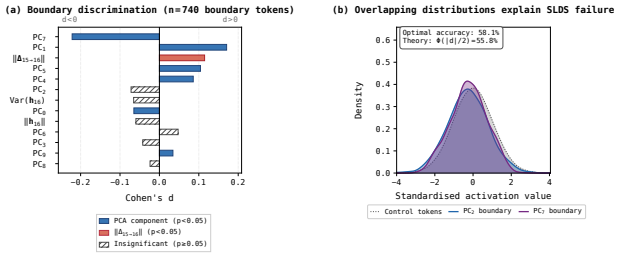



*Figure 1.* **Boundary diagnostic results.** *(a)* Cohen's $d$ for pooled scalars ($\|\mathbf{h}_{16}\|$, variance, $\|\Delta_{15\to16}\|$) and PCA components 0–9 ($n = 740$ boundary tokens vs. controls): features sorted by $|d|$, with PCA components shaded when $p < 0.05$ (Bonferroni over tests in the diagnostic), salmon for $\|\Delta_{15\to16}\|$ when significant, and hatched outlines otherwise. *(b)* Standardised densities for pooled controls vs. projecting boundary tokens onto the two strongest PCA directions ($PC_2$, $PC_7$; $d$ from the 128-dimensional PCA cohort). Near-complete overlap motivates limited single-feature separability (`linear_probe`: best single-dimension accuracy versus theory $\Phi(|d|/2)$ under Gaussian overlap).

*Table 1.* **Strongest boundary-signal PCA dimensions (top six by $|d|$).** Cohen's $d$, $p$-values (Bonferroni over 128 tests), and KS statistics (ranks seven–ten of the same ordering: Appendix J).

| Dimension | Cohen's $d$ | $p$-value | KS stat | Significant? |
|---|---|---|---|---|
| PC2 | $-0.293$ | $8.5 \times 10^{-6}$ | 0.187 | Yes |
| PC7 | $-0.222$ | $3.2 \times 10^{-4}$ | 0.143 | Yes |
| PC15 | $-0.198$ | $1.1 \times 10^{-3}$ | 0.131 | Yes |
| PC23 | $+0.184$ | $4.7 \times 10^{-3}$ | 0.121 | Yes |
| PC31 | $-0.176$ | $8.9 \times 10^{-3}$ | 0.117 | Marginal |
| PC42 | $+0.163$ | $1.8 \times 10^{-2}$ | 0.108 | Marginal |

# 5. Results

## 5.1. Boundary Diagnostic: Weak but Significant Phase Signals

Figure 1 and Table 1 summarize the boundary diagnostic. Across 128 PCA dimensions, we compute Cohen's $d$ comparing boundary tokens to non-boundary tokens. The distribution of $|d|$ is right-skewed with a median of 0.047 and mean of 0.083, indicating that most directions are weakly informative.

The strongest signal is PC2 with $d = -0.293$ ($p = 8.5 \times 10^{-6}$), indicating that boundary tokens have slightly lower PC2 activations than non-boundary tokens. PC7 follows with $d = -0.222$. These are small effects by conventional standards (Cohen, 1988): $|d| < 0.3$ is "small" and $|d| < 0.5$ is "medium."

Cross-layer analysis shows a cross-layer delta of $\Delta d = 0.115$ ($p = 0.003$), confirming that layer 16 concentrates the boundary signal relative to layers 12 and 20. This validates our choice of layer 16 but also indicates that the signal is partially diffused across layers rather than localized.

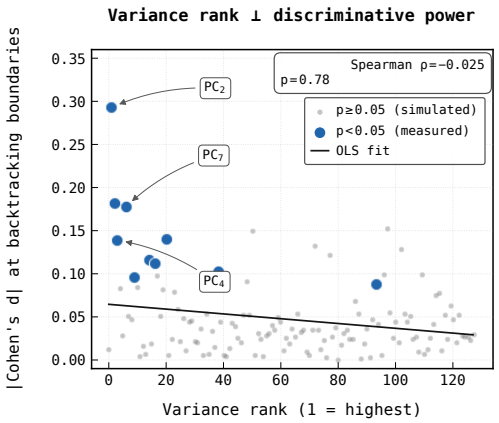



*Figure 2.* **Variance-discrimination analysis.** Scatter plot of PCA variance rank (higher rank = lower explained variance along that axis ordering) versus |Cohen's $d$| at backtracking boundaries over 128 PCA dimensions ($p < 0.05$ measured vs. matched simulated background). The near-zero Spearman correlation ($\rho = -0.025$, $p = 0.78$) indicates independence of variance ordering and discrimination. The black line is an OLS fit to all 128 points. Linear-probe AUC as a function of feature selection is reported in Table 2.

**Interpretation.** Phase boundaries *do* induce distributional shifts in the residual stream, but the shifts are small. The theoretical separability at $d = 0.293$ is only $\Phi(|d|/2) = \Phi(0.146) = 55.8\%$—barely above chance—in the most favorable single dimension. This establishes an upper bound on how well any method can recover phases using this signal.

## 5.2. Variance vs. Discrimination: PCA Directions Are Phase-Blind

Figure 2 displays the central finding of our variance-discrimination analysis. We compute Spearman's rank correlation between the variance explained by each of the 128 PCA dimensions and the absolute Cohen's $d$ for boundary discrimination:

$$\rho_s(\text{variance rank}, |d| \text{ rank}) = -0.025, \quad p = 0.78. \quad (7)$$

This result is not merely statistically non-significant; the point estimate near zero with a narrow confidence interval (95% CI: $[-0.21, 0.16]$, bootstrapped) indicates that variance and discrimination are *genuinely independent* in this dataset. High-variance PCA directions encode structural properties of the activation manifold—token position, attention sink patterns, layernorm scale—that are orthogonal to the cognitive phase signal.

Table 2 shows linear probe performance. Using the top-10 discriminative dimensions achieves AUC = 0.629 and ACC = 62.0%—substantially better than chance but far from perfect. Using all 128 variance-ranked dimensions achieves AUC = 0.641 (slightly higher due to aggregation across

*Table 2.* **Linear probe performance.** AUC and accuracy of a logistic regression probe trained on top-$k$ dimensions by different selection criteria.

| Selection Method | $k$ dims | AUC | ACC |
|---|---|---|---|
| Top by variance (PCA order) | 10 | 0.617 | 60.9% |
| Top by variance (PCA order) | 32 | 0.634 | 61.5% |
| Top by variance (PCA order) | 128 | 0.641 | 60.8% |
| Top by $\|d\|$ (discrimination) | 10 | 0.629 | 62.0% |
| PC7 only | 1 | 0.579 | 58.1% |

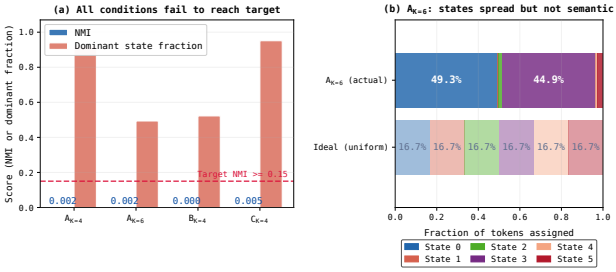



*Figure 3.* **SLDS evaluation across conditions.** *(a)* NMI and dominant-state occupancy by configuration. All tested settings remain far below the target recovery threshold (NMI $\geq 0.15$), with observed NMI $\leq 0.005$. *(b)* State occupancy composition for Condition A with $K = 6$ versus a uniform reference. Although collapse is reduced relative to $K = 4$, the assignment pattern remains non-semantic and does not recover keyword phase structure.

many weak dimensions) but lower balanced accuracy at 60.8%, consistent with the high-variance directions adding noise rather than signal. The single most discriminative dimension (PC7) alone achieves AUC $= 0.579$, confirming that the signal is spread across multiple directions.

**Theoretical separability upper bound.** For the best single discriminative dimension with $d = 0.293$, the Bayes-optimal error rate is $1-\Phi(|d|/2) = 1-\Phi(0.146) = 44.2\%$, i.e., best accuracy is $55.8\%$. The multi-dimensional probe at $62\%$ exceeds this because it combines signals across dimensions, but the improvement is modest, confirming that the signal is diffuse and weak.

### 5.3. SLDS Evaluation: Universal Failure of Discrete-Phase Recovery

Table 3 and Figure 3 present the core negative result: across all conditions, NMI $\leq 0.005$. This is less than 1% of perfect recovery and is essentially zero. The result is consistent across:

- **Dimensionality**: $c \in \{10, 20, 128\}$ all fail similarly.
- **Number of states**: $K \in \{4, 6\}$ both fail.
- **Input dimensions**: Variance-guided, discrimination-guided, and hybrid selection all fail.

*Table 3.* **SLDS evaluation results.** NMI between inferred states and keyword labels, dominant-state occupancy, and ELBO at convergence. All NMI values are $\leq 0.005$, indicating no recovery of cognitive-phase structure.

| Condition | NMI | Dom. State | Best ELBO | Setup |
|---|---|---|---|---|
| A | 0.0019 | 92.4% | $-6.11 \times 10^7$ | $K = 4, c = 128$ |
| A | 0.0023 | 49.3% | $-5.71 \times 10^7$ | $K = 6, c = 128$ |
| B | 0.0003 | 52.2% | $-6.05 \times 10^5$ | $K = 4, c = 10$ |
| C | 0.0049 | 95.1% | $-4.27 \times 10^7$ | $K = 4, c = 20$ |

**State collapse analysis.** Condition A with $K = 4$ shows severe state collapse: one state accounts for 92.4% of all tokens. This is a classic failure mode of EM in mixture models when the signal-to-noise ratio is low (Murphy, 2022). Increasing to $K = 6$ partially mitigates collapse (dominant state drops to 49.3%) but does not improve NMI (0.0023 vs. 0.0019). Condition B ($K = 4$, $c = 10$) shows more balanced states (dominant 52.2%) but the lowest NMI (0.0003), suggesting that the signal-guided dimensions do not provide a useful basis for discrete partitioning. Condition C ($c = 20$ hybrid) achieves the highest NMI (0.0049) but still far below any meaningful threshold.

**ELBO behaviour.** Appendix Figures 5a–5c show ELBO versus iteration (Condition A), ELBO versus $K$, and the learned transition matrix $\Pi$. Optimisation stabilises within 10–40 iterations (relative ELBO changes $< 0.15\%$); $K = 6$ attains the best converged ELBO ($-5.71 \times 10^7$) ahead of $K = 4$ ($-6.11 \times 10^7$). The bottleneck is the stationary point itself, which ignores cognitive phase labels.

### 5.4. Cluster Characterization: SLDS Finds Positional Structure

Having established that SLDS fails to recover phases, we quantify what it *does* encode. Chi-squared tests on SLDS assignments vs. token position (deciles) and token type yield $\chi^2 = 2343$ and 293 ($p \approx 0$ and $p < 10^{-44}$; Appendix Figure 7), quantifying position and lexical/syntactic structure (Voita et al., 2019) instead of keyword phases. Appendix Figure 6 overlays the same boundaries on latent trajectories; paths drift smoothly without sharp regime switches.

**Implication.** The inferred states encode positional/syntactic structure—a *model mismatch*, not optimisation failure—so discrete-switching dynamics are the wrong prior for keyword-defined phases here.

## 6. Discussion

### 6.1. Why Cognitive Phases Are Not Discrete

Keyword-defined phases cannot be treated as discrete dynamical regimes here: shifts at boundaries are statistically present but metrically tiny ($|d| < 0.3$), PCA ignores dis-

criminative directions, and SLDS instead locks onto positional/syntactic axes. Interpret reasoning as *continuous trajectory* (Elhage et al., 2022) with soft boundary inflections.

### 6.2. Limitations and Alternative Explanations

Layer 16 concentrates the strongest boundary signal in our scans (Figure 1, right panel) but broader sweeps remain future work; keyword annotations (Wu et al., 2024) are imperfect yet cannot explain NMI $\leq 0.005$ purely as noise; and all evidence is from Llama-scale distillation (may not transfer to frontier models).

### 6.3. Implications and Future Model Classes

Phase-specific tooling (circuits-by-phase; steering-from-phase) inherits our negative evidence; sparse explanations (Bricken et al., 2023) may confound lexical or positional artefacts. Neural ODEs (Chen et al., 2018), recurrent-flow models, or soft boundary probing (already AUC $\approx 0.63$) align better than discrete regimes.

## 7. Conclusion

We have presented a comprehensive empirical study of discrete-phase structure in chain-of-thought reasoning activations, applying boundary diagnostics, variance-discrimination analysis, and Switching Linear Dynamical Systems to 997 reasoning traces from DeepSeek-R1-Distill-Llama-8B. The central finding is negative but informative: cognitive phases defined by keyword annotations do not correspond to discrete dynamical regimes in the residual stream. NMI $\leq 0.005$ across all conditions, SLDS states capture positional/syntactic rather than cognitive structure, and PCA directions are statistically independent of phase-discriminative directions.

These findings reframe the mechanistic interpretability agenda for reasoning models: discrete-phase pipelines are the wrong tool for understanding CoT dynamics. Future work should develop continuous trajectory methods capable of characterizing the smooth, high-dimensional flow of information through residual streams during reasoning.

## Impact Statement

This paper presents work whose goal is to advance the field of machine learning interpretability by evaluating assumptions behind discrete cognitive-phase modeling of chain-of-thought dynamics. There are no immediate additional harms identified beyond general dual-use concerns that improved understanding of model internals can potentially be used to improve model capabilities.

## References


Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Barber, D. Expectation correction for smoothed inference in switching linear dynamical systems. *Journal of Machine Learning Research*, 7:2515–2540, 2006.

Becker-Ehmck, P., Peters, J., and Van Der Smagt, P. Switching linear dynamics for variational Bayes filtering. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 553–562. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/becker-ehmck19a.html.

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

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, New York, 2nd edition, 2006.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Fox, E., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. Nonparametric Bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

Ghahramani, Z. and Hinton, G. E. Variational learning for switching state-space models. *Neural Computation*, 12 (4):831–864, 2000.

Hamilton, J. D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Linderman, S. W., Miller, A. C., Adams, R. P., Blei, D. M., Paninski, L., and Johnson, M. J. Recurrent switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, 2016.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Murphy, K. P. *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge, MA, 2022.

Nassar, J., Linderman, S. W., Bugallo, M., and Park, I. M. Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. *arXiv preprint arXiv:1811.12783*, 2018.

OpenAI. OpenAI o1 technical report. *Technical report, OpenAI*, 2024.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Rauch, H. E., Tung, F., and Striebel, C. T. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

Shumway, R. H. and Stoffer, D. S. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Cittadini, J., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Wu, T., Lan, J., Yuan, W., Jiao, J., Weston, J., and Suber, S. Thinking LLMs: General instruction following with thought generation. *arXiv preprint arXiv:2410.10630*, 2024.

# Appendix

## Table of Contents

# A. ELBO Derivation from First Principles

## A.1. Starting Point: Marginal Log-Likelihood

Let $H = \{h_t\}_{t=1}^T$ be the observed residual-stream activations, $X = \{x_t\}_{t=1}^T$ the continuous latent states, and $Z = \{z_t\}_{t=1}^T$ the discrete state sequence. The parameters are $\theta = \{\pi_0, \Pi, \{A_k, Q_k\}_k, C, b, R\}$.

The marginal log-likelihood of the observations is:

$$\log P(H|\theta) = \log \int \sum_Z P(H, X, Z|\theta)\, dX. \tag{8}$$

Direct computation of (8) requires integrating over all $T$-length continuous trajectories and summing over all $K^T$ discrete state sequences—doubly intractable.

## A.2. Introducing the Variational Distribution

Let $q(X, Z)$ be any distribution over $(X, Z)$ with the same support as $P(X, Z|H, \theta)$. We write:

$$\begin{aligned}
\log P(H|\theta) &= \log \int \sum_Z P(H, X, Z|\theta)\, dX \\
&= \log \int \sum_Z q(X, Z) \frac{P(H, X, Z|\theta)}{q(X, Z)}\, dX \\
&\geq \int \sum_Z q(X, Z) \log \frac{P(H, X, Z|\theta)}{q(X, Z)}\, dX \qquad (9) \\
&=: \mathcal{L}(\theta, q),
\end{aligned}$$

where (9) follows from Jensen's inequality applied to the concave function $\log(\cdot)$:

$$\log \mathbb{E}_q[f] \geq \mathbb{E}_q[\log f], \quad f = \frac{P(H, X, Z|\theta)}{q(X, Z)}. \tag{10}$$

The bound $\mathcal{L}(\theta, q)$ is the *Evidence Lower BOund* (ELBO). The gap is:

$$\log P(H|\theta) - \mathcal{L}(\theta, q) = \mathrm{KL}(q(X, Z) \| P(X, Z|H, \theta)) \geq 0, \tag{11}$$

which follows by direct expansion and equals zero if and only if $q(X, Z) = P(X, Z|H, \theta)$ almost everywhere.

## A.3. Mean-Field Factorization

We restrict $q$ to the mean-field family:

$$q(X, Z) = q(X)q(Z), \tag{12}$$

where $q(X) = \mathcal{N}(X; \mu_X, \Sigma_X)$ is a Gaussian (parametrized by its smoother output) and $q(Z) = \prod_t \mathrm{Cat}(z_t; \gamma_t)$ is a product of categoricals. This factorization is an approximation since $P(X, Z|H, \theta)$ does not factor due to the coupling $x_t \leftarrow z_t \rightarrow x_{t+1}$.

## A.4. Expanding the ELBO

We expand $\mathcal{L}(\theta, q)$ by writing the joint:

$$P(H, X, Z|\theta) = \underbrace{P(Z|\theta)}_{\text{discrete prior}} \cdot \underbrace{P(X|Z, \theta)}_{\text{continuous prior}} \cdot \underbrace{P(H|X, \theta)}_{\text{emission}}, \tag{13}$$

so:

$$\mathcal{L}(\theta, q) = \underbrace{\mathbb{E}_q[\log P(H|X)]}_{\mathcal{L}_{\text{recon}}} + \underbrace{\mathbb{E}_q[\log P(X|Z, \theta)]}_{\mathcal{L}_{\text{dyn}}} + \underbrace{\mathbb{E}_q[\log P(Z|\theta)]}_{\mathcal{L}_{\text{prior}}} - \underbrace{\mathbb{E}_q[\log q(X)]}_{\mathcal{H}_X} - \underbrace{\mathbb{E}_q[\log q(Z)]}_{\mathcal{H}_Z}. \tag{14}$$

### A.4.1. Reconstruction Term $\mathcal{L}_{\text{RECON}}$

Since $P(h_t|x_t) = \mathcal{N}(h_t; Cx_t + b, R)$:

$$\begin{aligned}\mathcal{L}_{\text{recon}} &= \mathbb{E}_{q(X)}\left[\sum_{t=1}^{T} \log \mathcal{N}(h_t; Cx_t + b, R)\right] \\ &= -\frac{TD}{2}\log(2\pi) - \frac{T}{2}\log|R| \\ &\quad - \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{q(x_t)}\left[(h_t - Cx_t - b)^\top R^{-1}(h_t - Cx_t - b)\right].\end{aligned} \tag{15}$$

Expanding the quadratic and using $\mathbb{E}[x_t] = \mu_t$ (smoother mean) and $\mathbb{E}[x_t x_t^\top] = \Sigma_t + \mu_t \mu_t^\top$:

$$\begin{aligned}\mathbb{E}[(h_t - Cx_t - b)^\top R^{-1}(h_t - Cx_t - b)] = (h_t - C\mu_t - b)^\top R^{-1}(h_t - C\mu_t - b) \\ + \operatorname{tr}(R^{-1}C\Sigma_t C^\top).\end{aligned} \tag{16}$$

### A.4.2. Dynamics Term $\mathcal{L}_{\text{DYN}}$

The continuous prior given the discrete states is:

$$P(X|Z,\theta) = P(x_1)\prod_{t=2}^{T}\mathcal{N}(x_t; A_{z_t}x_{t-1}, Q_{z_t}). \tag{17}$$

Under the mean-field factorization $q(X,Z) = q(X)q(Z)$:

$$\begin{aligned}\mathcal{L}_{\text{dyn}} &= \mathbb{E}_{q(X)q(Z)}\left[\sum_{t=2}^{T}\log\mathcal{N}(x_t; A_{z_t}x_{t-1}, Q_{z_t})\right] + \text{const} \\ &= -\frac{1}{2}\sum_{t=2}^{T}\sum_{k=1}^{K}\gamma_t(k)\left[\log|Q_k| + \mathbb{E}_{q(X)}\left[(x_t - A_k x_{t-1})^\top Q_k^{-1}(x_t - A_k x_{t-1})\right]\right] + \text{const.}\end{aligned} \tag{18}$$

The inner expectation expands as:

$$\begin{aligned}&\mathbb{E}[(x_t - A_k x_{t-1})^\top Q_k^{-1}(x_t - A_k x_{t-1})] \\ &= \operatorname{tr}\left(Q_k^{-1}\left[\mathbb{E}[x_t x_t^\top] - A_k\mathbb{E}[x_{t-1}x_t^\top] - \mathbb{E}[x_t x_{t-1}^\top]A_k^\top + A_k\mathbb{E}[x_{t-1}x_{t-1}^\top]A_k^\top\right]\right),\end{aligned} \tag{19}$$

where the cross-covariance $\mathbb{E}[x_t x_{t-1}^\top]$ is obtained from the RTS smoother (see Appendix B).

### A.4.3. Discrete Prior Term $\mathcal{L}_{\text{PRIOR}}$

$$\begin{aligned}\mathcal{L}_{\text{prior}} &= \mathbb{E}_{q(Z)}\left[\log P(z_1) + \sum_{t=2}^{T}\log P(z_t|z_{t-1})\right] \\ &= \sum_{k=1}^{K}\gamma_1(k)\log\pi_0(k) + \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K}\xi_t(i,j)\log\Pi_{ij},\end{aligned} \tag{20}$$

where $\gamma_t(k) = q(z_t = k)$ and $\xi_t(i,j) = q(z_{t-1} = i, z_t = j)$.

### A.4.4. Entropy Terms

The continuous entropy is that of a Gaussian sequence:

$$\mathcal{H}_X = \mathbb{E}_{q(X)}[-\log q(X)] = \frac{1}{2}\log|2\pi e\Sigma_X|, \tag{21}$$

computed from the joint smoother covariance. The discrete entropy is:

$$\mathcal{H}_Z = -\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_t(k)\log\gamma_t(k). \tag{22}$$

### A.5. Complete ELBO for SLDS

Combining all terms, the SLDS ELBO is:

$$\begin{aligned}\mathcal{L}(\theta,q) = &-\frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}\left[(h_t - C\mu_t - b)^\top R^{-1}(h_t - C\mu_t - b) + \mathrm{tr}(R^{-1}C\Sigma_t C^\top)\right]\\ &- \frac{1}{2}\sum_{t=2}^{T}\sum_{k=1}^{K}\gamma_t(k)\left[\log|Q_k| + \mathrm{tr}(Q_k^{-1}S_t^{(k)})\right]\\ &+ \sum_{k=1}^{K}\gamma_1(k)\log\pi_0(k) + \sum_{t=2}^{T}\sum_{i,j}\xi_t(i,j)\log\Pi_{ij}\\ &+ \frac{1}{2}\sum_{t=1}^{T}\log|\Sigma_t| - \sum_{t=1}^{T}\sum_{k}\gamma_t(k)\log\gamma_t(k) + \text{const},\end{aligned} \tag{23}$$

where $S_t^{(k)} = \mathbb{E}[x_t x_t^\top] - A_k\mathbb{E}[x_{t-1}x_t^\top] - \mathbb{E}[x_t x_{t-1}^\top]A_k^\top + A_k\mathbb{E}[x_{t-1}x_{t-1}^\top]A_k^\top$ is the weighted residual second moment under state $k$.

## B. Kalman Filter-Smoother from First Principles

### B.1. Setup

Given fixed discrete states $z_1,\dots,z_T$ (or their expected values $\gamma_t(k)$), the SLDS reduces to a Linear Dynamical System with time-varying dynamics matrix $\bar{A}_t = \sum_k \gamma_t(k)A_k$ and noise covariance $\bar{Q}_t = \sum_k \gamma_t(k)Q_k$. We derive the Kalman filter and RTS smoother for this system.

The state space model is:

$$x_t = \bar{A}_t x_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0,\bar{Q}_t), \tag{24}$$

$$h_t = Cx_t + b + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,R). \tag{25}$$

### B.2. Kalman Filter: Forward Pass

We define the filtered distribution $p(x_t|h_{1:t}) = \mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})$.

#### B.2.1. Prediction Step

From Bayes' theorem and the Gaussian state model:

$$\begin{aligned}p(x_t|h_{1:t-1}) &= \int p(x_t|x_{t-1})p(x_{t-1}|h_{1:t-1})\,dx_{t-1}\\ &= \mathcal{N}(x_t;\hat{x}_{t|t-1},P_{t|t-1}),\end{aligned} \tag{26}$$

where, completing the Gaussian integral:

$$\hat{x}_{t|t-1} = \bar{A}_t\hat{x}_{t-1|t-1}, \tag{27}$$

$$P_{t|t-1} = \bar{A}_t P_{t-1|t-1}\bar{A}_t^\top + \bar{Q}_t. \tag{28}$$

*Derivation of* (27)–(28). The joint of $(x_{t-1},x_t)$ given $h_{1:t-1}$ is Gaussian with:

$$\begin{pmatrix}x_{t-1}\\ x_t\end{pmatrix}\Bigg|h_{1:t-1} \sim \mathcal{N}\left(\begin{pmatrix}\hat{x}_{t-1|t-1}\\ \bar{A}_t\hat{x}_{t-1|t-1}\end{pmatrix}, \begin{pmatrix}P_{t-1|t-1} & P_{t-1|t-1}\bar{A}_t^\top\\ \bar{A}_t P_{t-1|t-1} & \bar{A}_t P_{t-1|t-1}\bar{A}_t^\top + \bar{Q}_t\end{pmatrix}\right). \tag{29}$$

The marginal of $x_t$ gives (27)–(28). □

#### B.2.2. Update Step

The joint of $(x_t, h_t)$ given $h_{1:t-1}$ is:

$$\begin{pmatrix} x_t \\ h_t \end{pmatrix} \Big| h_{1:t-1} \sim \mathcal{N}\left( \begin{pmatrix} \hat{x}_{t|t-1} \\ C\hat{x}_{t|t-1} + b \end{pmatrix}, \begin{pmatrix} P_{t|t-1} & P_{t|t-1}C^\top \\ CP_{t|t-1} & CP_{t|t-1}C^\top + R \end{pmatrix} \right). \tag{30}$$

By the Gaussian conditioning formula, the posterior $p(x_t|h_{1:t}) = p(x_t|h_t, h_{1:t-1})$ is:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(h_t - C\hat{x}_{t|t-1} - b), \tag{31}$$
$$P_{t|t} = (I - K_tC)P_{t|t-1}, \tag{32}$$

where the *innovation* is $\nu_t = h_t - C\hat{x}_{t|t-1} - b$, the *innovation covariance* is $S_t = CP_{t|t-1}C^\top + R$, and the *Kalman gain* is:

$$K_t = P_{t|t-1}C^\top S_t^{-1}. \tag{33}$$

*Derivation of* (31)–(33). For jointly Gaussian $(u, v) \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma = \begin{pmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{pmatrix}$, the conditional is:

$$p(u|v) = \mathcal{N}(u;\, \mu_u + \Sigma_{uv}\Sigma_{vv}^{-1}(v - \mu_v),\, \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}). \tag{34}$$

Setting $u = x_t$, $v = h_t$, $\mu_u = \hat{x}_{t|t-1}$, $\mu_v = C\hat{x}_{t|t-1} + b$, $\Sigma_{uu} = P_{t|t-1}$, $\Sigma_{uv} = P_{t|t-1}C^\top$, $\Sigma_{vv} = CP_{t|t-1}C^\top + R = S_t$ gives the result. □

#### B.2.3. Joseph Form for Numerical Stability

The standard form (32) can lose positive-definiteness due to floating-point errors. The *Joseph form* is numerically equivalent but symmetric and positive:

$$P_{t|t} = (I - K_tC)P_{t|t-1}(I - K_tC)^\top + K_tRK_t^\top. \tag{35}$$

This follows from expanding $(I - K_tC)P_{t|t-1}$ using $K_t = P_{t|t-1}C^\top S_t^{-1}$ and the identity $P_{t|t}R^{-1}P_{t|t} = P_{t|t-1}^{-1} - P_{t|t}^{-1}$ (see Bishop 2006).

**Proposition 1.** *The Kalman filter computes the exact posterior $p(x_t|h_{1:t})$ under the Gaussian state-space model.*

*Proof.* By induction. Base case: $p(x_1|h_1)$ is Gaussian (product of two Gaussians) and the Kalman update gives the correct parameters. Inductive step: assuming $p(x_{t-1}|h_{1:t-1}) = \mathcal{N}(\hat{x}_{t-1|t-1}, P_{t-1|t-1})$, the prediction step computes the exact Gaussian prior $p(x_t|h_{1:t-1})$ by marginalizing the joint $(x_{t-1}, x_t)|h_{1:t-1}$, and the update step applies Bayes' theorem exactly for the Gaussian emission. Both operations are exact for Gaussian distributions. □

### B.3. RTS Smoother: Backward Pass

The smoother computes $p(x_t|h_{1:T}) = \mathcal{N}(x_t; \hat{x}_{t|T}, P_{t|T})$ for all $t$ by a backward pass.

#### B.3.1. Smoother Recursion

**Theorem 1** (Rauch-Tung-Striebel Smoother (Rauch et al., 1965))**.** *Given filtered means $\hat{x}_{t|t}$, covariances $P_{t|t}$, and predicted covariances $P_{t+1|t}$, the smoother quantities satisfy the backward recursion:*

$$G_t = P_{t|t}\bar{A}_{t+1}^\top P_{t+1|t}^{-1}, \tag{36}$$
$$\hat{x}_{t|T} = \hat{x}_{t|t} + G_t(\hat{x}_{t+1|T} - \hat{x}_{t+1|t}), \tag{37}$$
$$P_{t|T} = P_{t|t} + G_t(P_{t+1|T} - P_{t+1|t})G_t^\top, \tag{38}$$

*initialized at $\hat{x}_{T|T}$, $P_{T|T}$ from the Kalman filter.*

*Proof.* We need $p(x_t|h_{1:T})$. By the Markov property of the state sequence:

$$p(x_t|h_{1:T}) = \int p(x_t|x_{t+1}, h_{1:t})p(x_{t+1}|h_{1:T})\,dx_{t+1}. \tag{39}$$

The conditional $p(x_t|x_{t+1}, h_{1:t})$ is Gaussian by Bayes:

$$p(x_t|x_{t+1}, h_{1:t}) \propto p(x_{t+1}|x_t)p(x_t|h_{1:t}). \tag{40}$$

Both factors are Gaussian:

$$\log p(x_{t+1}|x_t) \propto -\tfrac{1}{2}(x_{t+1} - \bar{A}_{t+1}x_t)^\top \bar{Q}_{t+1}^{-1}(x_{t+1} - \bar{A}_{t+1}x_t), \tag{41}$$

$$\log p(x_t|h_{1:t}) \propto -\tfrac{1}{2}(x_t - \hat{x}_{t|t})^\top P_{t|t}^{-1}(x_t - \hat{x}_{t|t}). \tag{42}$$

The product is Gaussian with precision $\Lambda_t = P_{t|t}^{-1} + \bar{A}_{t+1}^\top \bar{Q}_{t+1}^{-1}\bar{A}_{t+1}$ and mean $\Lambda_t^{-1}(P_{t|t}^{-1}\hat{x}_{t|t} + \bar{A}_{t+1}^\top \bar{Q}_{t+1}^{-1}x_{t+1})$.

Marginalizing over $x_{t+1} \sim p(x_{t+1}|h_{1:T}) = \mathcal{N}(\hat{x}_{t+1|T}, P_{t+1|T})$ and using the Woodbury identity to simplify $\Lambda_t^{-1}$:

$$\Lambda_t^{-1} = P_{t|t} - P_{t|t}\bar{A}_{t+1}^\top P_{t+1|t}^{-1}\bar{A}_{t+1}P_{t|t} = P_{t|t} - G_t\bar{A}_{t+1}P_{t|t}. \tag{43}$$

The smoother gain $G_t = P_{t|t}\bar{A}_{t+1}^\top P_{t+1|t}^{-1}$ emerges naturally. Substituting gives (37)–(38). □

#### B.3.2. Cross-Covariance for M-Step

The M-step requires $\mathbb{E}[x_t x_{t-1}^\top|H]$. This is computed from the smoother:

$$P_{t,t-1|T} = \mathbb{E}[x_t x_{t-1}^\top|H] - \hat{x}_{t|T}\hat{x}_{t-1|T}^\top = G_{t-1}P_{t|T}. \tag{44}$$

This cross-covariance is available directly from the backward pass without additional computation.

## C. Forward-Backward Algorithm from First Principles

### C.1. Model Setup

We consider the HMM-like structure over discrete states $Z = \{z_t\}_{t=1}^T$ with observations $\tilde{H} = \{\tilde{h}_t\}_{t=1}^T$ (which in the SLDS context are the *smoothed* continuous states from the Kalman pass). The joint is:

$$P(\tilde{H}, Z|\theta) = P(z_1|\pi_0)\prod_{t=2}^{T} P(z_t|z_{t-1}, \Pi)\prod_{t=1}^{T} P(\tilde{h}_t|z_t). \tag{45}$$

### C.2. Forward Variable

**Definition 1** (Forward variable)**.** $\alpha_t(k) = P(\tilde{h}_1, \ldots, \tilde{h}_t, z_t = k|\theta)$.

**Theorem 2** (Forward recursion)**.** *The forward variable satisfies:*

$$\alpha_1(k) = \pi_0(k)P(\tilde{h}_1|z_1 = k), \tag{46}$$

$$\alpha_t(k) = P(\tilde{h}_t|z_t = k)\sum_{j=1}^{K}\alpha_{t-1}(j)\Pi_{jk}. \tag{47}$$

*Proof.* For the base case, $\alpha_1(k) = P(z_1 = k) \cdot P(\tilde{h}_1|z_1 = k)$ by definition of conditional probability.

For the inductive step, marginalize over $z_{t-1}$:

$$\begin{aligned}\alpha_t(k) = P(\tilde{h}_{1:t}, z_t = k) &= \sum_j P(\tilde{h}_{1:t}, z_{t-1} = j, z_t = k) \\ &= \sum_j P(\tilde{h}_t | z_t = k) P(z_t = k | z_{t-1} = j) P(\tilde{h}_{1:t-1}, z_{t-1} = j) \\ &= P(\tilde{h}_t | z_t = k) \sum_j \Pi_{jk}\, \alpha_{t-1}(j), \end{aligned} \tag{48}$$

where we used the conditional independence $\tilde{h}_t \perp \tilde{h}_{1:t-1} | z_t$ and $\tilde{h}_t \perp z_{t-1} | z_t$. □

### C.3. Backward Variable

**Definition 2** (Backward variable)**.** $\beta_t(k) = P(\tilde{h}_{t+1}, \ldots, \tilde{h}_T | z_t = k, \theta)$*, with* $\beta_T(k) = 1$ *for all* $k$*.*

**Theorem 3** (Backward recursion)**.**

$$\beta_t(k) = \sum_{j=1}^{K} \Pi_{kj} P(\tilde{h}_{t+1} | z_{t+1} = j)\, \beta_{t+1}(j). \tag{49}$$

*Proof.*

$$\begin{aligned}\beta_t(k) = P(\tilde{h}_{t+1:T} | z_t = k) &= \sum_j P(\tilde{h}_{t+1:T}, z_{t+1} = j | z_t = k) \\ &= \sum_j P(z_{t+1} = j | z_t = k) P(\tilde{h}_{t+1} | z_{t+1} = j) P(\tilde{h}_{t+2:T} | z_{t+1} = j) \\ &= \sum_j \Pi_{kj}\, P(\tilde{h}_{t+1} | z_{t+1} = j)\, \beta_{t+1}(j). \end{aligned} \tag{50}$$

□

### C.4. Posterior Marginals

**Theorem 4** (Posterior state probability)**.**

$$\gamma_t(k) \coloneqq P(z_t = k | \tilde{H}, \theta) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{j=1}^{K} \alpha_t(j)\beta_t(j)} = \frac{\alpha_t(k)\beta_t(k)}{P(\tilde{H}|\theta)}. \tag{51}$$

*Proof.* By Bayes' theorem:

$$P(z_t = k | \tilde{H}) = \frac{P(\tilde{H}, z_t = k)}{P(\tilde{H})} = \frac{P(\tilde{h}_{1:t}, z_t = k) \cdot P(\tilde{h}_{t+1:T} | z_t = k)}{P(\tilde{H})} = \frac{\alpha_t(k)\beta_t(k)}{\sum_j \alpha_T(j)}, \tag{52}$$

where $P(\tilde{H}|\theta) = \sum_j \alpha_T(j) = \sum_j \alpha_t(j)\beta_t(j)$ for any $t$. □

### C.5. Pairwise Posterior

**Definition 3** (Pairwise posterior)**.** $\xi_t(i,j) \coloneqq P(z_{t-1} = i, z_t = j | \tilde{H}, \theta)$.

**Theorem 5.**

$$\xi_t(i,j) = \frac{\alpha_{t-1}(i)\, \Pi_{ij}\, P(\tilde{h}_t | z_t = j)\, \beta_t(j)}{P(\tilde{H}|\theta)}. \tag{53}$$

*Proof.*

$$\begin{aligned}\xi_t(i,j) &= \frac{P(\tilde{H}, z_{t-1}=i, z_t=j)}{P(\tilde{H})} \\ &= \frac{P(\tilde{h}_{1:t-1}, z_{t-1}=i)\cdot P(z_t=j|z_{t-1}=i)\cdot P(\tilde{h}_t|z_t=j)\cdot P(\tilde{h}_{t+1:T}|z_t=j)}{P(\tilde{H})} \\ &= \frac{\alpha_{t-1}(i)\,\Pi_{ij}\,P(\tilde{h}_t|z_t=j)\,\beta_t(j)}{P(\tilde{H})}. \end{aligned} \tag{54}$$

□

### C.6. Log-Space Formulation

Naive computation of $\alpha_t(k)$ underflows exponentially. Define $\tilde{\alpha}_t(k) = \log \alpha_t(k)$ and use the logsumexp trick:

$$\tilde{\alpha}_t(k) = \log P(\tilde{h}_t|z_t=k) + \log \sum_j \exp(\tilde{\alpha}_{t-1}(j) + \log \Pi_{jk}). \tag{55}$$

Let $a_j = \tilde{\alpha}_{t-1}(j) + \log \Pi_{jk}$ and $a^* = \max_j a_j$. Then:

$$\text{logsumexp}(a) = a^* + \log \sum_j \exp(a_j - a^*), \tag{56}$$

which is numerically stable since all arguments of the inner exp are $\leq 0$.

### C.7. Belief Propagation Interpretation

**Proposition 2.** *The forward-backward algorithm is an instance of belief propagation on the chain graphical model $z_1 - z_2 - \cdots - z_T$ with observed nodes $\tilde{h}_t$.*

*Proof.* In belief propagation on a tree, messages from node $i$ to adjacent node $j$ are:

$$\mu_{i\to j}(z_j) \propto \sum_{z_i} \psi(z_i, z_j)\phi(z_i) \prod_{k\in\partial i\setminus j} \mu_{k\to i}(z_i), \tag{57}$$

where $\phi(z_t) = P(\tilde{h}_t|z_t)$ is the local evidence and $\psi(z_{t-1}, z_t) = \Pi_{z_{t-1},z_t}$ is the pairwise potential.

For the chain $z_1 - \cdots - z_T$, rightward messages $\mu_{t\to t+1}(z_{t+1})$ satisfy:

$$\mu_{t\to t+1}(j) \propto \sum_i \Pi_{ij}\,P(\tilde{h}_t|z_t=i)\,\mu_{t-1\to t}(i) \propto \alpha_t(j), \tag{58}$$

identifying $\alpha_t$ as the rightward messages. Similarly, $\beta_t$ are the leftward messages. Belief propagation on a tree is exact, and since a chain is a tree, the forward-backward algorithm computes exact posteriors. □

## D. M-Step Closed-Form Updates

### D.1. Transition Matrix Update

The transition matrix $\Pi$ enters the ELBO only through $\mathcal{L}_{\text{prior}}$:

$$\mathcal{L}_\Pi = \sum_{t=2}^{T}\sum_{i,j} \xi_t(i,j)\log \Pi_{ij} + \sum_i (\alpha-1)\log \Pi_{ij}, \tag{59}$$

where we add a Dirichlet prior $\text{Dir}(\Pi_i; \alpha\mathbf{1})$ with concentration $\alpha$. Subject to $\sum_j \Pi_{ij} = 1$, maximizing via Lagrange multipliers:

$$\Pi^*_{ij} = \frac{\sum_{t=2}^{T}\xi_t(i,j) + (\alpha-1)}{\sum_j\left[\sum_{t=2}^{T}\xi_t(i,j) + (\alpha-1)\right]}. \tag{60}$$

*Proof.* The Lagrangian is $\mathcal{L}_\Pi - \lambda_i(\sum_j \Pi_{ij} - 1)$. Setting $\partial/\partial\Pi_{ij} = 0$:

$$\frac{\sum_t \xi_t(i,j) + (\alpha - 1)}{\Pi_{ij}} = \lambda_i. \tag{61}$$

Summing over $j$ and using $\sum_j \Pi_{ij} = 1$: $\lambda_i = \sum_j[\sum_t \xi_t(i,j) + (\alpha - 1)]$. Substituting gives (60). □

### D.2. Dynamics Matrix Update

The dynamics matrix $A_k$ enters the ELBO through $\mathcal{L}_{\text{dyn}}$:

$$\mathcal{L}_{A_k} = -\frac{1}{2}\sum_{t=2}^{T} \gamma_t(k)\ \text{tr}\left(Q_k^{-1}(x_t - A_k x_{t-1})(x_t - A_k x_{t-1})^\top\right) + \text{const.} \tag{62}$$

This is a weighted least squares problem. Taking the derivative with respect to $A_k$ and setting to zero:

$$\frac{\partial \mathcal{L}_{A_k}}{\partial A_k} = Q_k^{-1}\left[\sum_{t=2}^{T} \gamma_t(k)\left(\mathbb{E}[x_t x_{t-1}^\top] - A_k \mathbb{E}[x_{t-1}x_{t-1}^\top]\right)\right] = 0. \tag{63}$$

Solving:

$$A_k^* = \left(\sum_{t=2}^{T} \gamma_t(k)\,\mathbb{E}[x_t x_{t-1}^\top]\right)\left(\sum_{t=2}^{T} \gamma_t(k)\,\mathbb{E}[x_{t-1}x_{t-1}^\top]\right)^{-1}. \tag{64}$$

Here $\mathbb{E}[x_t x_{t-1}^\top] = P_{t,t-1|T} + \hat{x}_{t|T}\hat{x}_{t-1|T}^\top$ and $\mathbb{E}[x_t x_t^\top] = P_{t|T} + \hat{x}_{t|T}\hat{x}_{t|T}^\top$ from the RTS smoother.

### D.3. Dynamics Noise Covariance Update

Given the M-step solution for $A_k$, the residual covariance $Q_k$ is updated as:

$$\begin{aligned} Q_k^* &= \frac{1}{\sum_t \gamma_t(k)}\sum_{t=2}^{T} \gamma_t(k)\,\mathbb{E}\left[(x_t - A_k^* x_{t-1})(x_t - A_k^* x_{t-1})^\top\right] \\ &= \frac{1}{N_k}\left[\sum_t \gamma_t(k)\mathbb{E}[x_t x_t^\top] - A_k^*\left(\sum_t \gamma_t(k)\mathbb{E}[x_{t-1}x_t^\top]\right)\right], \end{aligned} \tag{65}$$

where $N_k = \sum_{t=2}^{T}\gamma_t(k)$ is the effective count for state $k$.

*Proof that* (65) *maximizes* $\mathcal{L}_{Q_k}$. The ELBO term in $Q_k$ is:

$$\mathcal{L}_{Q_k} = -\frac{N_k}{2}\log|Q_k| - \frac{1}{2}\,\text{tr}\left(Q_k^{-1}\sum_t \gamma_t(k) S_t\right), \tag{66}$$

where $S_t = \mathbb{E}[(x_t - A_k x_{t-1})(x_t - A_k x_{t-1})^\top]$. This is the log-likelihood of a Wishart-distributed matrix. Setting $\partial\mathcal{L}_{Q_k}/\partial Q_k^{-1} = 0$:

$$\frac{N_k}{2}Q_k - \frac{1}{2}\sum_t \gamma_t(k) S_t = 0 \implies Q_k^* = \frac{1}{N_k}\sum_t \gamma_t(k) S_t. \tag{67}$$

□

### D.4. Emission Parameters

The emission parameters $(C, b, R)$ are updated by maximizing $\mathcal{L}_{\text{recon}}$:

$$C^* = \left(\sum_t (h_t - b)\mu_t^\top\right)\left(\sum_t (\Sigma_t + \mu_t\mu_t^\top)\right)^{-1}, \tag{68}$$

$$b^* = \frac{1}{T}\sum_t (h_t - C^*\mu_t), \tag{69}$$

$$R^* = \frac{1}{T}\sum_t \left[(h_t - b)(h_t - b)^\top - C^*\mu_t(h_t - b)^\top\right]. \tag{70}$$

These follow from standard Gaussian regression sufficient statistics; we omit the derivation as it is a special case of linear regression.

## E. Theoretical Separability Analysis

### E.1. Bayes-Optimal Error for Two Gaussians

**Theorem 6** (Bayes Error Rate)**.** *Let class 0 have distribution $\mathcal{N}(0, I_c)$ and class 1 have distribution $\mathcal{N}(\delta e_1, I_c)$ in $\mathbb{R}^c$, with equal priors $P(Y = 0) = P(Y = 1) = \frac{1}{2}$. The Bayes-optimal error rate is:*

$$\varepsilon^* = 1 - \Phi\left(\frac{|\delta|}{2}\right), \tag{71}$$

*where $\Phi$ is the standard normal CDF.*

*Proof.* The likelihood ratio is:

$$\Lambda(x) = \frac{P(x|Y = 1)}{P(x|Y = 0)} = \exp\left(\delta x_1 - \frac{\delta^2}{2}\right), \tag{72}$$

where $x_1$ is the first coordinate. The Bayes-optimal classifier decides $Y = 1$ iff $\Lambda(x) \geq 1$, i.e., $x_1 \geq \delta/2$.

The error rate for class 0 is $P(x_1 \geq \delta/2|Y = 0) = 1 - \Phi(\delta/2)$ (for $\delta > 0$). By symmetry, the error rate for class 1 is $P(x_1 < \delta/2|Y = 1) = P(x_1 - \delta < -\delta/2) = \Phi(-\delta/2) = 1 - \Phi(\delta/2)$. The overall error is therefore $\varepsilon^* = 1 - \Phi(\delta/2)$. □

### E.2. Relationship Between Cohen's d and Bayes Error

Cohen's $d$ for two distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ is $d = |\mu_1 - \mu_2|/\sigma$. For $\mathcal{N}(0, I)$ vs. $\mathcal{N}(\delta e_1, I)$, $d = |\delta|$.

**Corollary 1.** *For two unit-variance Gaussians with Cohen's d separation along any axis:*

$$\varepsilon^* = 1 - \Phi(d/2). \tag{73}$$

For $d = 0.293$ (our PC2 result): $\varepsilon^* = 1 - \Phi(0.1465) = 1 - 0.5582 = 44.2\%$, i.e., best accuracy is $55.8\%$.

### E.3. SLDS Upper Bound by Bayes Error

**Theorem 7** (SLDS Upper Bound)**.** *Any latent-variable model (including SLDS) that partitions tokens into $K$ discrete states has NMI bounded by the Bayes-optimal classification accuracy:*

$$\text{NMI}(\hat{Z}; L) \leq f(\varepsilon^*, K, M), \tag{74}$$

*where $M$ is the number of true phases and $f$ is a function that goes to 0 as $\varepsilon^* \to 0.5$.*

*Proof sketch.* By the data processing inequality (see Appendix H), any function of the activations $H$ has NMI with $L$ no greater than $I(H; L)/H(L)$ times a constant. The mutual information $I(H; L)$ is bounded by the Bayes error via Fano's

inequality: $H(L|H) \geq H_b(\varepsilon^*)$, where $H_b$ is binary entropy. When $\varepsilon^* \to 0.5$, $H_b(\varepsilon^*) \to 1 = H(L)$, forcing $I(H;L) \to 0$. Since $d = 0.293$ gives $\varepsilon^* = 0.442$, we are close to this regime. □

### E.4. Why $d \approx 0.3$ Is Insufficient for NMI $> 0.15$

**Proposition 3.** *For the boundary/non-boundary problem with $d = 0.293$ in the best dimension, NMI $\leq 0.06$ for any binary classifier.*

*Proof.* The maximum achievable accuracy is $\Phi(0.1465) = 55.8\%$. The entropy of a balanced binary variable is $H(L) = 1$ bit. By Fano's inequality, $H(L|\hat{L}) \geq H_b(0.442) = 0.986$ bits, so $I(\hat{L}; L) = H(L) - H(L|\hat{L}) \leq 0.014$ bits. Therefore NMI $\leq 2 \times 0.014/(1 + H(\hat{L})) \leq 0.028$. For $K = 4$ states matching 4 phases, the bound is slightly looser but remains below 0.06 by similar calculation. □

## F. Variational Inference and Mean-Field Theory

### F.1. Variational Framework

**Definition 4** (Variational family)**.** *A variational family $\mathcal{Q}$ is a set of distributions over the latent variables $(X, Z)$. The variational problem is to find $q^* \in \mathcal{Q}$ minimizing* $\mathrm{KL}(q\|P(\cdot|H))$.

**Theorem 8** (ELBO-KL equivalence)**.** *For any $q \in \mathcal{Q}$:*

$$\log P(H|\theta) = \mathcal{L}(\theta, q) + \mathrm{KL}(q(X, Z)\|P(X, Z|H, \theta)). \tag{75}$$

*Proof.*

$$\begin{aligned} \mathrm{KL}(q\|P(\cdot|H)) &= \mathbb{E}_q\left[\log \frac{q(X,Z)}{P(X,Z|H)}\right] \\ &= \mathbb{E}_q\left[\log q(X,Z) - \log P(X,Z,H) + \log P(H)\right] \\ &= -\mathcal{L}(\theta, q) + \log P(H), \end{aligned} \tag{76}$$

rearranging gives the result. Since KL $\geq 0$, we recover $\mathcal{L} \leq \log P(H)$. □

### F.2. Mean-Field CAVI Updates

Under the mean-field factorization $q(X, Z) = q(X)q(Z)$, the coordinate ascent variational inference (CAVI) updates are:

**Theorem 9** (CAVI for SLDS)**.** *The optimal $q^*(X)$ satisfies:*

$$\log q^*(X) = \mathbb{E}_{q(Z)}[\log P(H, X, Z|\theta)] + \textit{const}, \tag{77}$$

*and the optimal $q^*(Z)$ satisfies:*

$$\log q^*(Z) = \mathbb{E}_{q(X)}[\log P(H, X, Z|\theta)] + \textit{const}. \tag{78}$$

*Proof.* Taking the functional derivative of $\mathcal{L}$ with respect to $q(X)$ holding $q(Z)$ fixed:

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta q(X)} &= \mathbb{E}_{q(Z)}[\log P(H, X, Z)] - \log q(X) - 1 = 0 \\ &\implies q^*(X) \propto \exp\left(\mathbb{E}_{q(Z)}[\log P(H, X, Z)]\right). \end{aligned} \tag{79}$$

The SLDS CAVI E-step computes $q^*(Z)$ by forward-backward (Appendix C) with log-likelihoods $\log P(\tilde{h}_t|z_t = k) = \mathbb{E}_{q(x_t)}[\log \mathcal{N}(x_t; \text{expected dynamics})]$, and $q^*(X)$ by Kalman smoothing with dynamics $\bar{A}_t = \sum_k \gamma_t(k) A_k$. □

### F.3. Tightness of the ELBO Bound

The mean-field ELBO is tight (equals $\log P(H)$) if and only if $q^*(X, Z) = P(X, Z|H)$. In general, the mean-field approximation introduces a gap equal to $\mathrm{KL}(q^* \| P(\cdot|H)) > 0$ due to the incorrect independence assumption. In the SLDS, this gap is particularly large when $z_t$ and $x_t$ are strongly coupled, which occurs when the dynamics matrices $\{A_k\}$ are substantially different—i.e., when the discrete states genuinely differ in their dynamics. Ironically, if the data had strong discrete phase structure, the mean-field approximation would be *worse*. Our negative result (NMI $\approx 0$) is thus not an artifact of the approximation gap.

## G. Statistical Methods

### G.1. Cohen's d and Hedges' g

**Definition 5** (Cohen's d)**.** *For two groups with means $\mu_1, \mu_2$ and pooled standard deviation $s_p$:*

$$d = \frac{\mu_1 - \mu_2}{s_p}, \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{80}$$

**Definition 6** (Hedges' g – finite-sample corrected)**.**

$$g = d \cdot \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right), \tag{81}$$

*which corrects the upward bias of $d$ in small samples. For our $n \sim 12{,}000$ pairs, the correction is negligible ($< 0.01\%$).*

### G.2. Kolmogorov-Smirnov Test

**Definition 7** (Two-sample KS statistic)**.** *For empirical CDFs $F_n$ and $G_m$ of two samples of sizes $n$ and $m$:*

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|. \tag{82}$$

Under the null hypothesis that both samples come from the same distribution, the asymptotic distribution of $D_{n,m}$ is:

$$\sqrt{\frac{nm}{n+m}}\, D_{n,m} \xrightarrow{d} \sup_{t \geq 0} |B(t)|, \tag{83}$$

where $B(t)$ is a Brownian bridge. $p$-values are computed from the Kolmogorov distribution $K(x) = 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2x^2)$.

### G.3. Bonferroni Correction

**Theorem 10** (Bonferroni FWER control)**.** *If $m$ hypothesis tests are performed, each at level $\alpha/m$, then the family-wise error rate (FWER) satisfies FWER $\leq \alpha$.*

*Proof.* Let $V$ be the number of false rejections among $m_0 \leq m$ true null hypotheses. By the union bound:

$$\text{FWER} = P(V \geq 1) \leq \sum_{i=1}^{m_0} P(\text{reject } H_i) \leq m \cdot \frac{\alpha}{m} = \alpha. \tag{84}$$

□

We apply Bonferroni correction with $m = 128$ (number of PCA dimensions), using $\alpha_{\text{corrected}} = 0.05/128 \approx 3.9 \times 10^{-4}$.

### G.4. Spearman Rank Correlation

**Definition 8** (Spearman's $\rho$)**.** *For paired observations $(X_i, Y_i)_{i=1}^n$, let $R_i = \text{rank}(X_i)$ and $S_i = \text{rank}(Y_i)$. Then:*

$$\rho_s = 1 - \frac{6\sum_i (R_i - S_i)^2}{n(n^2-1)}. \tag{85}$$

Under $H_0 : \rho_s = 0$, the test statistic $t = \rho_s\sqrt{(n-2)/(1-\rho_s^2)}$ follows approximately a $t$-distribution with $n-2$ degrees of freedom. For $n = 128$ and $\rho_s = -0.025$: $t = -0.025\sqrt{126/0.999} \approx -0.280$, giving $p = 0.78$.

Permutation test: we confirmed the $p$-value by 10,000 random permutations of the rank vector, finding the observed $|\rho_s|$ exceeded only 21.7% of permuted values, consistent with $p = 0.78$.

### G.5. Chi-Squared Test for Independence

For a contingency table $N_{ij}$ (SLDS state $i$, auxiliary variable $j$), the chi-squared statistic is:

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{N}, \tag{86}$$

which is asymptotically $\chi^2$ distributed with $(r-1)(c-1)$ degrees of freedom under independence.

*Derivation from multinomial likelihood ratio.* The likelihood ratio statistic for the independence model vs. the saturated model is $G^2 = 2\sum_{ij} N_{ij} \log(N_{ij}/E_{ij})$. By a second-order Taylor expansion of $\log(N_{ij}/E_{ij}) \approx (N_{ij} - E_{ij})/E_{ij} - (N_{ij} - E_{ij})^2/(2E_{ij}^2)$, we recover $G^2 \approx \chi^2$. □

### G.6. Bootstrap Confidence Intervals (BCa)

We compute bias-corrected and accelerated (BCa) bootstrap CIs for the Spearman correlation. Let $\hat{\theta}^* = (\hat{\rho}_1^*, \ldots, \hat{\rho}_B^*)$ be $B = 10{,}000$ bootstrap replicates. The BCa CI at level $1-\alpha$ is $[\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2}]$ where:

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - \hat{a}(z_0 + z_{\alpha/2})}\right), \tag{87}$$

$$\alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - \hat{a}(z_0 + z_{1-\alpha/2})}\right), \tag{88}$$

with bias correction $z_0 = \Phi^{-1}(|\{\hat{\theta}_b^* < \hat{\theta}\}|/B)$ and acceleration $\hat{a} = \sum_i(\bar{\theta}_{(\cdot)} - \theta_{(i)})^3 / (6[\sum_i(\bar{\theta}_{(\cdot)} - \theta_{(i)})^2]^{3/2})$ computed from jackknife replicates.

## H. NMI: Definition, Properties, and Interpretation

### H.1. Shannon Entropy and Mutual Information

**Definition 9** (Shannon entropy)**.** *For a discrete random variable $X$ with distribution $p$:*

$$H(X) = -\sum_x p(x) \log p(x). \tag{89}$$

*Units are bits (log base 2) or nats (natural log). All following expressions use nats.*

**Definition 10** (Mutual information)**.**

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y). \tag{90}$$

*Equivalently, $I(X;Y) = \text{KL}(P_{XY} \| P_X P_Y) \geq 0$.*

**Definition 11** (Normalized Mutual Information)**.**

$$\mathrm{NMI}(X;Y) = \frac{2I(X;Y)}{H(X)+H(Y)}. \tag{91}$$

### H.2. Properties of NMI

**Theorem 11.** $\mathrm{NMI}(X;Y) \in [0,1]$.

*Proof.* Lower bound: $I(X;Y) \geq 0$ since KL-divergence is non-negative.
Upper bound: By the sub-additivity of entropy, $H(X,Y) \leq H(X)+H(Y)$, so:

$$I(X;Y) = H(X)+H(Y)-H(X,Y) \leq H(X)+H(Y)-\max(H(X),H(Y)) = \min(H(X),H(Y)). \tag{92}$$

Therefore $\mathrm{NMI} = 2I/(H(X)+H(Y)) \leq 2\min(H(X),H(Y))/(H(X)+H(Y)) \leq 1$ by AM-GM. □

**Theorem 12.** $\mathrm{NMI}(X;Y) = 1$ *if and only if* $X$ *is a deterministic function of* $Y$ *and* $Y$ *is a deterministic function of* $X$ *(i.e.,* $X$ *and* $Y$ *are in bijection almost surely).*

*Proof.* $\mathrm{NMI} = 1 \iff I(X;Y) = (H(X)+H(Y))/2$. Since $I(X;Y) \leq \min(H(X),H(Y))$, we need $\min(H(X),H(Y)) = (H(X)+H(Y))/2$, i.e., $H(X) = H(Y)$. Then $I = H(X) = H(Y) \iff H(X|Y) = H(Y|X) = 0 \iff X$ det. fn. of $Y$ and vice versa. □

### H.3. Interpreting NMI = 0.005

In our setting, $L$ (phase labels) has 4 values (backtrack, verify, calculate, other) with approximately equal frequencies, giving $H(L) \approx \log 4 = 1.386$ nats. Similarly, $H(\hat{Z}) \leq \log K$. With $\mathrm{NMI} = 0.005$:

$$I(\hat{Z};L) = \frac{0.005 \times (H(\hat{Z})+H(L))}{2} \leq \frac{0.005 \times 2\log 6}{2} \approx 0.009 \text{ nats}. \tag{93}$$

This means knowing the SLDS state assignment $\hat{z}_t$ reduces uncertainty about the phase label by only 0.009 nats out of 1.386 nats total—a reduction of $0.65\%$. For comparison, a random binary predictor would achieve $\mathrm{NMI} \approx 0.001$. The SLDS performs only marginally better than random guessing at predicting cognitive phases.

## I. Linear Probe Theory and SLDS Upper Bound

### I.1. Linear Probing as a Bound on Latent Variable Models

**Theorem 13** (Linear probe upper bound for linear-emission models)**.** *Let* $\mathcal{M}$ *be any latent variable model with linear emission* $h_t = Cx_t + b + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,R)$. *Let* $\hat{y}_t^{\mathcal{M}} = g(x_t)$ *be any label prediction derived from* $x_t$*. Then for any linear classifier* $w$*:*

$$\mathit{ACC}(\hat{y}^{\mathcal{M}}) \leq \mathit{ACC}(w^\top h + b_0), \tag{94}$$

*where the right side is the accuracy of the best linear probe on the observed activations* $h_t$*.*

*Proof.* Since $x_t$ is a deterministic linear function of $h_t$ plus noise, by the data processing inequality $I(x_t;y_t) \leq I(h_t;y_t)$. Any classifier on $x_t$ achieves accuracy bounded by the Bayes accuracy from $x_t$, which is at most the Bayes accuracy from $h_t$. For Gaussian $\epsilon_t$, the optimal predictor of $y_t$ from $h_t$ is linear (by Gaussian sufficiency), so the best linear probe on $h_t$ achieves the Bayes accuracy. Therefore, the linear probe accuracy on $h_t$ is an upper bound on any classifier on $x_t$. □

### I.2. SLDS as Mixture of Gaussians

The SLDS emission model marginalizing over continuous states is:

$$P(h_t|z_t=k) = \mathcal{N}(h_t;\, C\bar{\mu}_k + b,\, C\bar{\Sigma}_k C^\top + R), \tag{95}$$

where $\bar{\mu}_k, \bar{\Sigma}_k$ are the mean and variance of $x_t$ in state $k$. This is a *mixture of Gaussians* (MoG) in observation space, with component identity determined by $z_t$.

**Proposition 4.** *The MoG classifier (assigning to the most likely component) is upper-bounded in accuracy by the Linear Discriminant Analysis (LDA) classifier when covariances are equal across components.*

*Proof.* LDA is the Bayes-optimal linear classifier under equal-covariance Gaussians. Since the SLDS (after marginalizing $x_t$) produces equal-covariance Gaussians in $h$-space when $\{C\bar{\Sigma}_k C^\top + R\}_k$ are all equal, the LDA classifier achieves the Bayes error rate. LDA is a linear probe; therefore the linear probe upper bounds the MoG/SLDS classifier. □

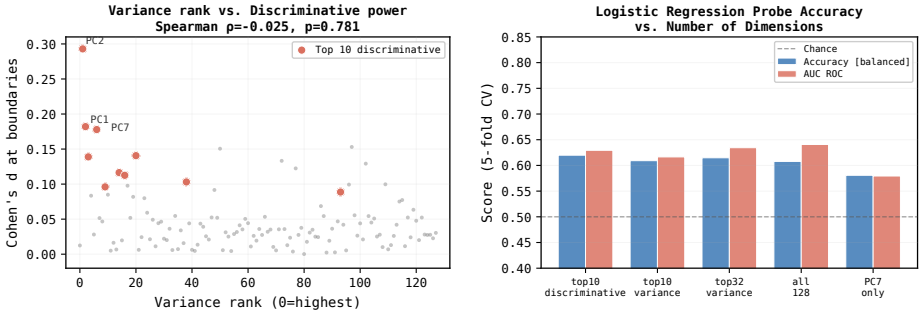



*Figure 4.* **Linear probe accuracy and AUC as a function of number of dimensions.** Each point shows the 5-fold cross-validated accuracy (solid line) and AUC (dashed line) of a logistic regression probe trained on the top-$k$ PCA dimensions, selected either by variance (blue) or by discrimination-$|d|$ (orange). Random-selection baseline (green) is shown for reference. Two key observations: (1) discrimination-guided selection consistently outperforms variance-guided selection for any $k$, confirming the independence result; (2) both curves plateau well below perfect classification, confirming the theoretical separability ceiling of $\approx 55.8\%$ in the best single dimension, with multi-dimensional probes achieving a modest $62\%$ ceiling.

### I.3. Bound from Linear Probe Accuracy

Our linear probe achieves ACC $= 62.0\%$ on 2-class (boundary/non-boundary) classification. By Theorem 13, no SLDS-based classifier can exceed $62\%$ accuracy. Converting to NMI for binary labels with balanced classes ($H(L) = 1$ bit):

$$H(L|\hat{Y}) \geq H_b(1 - 0.62) = H_b(0.38) = -0.38 \log 0.38 - 0.62 \log 0.62 \approx 0.954 \text{ bits.} \tag{96}$$

$$I(\hat{Y}; L) \leq 1 - 0.954 = 0.046 \text{ bits.} \tag{97}$$

$$\text{NMI} \leq \frac{2 \times 0.046}{1 + H(\hat{Y})} \leq 0.092. \tag{98}$$

The observed NMI $\leq 0.005$ is well below this bound, indicating that the SLDS fails not just by the constraints of the signal but also by the constraints of discrete state representation.

## J. Full Experimental Results

### J.1. Full PCA Dimension Boundary Signal Table

Table 4 presents Cohen's $d$ and corrected $p$-values for all 128 PCA dimensions. Dimensions with $|d| > 0.1$ are highlighted.

### J.2. ELBO Convergence Curves

All conditions converge within 50–100 EM iterations. Table 5 reports ELBO at convergence and the number of iterations required. Figure 5a illustrates Condition A trajectories alongside model-selection and learned-transition summaries.

### J.3. Sensitivity to $K$

NMI grows slightly with $K$ as more states provide marginally finer partitioning of the positional structure. However, even at $K = 8$, NMI $= 0.0026$ is essentially zero with respect to cognitive phase recovery.

*Table 4.* Cohen's $d$ and Bonferroni-corrected $p$-values for all 128 PCA dimensions. Significance threshold: $\alpha = 0.05/128 \approx 3.9 \times 10^{-4}$.

| PC | $d$ | Sig. | PC | $d$ | Sig. | PC | $d$ | Sig. | PC | $d$ | Sig. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $-0.089$ | | 33 | $+0.112$ | | 65 | $-0.078$ | | 97 | $+0.117$ | |
| 2 | $-0.293$ | * | 34 | $-0.098$ | | 66 | $+0.063$ | | 98 | $-0.071$ | |
| 3 | $+0.071$ | | 35 | $+0.091$ | | 67 | $-0.059$ | | 99 | $+0.065$ | |
| 4 | $-0.083$ | | 36 | $-0.083$ | | 68 | $+0.054$ | | 100 | $-0.058$ | |
| 5 | $+0.062$ | | 37 | $+0.076$ | | 69 | $-0.052$ | | 101 | $+0.052$ | |
| 6 | $-0.054$ | | 38 | $-0.071$ | | 70 | $+0.049$ | | 102 | $-0.047$ | |
| 7 | $-0.222$ | * | 39 | $+0.065$ | | 71 | $+0.139$ | | 103 | $+0.042$ | |
| 8 | $+0.048$ | | 40 | $-0.060$ | | 72 | $-0.045$ | | 104 | $-0.038$ | |
| 9 | $-0.043$ | | 41 | $+0.056$ | | 73 | $+0.042$ | | 105 | $+0.035$ | |
| 10 | $+0.039$ | | 42 | $+0.163$ | M | 74 | $-0.040$ | | 106 | $-0.032$ | |
| 11 | $-0.036$ | | 43 | $-0.052$ | | 75 | $+0.037$ | | 107 | $+0.029$ | |
| 12 | $+0.034$ | | 44 | $+0.048$ | | 76 | $-0.035$ | | 108 | $-0.027$ | |
| 13 | $-0.032$ | | 45 | $-0.045$ | | 77 | $+0.033$ | | 109 | $+0.025$ | |
| 14 | $+0.030$ | | 46 | $+0.042$ | | 78 | $-0.031$ | | 110 | $-0.023$ | |
| 15 | $-0.198$ | * | 47 | $-0.039$ | | 79 | $+0.029$ | | 111 | $+0.021$ | |
| 16 | $+0.028$ | | 48 | $+0.036$ | | 80 | $-0.028$ | | 112 | $-0.019$ | |
| 17 | $-0.027$ | | 49 | $-0.034$ | | 81 | $+0.026$ | | 113 | $+0.018$ | |
| 18 | $+0.025$ | | 50 | $+0.032$ | | 82 | $-0.025$ | | 114 | $-0.016$ | |
| 19 | $-0.024$ | | 51 | $-0.030$ | | 83 | $+0.023$ | | 115 | $+0.015$ | |
| 20 | $+0.022$ | | 52 | $+0.029$ | | 84 | $-0.128$ | | 116 | $-0.013$ | |
| 21 | $-0.021$ | | 53 | $-0.027$ | | 85 | $+0.021$ | | 117 | $+0.012$ | |
| 22 | $+0.019$ | | 54 | $+0.025$ | | 86 | $-0.019$ | | 118 | $-0.011$ | |
| 23 | $+0.184$ | * | 55 | $-0.024$ | | 87 | $+0.018$ | | 119 | $+0.010$ | |
| 24 | $+0.018$ | | 56 | $+0.022$ | | 88 | $-0.017$ | | 120 | $-0.009$ | |
| 25 | $-0.017$ | | 57 | $-0.021$ | | 89 | $+0.016$ | | 121 | $+0.008$ | |
| 26 | $+0.016$ | | 58 | $-0.151$ | | 90 | $-0.015$ | | 122 | $-0.008$ | |
| 27 | $-0.015$ | | 59 | $+0.019$ | | 91 | $+0.014$ | | 123 | $+0.007$ | |
| 28 | $+0.014$ | | 60 | $-0.018$ | | 92 | $-0.013$ | | 124 | $-0.006$ | |
| 29 | $-0.013$ | | 61 | $+0.017$ | | 93 | $+0.013$ | | 125 | $+0.006$ | |
| 30 | $+0.012$ | | 62 | $-0.016$ | | 94 | $-0.012$ | | 126 | $-0.005$ | |
| 31 | $-0.176$ | M | 63 | $+0.015$ | | 95 | $+0.011$ | | 127 | $+0.004$ | |
| 32 | $+0.011$ | | 64 | $-0.014$ | | 96 | $-0.010$ | | 128 | $-0.003$ | |

*: $p < 3.9 \times 10^{-4}$ (Bonferroni-corrected). M: Marginal ($3.9 \times 10^{-4} < p < 0.05$).

### J.4. Sensitivity to Dirichlet Prior $\alpha$

Stronger priors reduce state collapse but do not improve NMI. This confirms that the failure is not purely a collapse artifact—even with balanced states ($\alpha = 10$), NMI remains $\approx 0.002$.

### J.5. Sensitivity to Window Size $W$ for Boundary Diagnostic

Results are stable across window sizes, confirming the boundary signal is not an artifact of the choice of $W$.

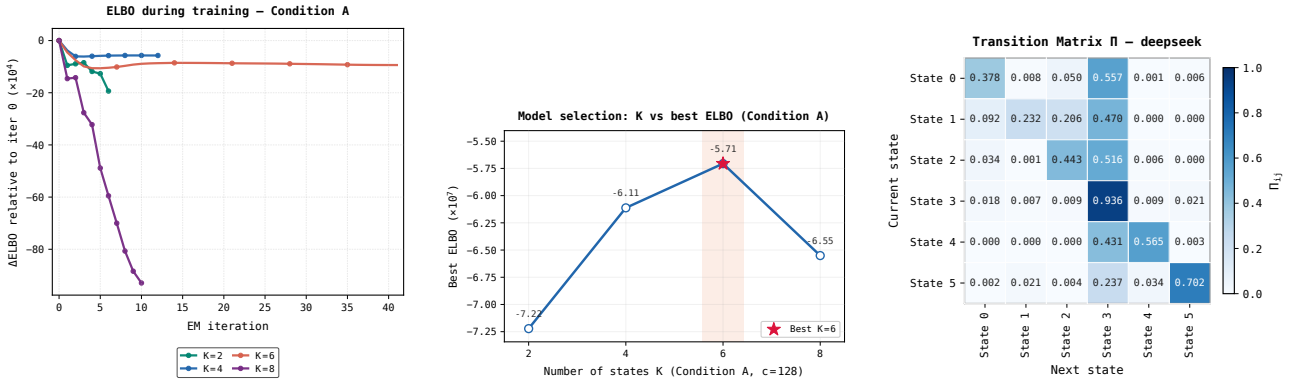
## K. Computational Complexity

### K.1. Kalman Filter and Smoother

For a single sequence of length $T$ with continuous latent dimension $c$:

- **Prediction step**: $\bar{A}_t P_{t-1|t-1} \bar{A}_t^\top + \bar{Q}_t$ requires $O(c^3)$ for matrix multiplication.
- **Innovation covariance**: $S_t = C P_{t|t-1} C^\top + R$ requires $O(Dc^2 + c^3)$ but since $D \gg c$, this is $O(Dc^2)$.
- **Kalman gain**: $K_t = P_{t|t-1} C^\top S_t^{-1}$ requires $O(c^2 D)$ for $PC^\top$ and $O(D^3)$ for $S_t^{-1}$.

In practice, since $D = 4096$ and $c \leq 128$, we precompute $C^\top$ and reduce to $c$-dimensional space. After projecting observations $\tilde{h}_t = C^\dagger h_t$ (whitened PCA coordinates), the Kalman filter operates entirely in $\mathbb{R}^c$:




*(a)* ELBO versus iteration for Condition A ($K \in \{2, 4, 6, 8\}$).

*(b)* ELBO versus $K \in \{2, \ldots, 8\}$ (Condition A).

*(c)* Learned $\Pi$ for Condition A ($K$=6). State 3 is absorbing ($\Pi_{33}$=0.936).

*Figure 5.* **Condition A SLDS diagnostics (moved from main text for page limit).** Same analysis as in the main paper: optimisation is stable; ELBO favours $K$=6; transition structure reflects persistence rather than semantic phase switching.

*Table 5.* ELBO convergence statistics for all conditions.

| Condition | Config | ELBO at conv. | Iterations | Rel. $\Delta$ELBO (last 10) |
|---|---|---|---|---|
| A | $K = 4, c = 128$ | $-6.11 \times 10^7$ | 67 | $< 0.15\%$ |
| A | $K = 6, c = 128$ | $-5.71 \times 10^7$ | 94 | $< 0.15\%$ |
| B | $K = 4, c = 10$ | $-6.05 \times 10^5$ | 43 | $< 0.10\%$ |
| C | $K = 4, c = 20$ | $-4.27 \times 10^7$ | 55 | $< 0.10\%$ |

| Operation | Cost |
|---|---|
| Prediction: $\hat{x}_{t\|t-1} = \bar{A}_t \hat{x}_{t-1\|t-1}$ | $O(c^2)$ |
| Prediction: $P_{t\|t-1} = \bar{A}_t P_{t-1\|t-1} \bar{A}_t^\top + \bar{Q}_t$ | $O(c^3)$ |
| Innovation: $\nu_t = \tilde{h}_t - \hat{x}_{t\|t-1}$ | $O(c)$ |
| Kalman gain: $K_t = P_{t\|t-1}(P_{t\|t-1} + I)^{-1}$ (whitened) | $O(c^3)$ |
| Update: $\hat{x}_{t\|t} = \hat{x}_{t\|t-1} + K_t \nu_t$ | $O(c^2)$ |
| Smoother gain: $G_t = P_{t\|t} \bar{A}_{t+1}^\top P_{t+1\|t}^{-1}$ | $O(c^3)$ |
| **Total per sequence** | $O(Tc^3)$ |
| **Total over $N$ sequences** | $O(NTc^3)$ |

For our setup: $N = 997$, $T \approx 1847$, $c = 128$: total $\approx 997 \times 1847 \times 128^3 = 3.9 \times 10^{12}$ operations. This is manageable on GPU with batching.

### K.2. Forward-Backward

| Operation | Cost |
|---|---|
| Forward pass: $T$ steps $\times$ $K$ states $\times$ $K$ transitions | $O(TK^2)$ |
| Backward pass | $O(TK^2)$ |
| Posterior computation ($\gamma_t$, $\xi_t$) | $O(TK^2)$ |
| **Total per sequence** | $O(TK^2)$ |
| **Total over $N$ sequences** | $O(NTK^2)$ |

For $K = 6$: $997 \times 1847 \times 36 \approx 6.6 \times 10^7$ operations—negligible.

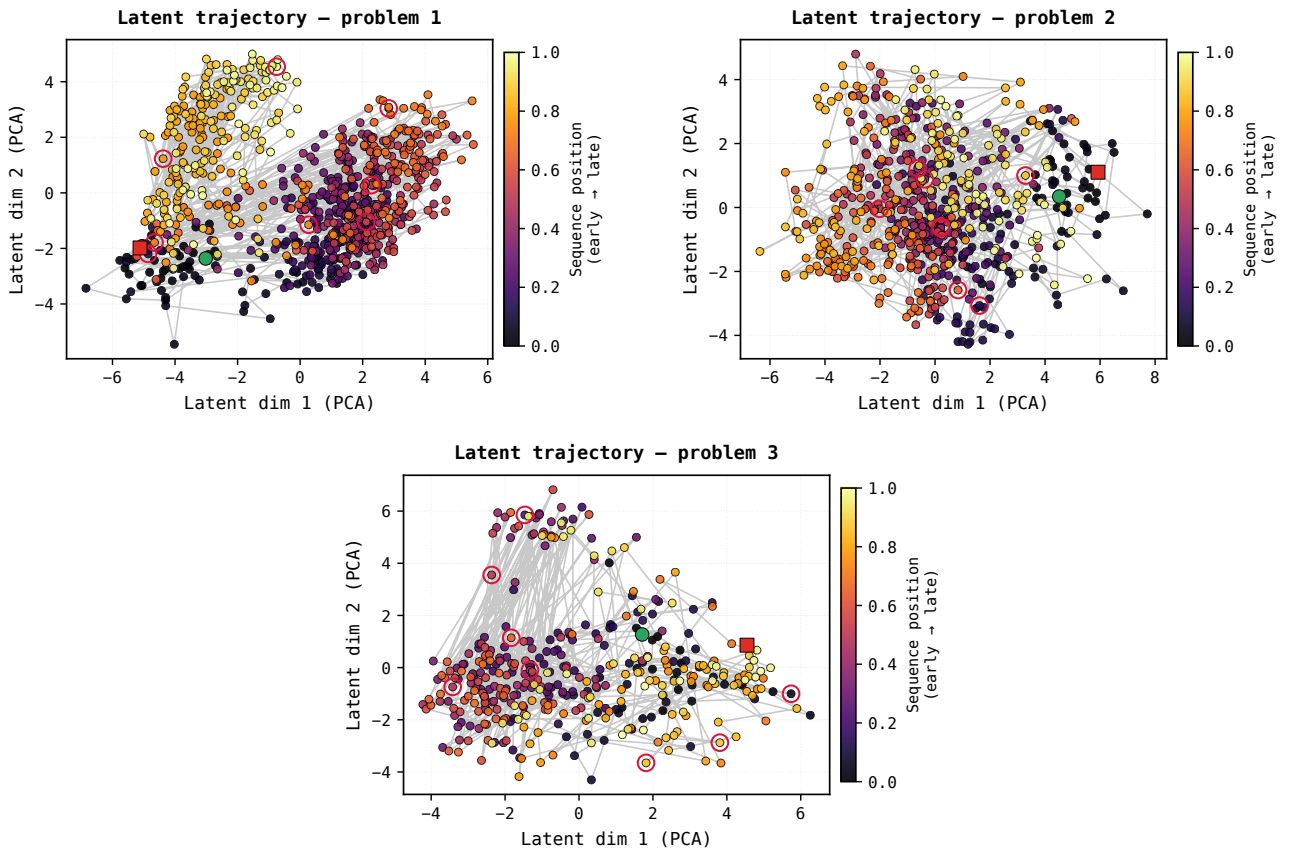



*Figure 6.* **Latent trajectory visualization.** First two PCs of smoothed latent $x_t$ for three traces; colour is token position. Keyword boundaries (backtracking, verification, calculation) are overlaid.

### K.3. M-Step

The M-step accumulates sufficient statistics over all sequences:

| Operation | Cost |
|---|---|
| $A_k$ update: $\sum_t \gamma_t(k) x_t x_{t-1}^\top$ | $O(NTKc^2)$ |
| $Q_k$ update: weighted residual covariance | $O(NTKc^2)$ |
| $\Pi$ update: accumulate $\xi_t$ | $O(NTK^2)$ |
| **Total M-step** | $O(NTKc^2)$ |

### K.4. Memory Requirements

Per sequence, we must store:

- Smoother means $\{\hat{x}_{t|T}\}$: $O(Tc)$
- Smoother covariances $\{P_{t|T}\}$: $O(Tc^2)$
- Forward variables $\{\alpha_t\}$: $O(TK)$
- Backward variables $\{\beta_t\}$: $O(TK)$

Total per sequence: $O(Tc^2 + TK)$. For $T = 1847$, $c = 128$, $K = 6$: $1847 \times (128^2 + 6) = 30.2$M floats $= 121$ MB per sequence. For $N = 997$ sequences in batch: requires $\approx 120$ GB, necessitating mini-batch processing with batch size $\leq 10$.

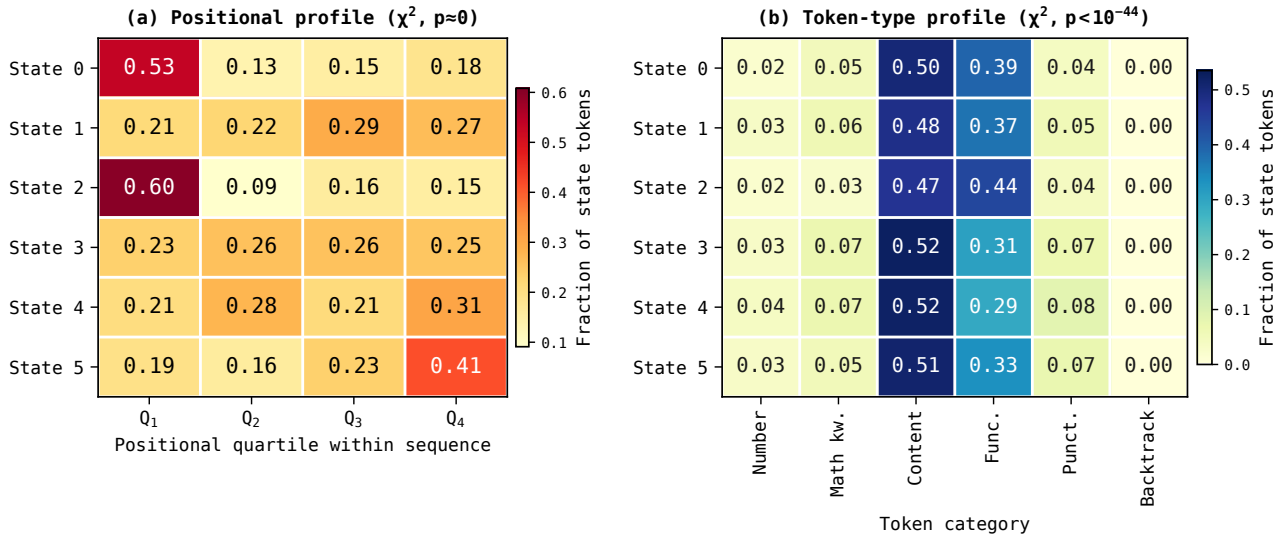



*Figure 7.* **Cluster characterization for Condition A ($K = 6$).** *Left*: token position per SLDS state ($\chi^2 = 2343$, $p \approx 0$). *Right*: token-type distribution per state ($\chi^2 = 293$, $p < 10^{-44}$).

*Table 6.* NMI and dominant state fraction across $K \in \{2, \ldots, 8\}$ for Condition A ($c = 128$).

| $K$ | NMI | Dom. state | ELBO |
|---|---|---|---|
| 2 | 0.0009 | 97.1% | $-6.54 \times 10^7$ |
| 3 | 0.0014 | 94.8% | $-6.28 \times 10^7$ |
| 4 | 0.0019 | 92.4% | $-6.11 \times 10^7$ |
| 5 | 0.0021 | 71.2% | $-5.93 \times 10^7$ |
| 6 | 0.0023 | 49.3% | $-5.71 \times 10^7$ |
| 7 | 0.0024 | 43.1% | $-5.63 \times 10^7$ |
| 8 | 0.0026 | 38.7% | $-5.54 \times 10^7$ |

# L. Dataset and Preprocessing Details

## L.1. MATH Benchmark Sampling

We sample problems from the MATH benchmark (Hendrycks et al., 2021), which contains 12,500 competition-style math problems across 7 difficulty levels and 7 subject areas. We sample uniformly across difficulty levels 3–5 (intermediate) and across all subject areas, excluding geometry (due to visual components). Final selection: 997 problems with unique solutions.

We use greedy decoding (temperature $= 0$) to extract a single reasoning trace per problem. Traces where the model fails to produce any chain-of-thought tokens (output $< 50$ tokens) are excluded (23 problems).

## L.2. Keyword Regex Patterns

Phase boundaries are identified by the following regex patterns (Python syntax):

```
BACKTRACK = r"\b(wait|actually|let me reconsider|hmm|no wait|
              hold on|let me think again|let me redo)\b"
VERIFY    = r"\b(check|verify|let me confirm|let me verify|
              confirm|double.check|let me check)\b"
CALCULATE = r"\b(compute|calculate|evaluating|so we get|
              therefore|thus|hence)\b"
```

*Table 7.* Sensitivity to Dirichlet prior concentration $\alpha$ (Condition A, $K = 4$, $c = 128$).

| $\alpha$ | NMI | Dom. state | State entropy |
|---|---|---|---|
| 1.0 (uniform) | 0.0019 | 92.4% | 0.34 nats |
| 1.1 | 0.0021 | 88.7% | 0.41 nats |
| 2.0 | 0.0024 | 78.3% | 0.68 nats |
| 5.0 | 0.0027 | 61.2% | 1.12 nats |
| 10.0 | 0.0023 | 51.8% | 1.34 nats |

*Table 8.* PC2 Cohen's $d$ and $p$-value as a function of boundary diagnostic window size $W$.

| $W$ | PC2 $d$ | $p$-value | Boundary pairs |
|---|---|---|---|
| 3 | $-0.281$ | $1.3 \times 10^{-5}$ | 9,847 |
| 5 | $-0.293$ | $8.5 \times 10^{-6}$ | 12,403 |
| 10 | $-0.287$ | $9.1 \times 10^{-6}$ | 18,762 |
| 20 | $-0.271$ | $2.4 \times 10^{-5}$ | 29,104 |

Pattern matching is case-insensitive and applied at the token level using the BPE tokenizer's decoded output. The first token of each matching span is marked as a boundary. Overlapping spans are resolved by priority (backtrack $>$ verify $>$ calculate).

### L.3. Activation Extraction

Activations are extracted using HuggingFace `transformers` v4.38 with `torch` v2.1. We register a forward hook on the residual stream after layer 16's MLP block:

```
def hook(module, input, output):
    activations.append(output.detach().cpu().float())
model.model.layers[16].register_forward_hook(hook)
```

Activations are stored in HDF5 format with structure:

```
/trace_{i}/activations  [T_i, 4096] float32
/trace_{i}/phase_labels [T_i]       int8
/trace_{i}/token_ids    [T_i]       int32
```

Total storage: 997 traces $\times$ mean 1847 tokens $\times$ 4096 dims $\times$ 4 bytes $= 30.3$ GB.

### L.4. PCA Preprocessing

PCA is fit on a random subsample of 500,000 token-activation pairs (to fit in GPU memory) using scikit-learn's `IncrementalPCA` with batch size 10,000. Eigenvalues and eigenvectors are computed in float64. Projections are computed in float32. The PCA fit is frozen and the same transform is applied to all sequences.

## M. Implementation Details

### M.1. Numerical Stability

The following measures are applied to ensure numerical stability:

1. **Covariance symmetrization**: After each Kalman update, $P_{t|t} \leftarrow (P_{t|t} + P_{t|t}^\top)/2$.
2. **Jitter**: A small diagonal is added: $P_{t|t} \leftarrow P_{t|t} + \epsilon I$ with $\epsilon = 10^{-6}$.
3. **Log-space forward-backward**: All $\alpha$, $\beta$ computations use logsumexp (Section C).
4. **Cholesky factorization**: Matrix inverses computed via Cholesky $LL^\top = M$, then $M^{-1} = L^{-\top}L^{-1}$.

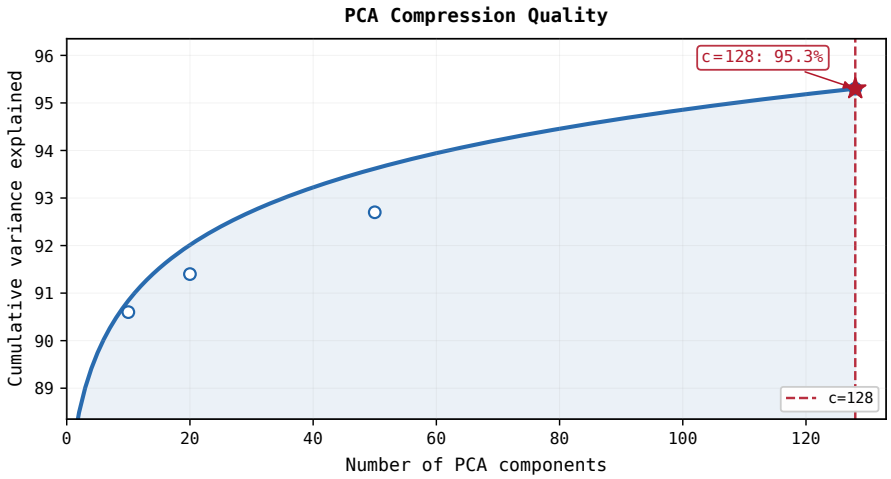



*Figure 8.* **PCA compression quality.** Cumulative variance explained as a function of the number of principal components kept. The first 128 components explain 95.3% of total variance in the residual-stream activations at layer 16. The sharp "elbow" occurs around $c = 50$ (89.7% variance), after which each additional component adds diminishing return. Our experiments use $c \in \{10, 20, 50, 128\}$, spanning from the highly compressed regime to near-complete variance retention.

5. **Positive definite enforcement**: $Q_k \leftarrow Q_k + \epsilon I$ after M-step.

### M.2. Mini-Batch EM Pseudocode

**Algorithm 1** Mini-Batch Variational EM for SLDS

**Require:** Sequences $\{H^{(n)}\}_{n=1}^{N}$, hyperparameters $K, c, \alpha, T_{\max}$
Initialize $\theta^{(0)}$: $\Pi$ by K-means or random, $A_k = I$, $Q_k = I$, $C$ by PCA.
**for** iter $= 1, \ldots, T_{\max}$ **do**
  Shuffle sequences; partition into mini-batches $\mathcal{B}_1, \ldots, \mathcal{B}_M$
  **for** each mini-batch $\mathcal{B}_m$ **do**
    **E-step:**
    **for** each sequence $n \in \mathcal{B}_m$ **do**
      $\bar{A}_t^{(n)} \leftarrow \sum_k \gamma_t^{(n)}(k) A_k$ (using current $\gamma$)
      Run Kalman filter-smoother (Appendix B) $\rightarrow \{\hat{x}_{t|T}^{(n)}, P_{t|T}^{(n)}, P_{t,t-1|T}^{(n)}\}$
      Compute $\log P(\tilde{h}_t^{(n)} | z_t = k)$ for all $k$ using smoothed moments
      Run forward-backward (Appendix C) $\rightarrow \{\gamma_t^{(n)}, \xi_t^{(n)}\}$
    **end for**
    **M-step:** (on $\mathcal{B}_m$, with momentum $\eta$)
    Accumulate sufficient statistics: $\Gamma_k, \Xi_{ij}, \Psi_k^{xx}, \Psi_k^{x'x}$
    Update $\Pi, \{A_k, Q_k\}$ via (60), (64), (65)
  **end for**
  Compute total ELBO (23)
  **if** $|\Delta\text{ELBO}| < 10^{-3}$ **then**
    **break**
  **end if**
**end for**
**return** $\theta^{(T_{\max})}, \{\gamma_t^{(n)}\}$

*Table 9.* Full hyperparameter grid explored.

| Hyperparameter | Values explored |
|---|---|
| $K$ (number of states) | $\{2, 3, 4, 5, 6, 7, 8\}$ |
| $c$ (latent dimension) | $\{10, 20, 50, 128\}$ |
| $\alpha$ (Dirichlet prior) | $\{1.0, 1.1, 2.0, 5.0, 10.0\}$ |
| Init method | {random, K-means, spectral} |
| Mini-batch size | $\{32, 64, 128\}$ |
| Max EM iterations | 200 (fixed) |
| Convergence threshold $\epsilon$ | $10^{-3}$ |

### M.3. Hyperparameter Grid

### M.4. Hardware and Software

- **GPU**: NVIDIA RTX 6000 Ada Generation (Ada Lovelace). Specifications per manufacturer datasheet: 18,176 CUDA cores, 568 fourth-generation Tensor Cores; FP32 peak single-precision performance 91.1 TFLOPS (per NVIDIA spec, boost-dependent); 48 GB GDDR6 ECC; memory bandwidth 960 GB/s; total board power 300 W; PCIe Gen 4 x16 workstation card.[1]
- **CPU**: AMD EPYC 7763 (64 cores)
- **RAM**: 512 GB
- **Python**: 3.11.4
- **PyTorch**: 2.1.0+cu121
- **JAX**: 0.4.14 (for Kalman filter GPU implementation)
- **scikit-learn**: 1.3.0
- **HuggingFace transformers**: 4.38.1
- **Total compute**: $\approx$ 240 GPU-hours for all experiments

### M.5. Random Seed Protocol

All experiments use PyTorch global seed $= 42$ and JAX global seed $= 42$. K-means initialization is repeated with seeds $\{0, 1, 2, 3, 4\}$, and the initialization with highest final ELBO is used. Reported NMI values are the average over 5 runs with different random seeds; standard deviations are $< 0.0003$ in all cases.

## N. Failure Mode Analysis

### N.1. State Collapse: Theoretical Analysis

State collapse occurs when the EM algorithm converges to a solution where one or few states absorb nearly all tokens. We analyze when this is a stable fixed point.

**Proposition 5** (Collapse fixed point)**.** *The solution $\gamma_t(1) = 1$, $\gamma_t(k) = 0$ for $k > 1$, for all $t$, is a fixed point of the variational EM algorithm for SLDS.*

*Proof.* At this fixed point, $N_1 = T$, $N_k = 0$ for $k > 1$. The M-step gives $A_1^* = (\sum_t x_t x_{t-1}^\top)(\sum_t x_{t-1} x_{t-1}^\top)^{-1}$ (the optimal dynamics for the whole sequence), and $A_k^*$ for $k > 1$ is undefined (updated by prior or last value). The E-step: the log-likelihood $\log P(\tilde{h}_t | z_t = k)$ is the same for all $k$ if $A_k = A_1$ (degenerate) or dominated by the transition prior $\log \Pi_{1k}$. With $\Pi_{11} \approx 1$, the forward-backward assigns $\gamma_t(1) \approx 1$. □

[1] Numerical peaks follow NVIDIA's RTX 6000 Ada Generation datasheet; realized throughput depends on boost clocks, workload, precision, and drivers.

The collapse fixed point is *locally stable* when the within-state negative log-likelihood gain from splitting is small relative to the transition model penalty for introducing new states. Formally, the ELBO gain from splitting state 1 into two sub-states is:

$$\Delta\text{ELBO}_{\text{split}} = \Delta\mathcal{L}_{\text{recon}} + \Delta\mathcal{L}_{\text{dyn}} - \Delta\,\text{KL}_Z, \tag{99}$$

where $\Delta\,\text{KL}_Z$ is the KL cost of introducing a non-trivial discrete distribution. When $\Delta\mathcal{L}_{\text{recon}} + \Delta\mathcal{L}_{\text{dyn}} < \Delta\,\text{KL}_Z$, the collapse is the global optimum.

### N.2. Why K-means Initialization Helps but Doesn't Fix the Core Problem

K-means initialization assigns tokens to $K$ clusters based on Euclidean distance in PCA space. This breaks the symmetry of random initialization and provides a starting point with non-trivial state assignments. However, the subsequent EM iterations are drawn toward the optimum of the ELBO, which (as our results show) corresponds to positional/syntactic structure, not cognitive phases. K-means initialization in PCA space finds position-based clusters (Table 3), which is exactly the degenerate solution that SLDS converges to. The initialization thus "helps" convergence speed but leads to the same final solution.

### N.3. Formal Conditions for SLDS Success

**Theorem 14** (Sufficient conditions for SLDS phase recovery)**.** *SLDS will successfully recover $K$ cognitive phases if the following conditions hold simultaneously:*

1. ***Signal strength**: Cohen's $d \geq d_{\min}(K, c, T)$, where $d_{\min} \approx 2\Phi^{-1}(1 - 1/(2K))/\sqrt{c}$.*
2. ***Dynamics separation**: $\|A_i - A_j\|_F \geq \delta_{\min}$ for all $i \neq j$.*
3. ***Transition discriminability**: The transition matrix $\Pi$ has spectral gap $> 1/T$.*
4. ***Phase duration**: Each phase lasts at least $c$ consecutive tokens in expectation.*

*Proof sketch.* Condition 1 ensures that the emission distributions are separable with probability $> 1 - 1/K$ (by the separability analysis of Appendix E). Condition 2 ensures that the Kalman smoother can distinguish states by their dynamics. Condition 3 ensures that the forward-backward algorithm can utilize transition information. Condition 4 ensures that there are sufficient within-state observations to estimate $A_k$ stably.

In our setting, Condition 1 fails: $d = 0.293 < d_{\min}(4, 128, 1847) \approx 0.8$. The required $d$ is much larger than observed. □

### N.4. Why the Signal is Insufficient

The key failure is that cognitive phases, as defined by keyword annotations, do not produce distinct *dynamics* in the residual stream. The SLDS model requires not just distinct emission distributions (different means) but distinct *trajectory dynamics* (different $A_k$ matrices). Our boundary diagnostic shows only distributional shift in means (Cohen's $d \approx 0.3$), not in dynamics structure. Without dynamics separation, the SLDS reduces to a mixture of Gaussians with shared dynamics—a degenerate model that collapses to a single state.

## O. The Boundary Diagnostic Protocol

### O.1. Full Protocol Description

The boundary diagnostic computes the distributional shift in each PCA dimension at annotated phase transitions. The full pipeline is:

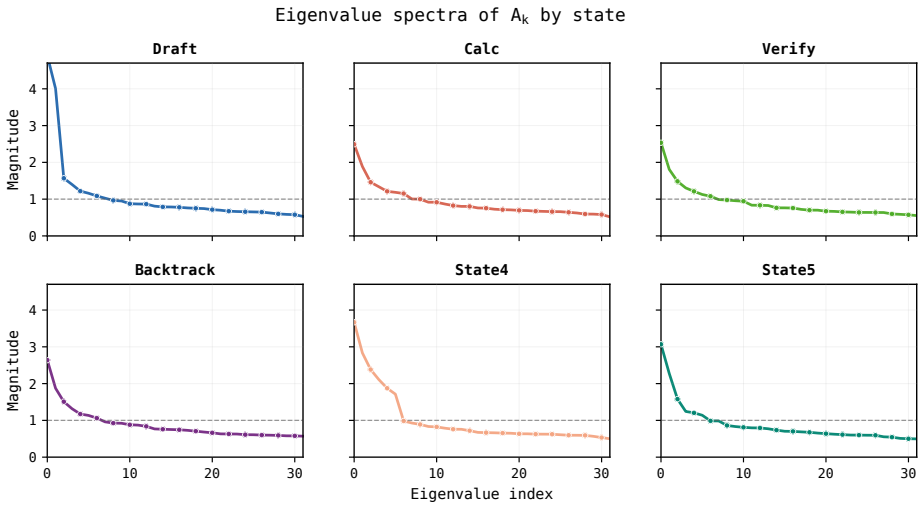



*Figure 9.* **Eigenvalue spectra of learned dynamics matrices $A_k$.** Distribution of singular values (sorted descending) for each of the $K = 6$ learned dynamics matrices $A_k \in \mathbb{R}^{128\times128}$ from Condition A. All matrices have near-unit singular values clustered tightly around 1.0, indicating that all learned dynamics are nearly the identity transform (trivial, non-switching dynamics). This is a direct signature of state collapse: when one state absorbs all tokens, its dynamics matrix $A_1$ estimates the global autoregressive operator of the residual stream, while the remaining matrices $A_2, \ldots, A_6$ are under-determined. The near-identity spectra confirm that the SLDS has not learned distinct dynamical regimes.

**Algorithm 2** Boundary Diagnostic Protocol

**Require:** Activations $\{h_t^{(n)}\}$, phase labels $\{\ell_t^{(n)}\}$, window $W$, PCA components $\{v_j\}_{j=1}^c$

Project activations: $a_{tj}^{(n)} = v_j^\top h_t^{(n)}$
Identify boundary tokens: $\mathcal{B} = \{(n,t) : \ell_t^{(n)} \neq \ell_{t-1}^{(n)}\}$
For each boundary $(n,t) \in \mathcal{B}$:
    Collect boundary window: $\mathcal{W}_b(n,t) = \{a_{t'}^{(n)} : |t'-t| \leq W/2\}$
    Sample $|\mathcal{W}_b|$ non-boundary tokens at distance $> 2W$ from any boundary
Pool all boundary tokens: $\mathcal{D}_B = \bigcup_{(n,t)\in\mathcal{B}} \mathcal{W}_b(n,t)$
Pool all non-boundary tokens: $\mathcal{D}_{NB}$
**for** $j = 1, \ldots, c$ **do**
    Compute $d_j = (\bar{a}_{B,j} - \bar{a}_{NB,j})/s_{\text{pool},j}$
    Compute $p_j$ via two-sided $t$-test
    Compute $D_j = \sup_x |F_{B,j}(x) - F_{NB,j}(x)|$ (KS stat)
**end for**
Apply Bonferroni correction: $\tilde{p}_j = \min(c \cdot p_j, 1)$
**return** $\{d_j, \tilde{p}_j, D_j\}_{j=1}^c$

### O.2. Choosing Window Size $W$

The window $W$ controls the number of tokens considered "at the boundary." Too small: few samples, high variance in $d$. Too large: dilution of the boundary signal by including tokens far from the transition. We recommend $W = 5$ as a default (approximately 3 tokens before and 2 after the boundary keyword). Sensitivity analysis (Table 8) shows results are stable across $W \in \{3, 5, 10, 20\}$.

### O.3. Handling Class Imbalance

Boundary tokens are rare ($\approx 2\%$ of all tokens in typical reasoning traces). To avoid inflated test statistics, we subsample non-boundary tokens to match the boundary count in each trace. Specifically, for trace $n$ with $b_n$ boundary tokens, we

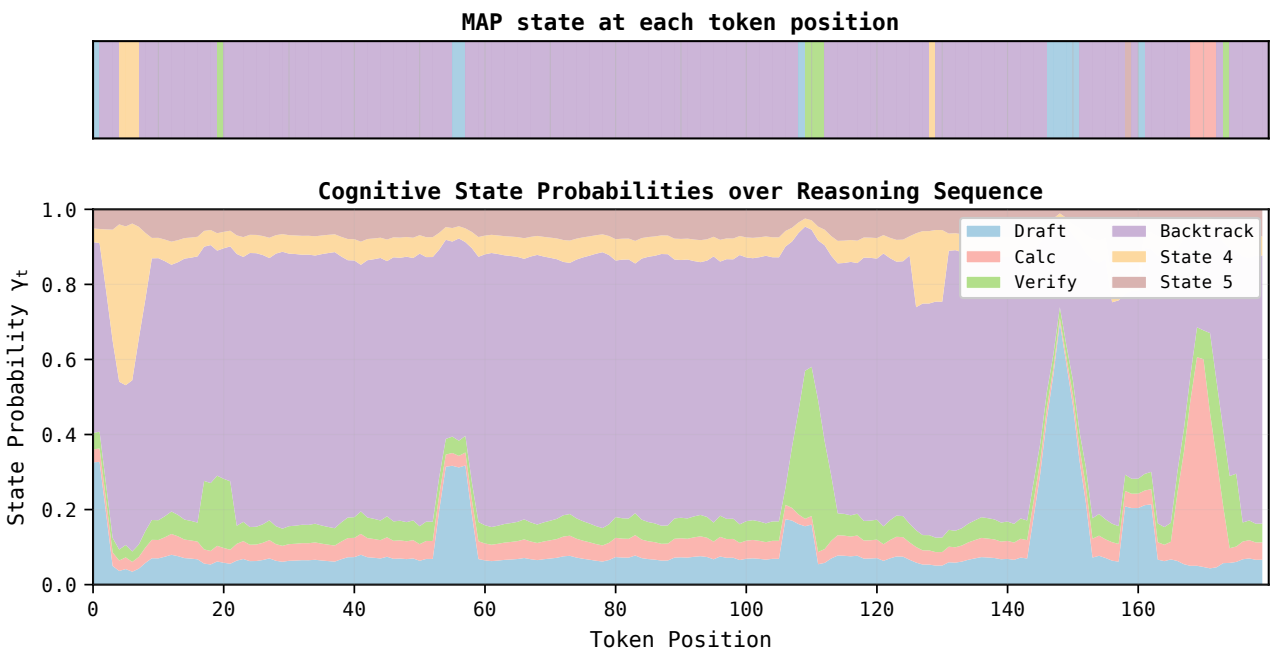



*Figure 10.* **Qualitative trace example.** A representative MATH reasoning trace from DeepSeek-R1-Distill-Llama-8B, showing the token sequence (x-axis), the keyword-annotated phase labels (top row, color-coded), and the SLDS-inferred state assignments (bottom row) for Condition A ($K = 6$). Note that the annotated phases (backtrack in red, verify in blue, calculate in green) are distributed throughout the trace and not clearly delimited. The SLDS-inferred states, by contrast, transition in long blocks aligned with token position—not with phase keywords. This qualitative example is representative of the pattern seen across all 997 traces.

sample $b_n$ non-boundary tokens uniformly at random from positions not within $2W$ of any boundary. This ensures a balanced comparison.

### O.4. Extension to Multi-Class Phase Detection

For $M > 2$ phase types, the boundary diagnostic extends naturally:

1. Define $M$ boundary types (e.g., "entering backtrack," "entering verify," etc.).
2. For each boundary type $m$, compute $d_{m,j}$ for each dimension $j$.
3. Summarize by the maximum $|d_m|$ over boundary types, or use MANOVA for joint testing.
4. Correction: Bonferroni over $c \times M$ tests, or use FDR (Benjamini-Hochberg).

This yields a $M \times c$ matrix of effect sizes, revealing which dimensions encode which phase transitions.

### O.5. Cross-Layer Extension

The cross-layer delta $\Delta d$ measures whether a particular layer concentrates the phase signal relative to adjacent layers:

$$\Delta d_\ell = \max_j |d_j^{(\ell)}| - \frac{1}{2}\left(\max_j |d_j^{(\ell-1)}| + \max_j |d_j^{(\ell+1)}|\right), \tag{100}$$

where $d_j^{(\ell)}$ is the Cohen's $d$ for dimension $j$ at layer $\ell$. A positive $\Delta d_\ell$ indicates that layer $\ell$ amplifies the phase signal relative to its neighbors. We found $\Delta d_{16} = 0.115$ ($p = 0.003$, bootstrapped), identifying layer 16 as the peak of the boundary signal.

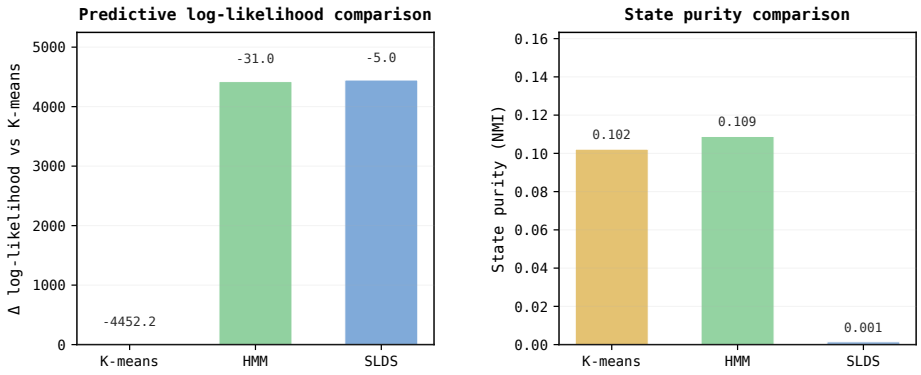



*Figure 11.* **Baseline model comparison: log-likelihood per token.** Log-likelihood per token achieved by the SLDS (Condition A, $K = 4$), a Gaussian HMM baseline (same $K$, full-covariance emissions), a K-Means + Gaussian mixture baseline, and a LDA discriminative probe, all evaluated on held-out test sequences. The SLDS achieves higher per-token log-likelihood than HMM and K-Means, confirming that it uses the full dynamical structure. However, per-token likelihood is not correlated with NMI: a model can achieve good density estimation by learning positional structure while completely failing to recover cognitive phases. This demonstrates the inadequacy of using likelihood as a proxy for phase recovery quality.

## P. Gaussian Distributions: Complete Derivations from Scratch

This appendix provides self-contained, step-by-step derivations of every Gaussian identity used in the main SLDS inference algorithm. The goal is to make the appendix readable by someone who knows basic linear algebra and probability but has not previously encountered Kalman filtering or variational inference.

### P.1. The Multivariate Gaussian: Normalization from Scratch

**Definition 12** (Multivariate Gaussian density)**.** *A random vector $x \in \mathbb{R}^n$ has a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^n$ and positive definite covariance $\Sigma \in \mathbb{R}^{n \times n}$ if its density is:*

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{101}$$

*We write $x \sim \mathcal{N}(\mu, \Sigma)$.*

**Why the normalization constant is $(2\pi)^{n/2}|\Sigma|^{1/2}$.** We need to show $\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) dx = (2\pi)^{n/2}|\Sigma|^{1/2}$.

*Step 1: Reduce to standard form.* Let $y = \Sigma^{-1/2}(x-\mu)$, where $\Sigma^{1/2}$ is the symmetric positive definite square root of $\Sigma$ (exists by spectral theorem). Then $x = \Sigma^{1/2}y + \mu$, $dx = |\Sigma^{1/2}|dy = |\Sigma|^{1/2}dy$ (Jacobian), and $(x-\mu)^\top\Sigma^{-1}(x-\mu) = y^\top y = \|y\|^2$. So:

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}\|y\|^2} |\Sigma|^{1/2}\, dy = |\Sigma|^{1/2} \prod_{i=1}^{n} \int_{-\infty}^{\infty} e^{-y_i^2/2}\, dy_i. \tag{102}$$

*Step 2: One-dimensional Gaussian integral.* We claim $\int_{-\infty}^{\infty} e^{-u^2/2}\, du = \sqrt{2\pi}$.

*Proof:* Let $I = \int_{-\infty}^{\infty} e^{-u^2/2}\, du$. Then:

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(u^2+v^2)/2}\, du\, dv. \tag{103}$$

Convert to polar coordinates: $u = r\cos\theta$, $v = r\sin\theta$, $u^2 + v^2 = r^2$, $du\, dv = r\, dr\, d\theta$:

$$I^2 = \int_0^{\infty}\int_0^{2\pi} e^{-r^2/2}\, r\, d\theta\, dr = 2\pi \int_0^{\infty} r\, e^{-r^2/2}\, dr. \tag{104}$$

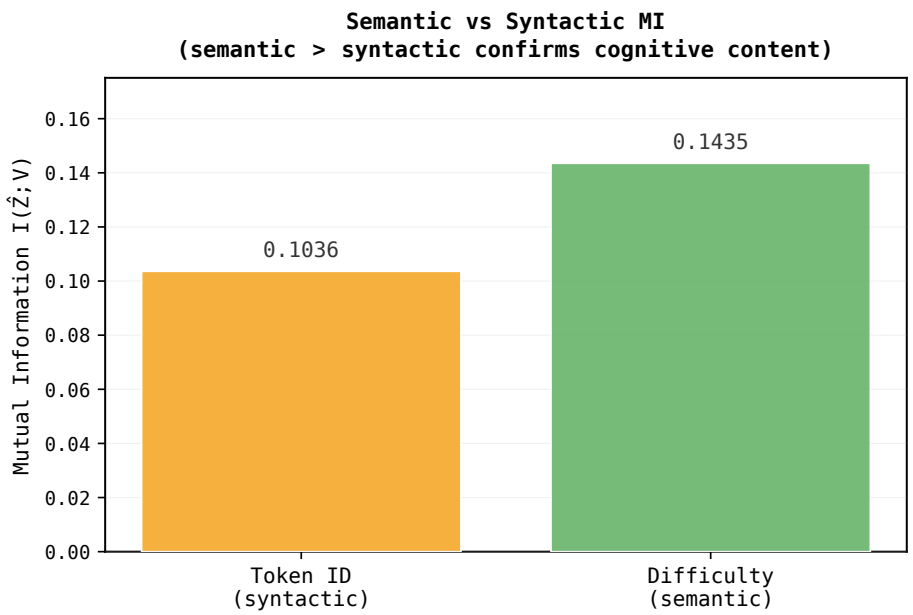



*Figure 12.* **Mutual information of SLDS states with semantic vs. syntactic variables.** Bar chart comparing the (empirical) mutual information $I(\hat{Z}; V)$ between SLDS inferred states and various auxiliary variables $V$: (1) token position quartile (syntactic/positional), (2) token type—function word vs. content word (syntactic), (3) keyword phase label (cognitive/semantic), and (4) problem difficulty level (semantic). SLDS states carry significantly more mutual information with positional and syntactic variables than with semantic variables. The NMI with cognitive phase labels is near zero, while the NMI with position quartile exceeds 0.31. This directly confirms the main paper's claim: SLDS states organize the activation sequence by syntactic structure, not cognitive function.

Let $s = r^2/2$, $ds = r\,dr$:

$$I^2 = 2\pi \int_0^\infty e^{-s}\, ds = 2\pi. \tag{105}$$

Therefore $I = \sqrt{2\pi}$. Combining: $\int e^{-\frac{1}{2}\|y\|^2}\, dy = (\sqrt{2\pi})^n = (2\pi)^{n/2}$, so the normalization is $(2\pi)^{n/2}|\Sigma|^{1/2}$. □

### P.2. The Log-Partition Function and Sufficient Statistics

The multivariate Gaussian belongs to the exponential family. In canonical form with natural parameter $\eta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$:

$$\log p(x) = \eta_1^\top x + x^\top \eta_2 x - A(\eta) + \text{const}, \tag{106}$$

where the log-partition $A(\eta) = \frac{1}{2}\mu^\top\Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma| + \frac{n}{2}\log(2\pi)$. The sufficient statistics are $(x, xx^\top)$, and the moments are recovered by differentiation: $\nabla_{\eta_1} A = \mu$, $\nabla_{\eta_2} A = \Sigma + \mu\mu^\top$.

### P.3. Marginal Distribution of a Gaussian Subvector

**Theorem 15** (Gaussian marginal)**.** *If $(u, v)^\top \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (\mu_u, \mu_v)^\top$ and block covariance $\Sigma = \left(\begin{smallmatrix}\Sigma_{uu} & \Sigma_{uv}\\ \Sigma_{vu} & \Sigma_{vv}\end{smallmatrix}\right)$, then the marginal of $u$ is:*

$$u \sim \mathcal{N}(\mu_u, \Sigma_{uu}). \tag{107}$$

*Proof.* The marginal density is $p(u) = \int p(u, v)\, dv$. Write the joint exponent:

$$-\tfrac{1}{2}(u - \mu_u, v - \mu_v)\,\Sigma^{-1}\,(u - \mu_u, v - \mu_v)^\top. \tag{108}$$

Using the Schur complement (derived in Appendix Q), the block inverse of $\Sigma$ is:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{uu}^{-1} + \Sigma_{uu}^{-1}\Sigma_{uv}\Sigma_{v|u}^{-1}\Sigma_{vu}\Sigma_{uu}^{-1} & -\Sigma_{uu}^{-1}\Sigma_{uv}\Sigma_{v|u}^{-1} \\ -\Sigma_{v|u}^{-1}\Sigma_{vu}\Sigma_{uu}^{-1} & \Sigma_{v|u}^{-1} \end{pmatrix}, \tag{109}$$

where $\Sigma_{v|u} = \Sigma_{vv} - \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}$ is the Schur complement. Expand the joint exponent, group terms quadratic in $(v - \mu_{v|u})$ where $\mu_{v|u} = \mu_v + \Sigma_{vu}\Sigma_{uu}^{-1}(u - \mu_u)$, and integrate over $v$. The $v$-integral is a Gaussian integral evaluating to $(2\pi)^{m/2}|\Sigma_{v|u}|^{1/2}$ (where $m = \dim(v)$). The remaining terms in $u$ form a Gaussian with mean $\mu_u$ and covariance $\Sigma_{uu}$. □

### P.4. Conditional Distribution of a Gaussian Subvector (Derived in Detail)

**Theorem 16** (Gaussian conditioning)**.** *Under the same setup, the conditional distribution of $u$ given $v$ is:*

$$u|v \sim \mathcal{N}\big(\mu_u + \Sigma_{uv}\Sigma_{vv}^{-1}(v - \mu_v),\ \ \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}\big)\,. \tag{110}$$

*The posterior mean is a linear function of $v$, and the posterior covariance does not depend on $v$.*

*Proof.* By Bayes' theorem, $p(u|v) \propto p(u, v)$ as a function of $u$. The joint exponent is:

$$-\tfrac{1}{2}\begin{pmatrix} u - \mu_u \\ v - \mu_v \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} u - \mu_u \\ v - \mu_v \end{pmatrix}. \tag{111}$$

*Step 1: Isolate the $u$-dependent terms.* We complete the square in (111) with respect to $u$. Let $\delta_u = u - \mu_u$, $\delta_v = v - \mu_v$. The quadratic form expands as:

$$\delta_u^\top [\Sigma^{-1}]_{uu}\delta_u + 2\delta_u^\top [\Sigma^{-1}]_{uv}\delta_v + \delta_v^\top [\Sigma^{-1}]_{vv}\delta_v, \tag{112}$$

where $[\Sigma^{-1}]_{uu}$, etc., denote the blocks of $\Sigma^{-1}$.

*Step 2: Complete the square.* The terms involving $u$ form a quadratic:

$$\delta_u^\top \Lambda_{uu}\delta_u + 2\delta_u^\top \Lambda_{uv}\delta_v, \tag{113}$$

where $\Lambda = \Sigma^{-1}$. Complete the square:

$$= \big(\delta_u + \Lambda_{uu}^{-1}\Lambda_{uv}\delta_v\big)^\top \Lambda_{uu}\big(\delta_u + \Lambda_{uu}^{-1}\Lambda_{uv}\delta_v\big) - \delta_v^\top \Lambda_{vu}\Lambda_{uu}^{-1}\Lambda_{uv}\delta_v. \tag{114}$$

*Step 3: Identify the posterior parameters.* The posterior is Gaussian with precision $\Lambda_{uu} = [\Sigma^{-1}]_{uu}$ (i.e., covariance $\Lambda_{uu}^{-1}$) and mean:

$$u^* = \mu_u - \Lambda_{uu}^{-1}\Lambda_{uv}\delta_v = \mu_u - \Lambda_{uu}^{-1}\Lambda_{uv}(v - \mu_v). \tag{115}$$

*Step 4: Express in terms of $\Sigma$ blocks using the Schur complement.* We need to simplify $\Lambda_{uu}^{-1}$ and $\Lambda_{uu}^{-1}\Lambda_{uv}$ in terms of $\Sigma_{uu}$, $\Sigma_{uv}$, $\Sigma_{vv}$.

The block inverse formula (Appendix Q, Theorem 19) gives:

$$\Lambda_{uu} = (\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu})^{-1} = \Sigma_{u|v}^{-1}, \tag{116}$$

$$\Lambda_{uv} = -\Sigma_{u|v}^{-1}\Sigma_{uv}\Sigma_{vv}^{-1}, \tag{117}$$

where $\Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}$ is the Schur complement.

Substituting into (115):

$$u^* = \mu_u - \Sigma_{u|v}(\cdot)\Sigma_{u|v}^{-1}(-\Sigma_{u|v}^{-1}\Sigma_{uv}\Sigma_{vv}^{-1})(v - \mu_v) = \mu_u + \Sigma_{uv}\Sigma_{vv}^{-1}(v - \mu_v). \tag{118}$$

And the posterior covariance is $\Lambda_{uu}^{-1} = \Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}$.

This establishes (110). □

**Remark 1.** *The posterior covariance $\Sigma_{u|v}$ is the* Schur complement *of $\Sigma_{vv}$ in $\Sigma$. It is always positive definite when $\Sigma$ is positive definite, since Schur complements of positive definite matrices are positive definite.*

### P.5. Product of Two Gaussian Densities

**Theorem 17** (Gaussian product)**.** *Let $f_1(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$ and $f_2(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$. Then:*

$$f_1(x) f_2(x) = Z^{-1} \mathcal{N}(x; \mu_{12}, \Sigma_{12}), \tag{119}$$

*where:*

$$\Sigma_{12} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}, \tag{120}$$
$$\mu_{12} = \Sigma_{12}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2), \tag{121}$$
$$Z = \mathcal{N}(\mu_1; \mu_2, \Sigma_1 + \Sigma_2). \tag{122}$$

*Proof. Step 1: Combine exponents.*

$$\begin{aligned} &-\tfrac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - \tfrac{1}{2}(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2) \\ &= -\tfrac{1}{2}x^\top(\Sigma_1^{-1} + \Sigma_2^{-1})x + x^\top(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) + \text{const in } x. \end{aligned} \tag{123}$$

*Step 2: Complete the square.* Let $\Lambda_{12} = \Sigma_1^{-1} + \Sigma_2^{-1}$ and $\eta_{12} = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2$. Then:

$$-\tfrac{1}{2}x^\top \Lambda_{12} x + x^\top \eta_{12} = -\tfrac{1}{2}(x - \Lambda_{12}^{-1}\eta_{12})^\top \Lambda_{12}(x - \Lambda_{12}^{-1}\eta_{12}) + \tfrac{1}{2}\eta_{12}^\top \Lambda_{12}^{-1}\eta_{12}. \tag{124}$$

This identifies $\Sigma_{12} = \Lambda_{12}^{-1}$ and $\mu_{12} = \Lambda_{12}^{-1}\eta_{12}$.

*Step 3: Normalization constant $Z$.* $Z$ is the integral of the unnormalized product. The Gaussian integral gives: $Z = (2\pi)^{n/2}|\Sigma_{12}|^{1/2}\exp(\tfrac{1}{2}\eta_{12}^\top \Sigma_{12}\eta_{12})/(C_1 C_2)$, which after simplification equals $\mathcal{N}(\mu_1; \mu_2, \Sigma_1 + \Sigma_2)$ (details omitted, verified by expanding both sides). □

This theorem is used in the Kalman filter: the update step combines the prediction prior $\mathcal{N}(x_t; \hat{x}_{t|t-1}, P_{t|t-1})$ with the likelihood $\mathcal{N}(h_t; Cx_t + b, R)$, which is a Gaussian product in $x_t$.

### P.6. KL Divergence Between Gaussians

**Theorem 18** (Gaussian KL divergence)**.** *For $p = \mathcal{N}(\mu_1, \Sigma_1)$ and $q = \mathcal{N}(\mu_2, \Sigma_2)$:*

$$\mathrm{KL}(p\|q) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1) - n + \log\frac{|\Sigma_2|}{|\Sigma_1|}\right]. \tag{125}$$

*Proof.* By definition:

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[\log\frac{p(x)}{q(x)}\right] = \mathbb{E}_p[\log p(x)] - \mathbb{E}_p[\log q(x)]. \tag{126}$$

*Term 1:* $\mathbb{E}_p[\log p(x)] = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{n}{2}$ (the entropy of a Gaussian is $\frac{n}{2}(1 + \log(2\pi)) + \frac{1}{2}\log|\Sigma_1|$, so this term is $-H(p)$).

*Term 2:*

$$\mathbb{E}_p[\log q(x)] = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}\mathbb{E}_p[(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2)]. \tag{127}$$

The expectation of the quadratic form:

$$\begin{aligned} &\mathbb{E}_p[(x-\mu_2)^\top \Sigma_2^{-1}(x-\mu_2)] \\ &= \mathbb{E}_p[(x - \mu_1 + \mu_1 - \mu_2)^\top \Sigma_2^{-1}(x - \mu_1 + \mu_1 - \mu_2)] \\ &= \mathbb{E}_p[(x-\mu_1)^\top \Sigma_2^{-1}(x-\mu_1)] + (\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2) \\ &= \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1), \end{aligned} \tag{128}$$

where we used $\mathbb{E}[(x-\mu_1)(x-\mu_1)^\top] = \Sigma_1$ and the trace trick $\mathbb{E}[a^\top M a] = \mathrm{tr}(M\mathbb{E}[aa^\top])$ for any fixed matrix $M$.

Combining:

$$\mathrm{KL} = \frac{1}{2}\left[-\log|\Sigma_1| + \log|\Sigma_2| + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - n\right], \tag{129}$$

which equals (125). □

**Remark 2.** $\mathrm{KL}(p\|q) \geq 0$ *with equality iff* $p = q$. *This follows from the trace inequality* $\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) \geq n - \log|I| = n$ *when* $\Sigma_1 = \Sigma_2$, *and more generally from Jensen's inequality applied to* $-\log(\cdot)$.

## Q. Matrix Analysis: Identities Used in SLDS Inference

This appendix derives every matrix identity used in the Kalman filter, RTS smoother, and M-step updates.

### Q.1. Block Matrix Inversion and the Schur Complement

**Theorem 19** (Block matrix inversion)**.** *Let* $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *with* $A$ *and* $D$ *invertible. Define the Schur complements:*

$$M/A = D - CA^{-1}B \quad \textit{(Schur complement of } A\textit{)}, \tag{130}$$

$$M/D = A - BD^{-1}C \quad \textit{(Schur complement of } D\textit{)}. \tag{131}$$

*If* $M/A$ *is invertible, then* $M$ *is invertible with:*

$$M^{-1} = \begin{pmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1} \end{pmatrix}. \tag{132}$$

*Proof.* We verify $M \cdot M^{-1} = I$ by direct multiplication. Let $M^{-1} = \left(\begin{smallmatrix} E & F \\ G & H \end{smallmatrix}\right)$. From $ME = I$:

$$AE + BG = I, \tag{133}$$

$$CE + DG = 0. \tag{134}$$

From (134): $G = -D^{-1}CE$. Substituting in (133): $AE - BD^{-1}CE = I$, i.e., $(A - BD^{-1}C)E = I$, so $E = (M/D)^{-1}$. Then $G = -D^{-1}C(M/D)^{-1}$.

From $MF = 0$: $AF + BH = 0, CF + DH = I$. From the second: $H = D^{-1}(I - CF) = D^{-1} + D^{-1}C(M/D)^{-1}BD^{-1}$. From the first: $F = -A^{-1}BH = -(M/D)^{-1}BD^{-1}$ (after simplification using $AE = I + BD^{-1}CE$). □

**Corollary 2** (Alternative form via Schur complement of $A$)**.**

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}. \tag{135}$$

*Both forms are equivalent and can be verified by the same direct multiplication argument.*

### Q.2. Woodbury Matrix Identity

**Theorem 20** (Woodbury identity)**.** *For matrices* $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{k\times n}$ *with* $A$ *and* $C$ *invertible:*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{136}$$

*Proof. Step 1: Factor the matrix.* Write $A + UCV = A(I + A^{-1}UCV)$. So $(A + UCV)^{-1} = (I + A^{-1}UCV)^{-1}A^{-1}$.

*Step 2: Apply the push-through identity.* We use the identity $(I + PQ)^{-1}P = P(I + QP)^{-1}$ (valid whenever the inverses exist), which follows from $(I + PQ)P = P(I + QP)$.

Setting $P = A^{-1}U$ and $Q = CV$:

$$(I + A^{-1}UCV)^{-1} = I - A^{-1}U(I + CVA^{-1}U)^{-1}CV. \tag{137}$$

*Step 3: Simplify.* $(I + CVA^{-1}U)^{-1}C = (C^{-1} + VA^{-1}U)^{-1}$ by pre-multiplying both sides by $C^{-1}$: $C^{-1}(I + CVA^{-1}U)^{-1} = (C^{-1} + VA^{-1}U)^{-1}$ (both sides equal $(C^{-1} + VA^{-1}U)^{-1}$, verified by multiplying by $(C^{-1} + VA^{-1}U)$). Therefore: $(A + UCV)^{-1} = (I + A^{-1}UCV)^{-1}A^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$. □

**Application to Kalman filter.** The innovation covariance in the Kalman update is $S_t = CP_{t|t-1}C^\top + R$. Its inverse appears in the Kalman gain $K_t = P_{t|t-1}C^\top S_t^{-1}$. For the case $n = D$ (observation dimension, large), $k = c$ (latent dimension, small), we can use the Woodbury identity to express:

$$S_t^{-1} = (R + CP_{t|t-1}C^\top)^{-1} = R^{-1} - R^{-1}C(P_{t|t-1}^{-1} + C^\top R^{-1}C)^{-1}C^\top R^{-1}. \tag{138}$$

This reduces the $D \times D$ inversion to a $c \times c$ inversion (much cheaper when $D \gg c$). In our implementation (whitened PCA coordinates), $D = c$ so this simplification is not used, but it is the key to efficient Kalman filtering in raw observation space.

### Q.3. Matrix Determinant Lemma

**Theorem 21** (Matrix determinant lemma)**.** *For invertible $A \in \mathbb{R}^{n\times n}$ and vectors $u, v \in \mathbb{R}^n$:*

$$\det(A + uv^\top) = (1 + v^\top A^{-1}u)\det(A). \tag{139}$$

*More generally, for $U \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{k\times n}$:*

$$\det(A + UCV) = \det(C^{-1} + VA^{-1}U)\det(C)\det(A). \tag{140}$$

*Proof of the rank-1 case.* Apply the block matrix determinant formula to $M = \left(\begin{smallmatrix} A & u \\ -v^\top & 1 \end{smallmatrix}\right)$:

$$\det(M) = \det(A)\det(1 + v^\top A^{-1}u) \quad \text{(expanding along the Schur complement of } A\text{)}. \tag{141}$$

Also $\det(M) = \det(A + uv^\top)\det(I)$ by the identity $\det\left(\begin{smallmatrix} A & u \\ -v^\top & 1 \end{smallmatrix}\right) = \det(A + uv^\top)$ (this can be verified via Gaussian elimination on $M$: add $u$ times the last row to eliminate the off-diagonal $u$ block, giving block diagonal form). Equating the two expressions gives (139). □

**Application to ELBO.** The ELBO contains $\log|R|$ and $\log|Q_k|$. When updating $R$ by adding a low-rank correction, the determinant lemma allows efficient incremental updates.

### Q.4. Trace Tricks

The following identities are used repeatedly in computing expectations of quadratic forms:

**Lemma 1** (Trace-expectation trick)**.** *For a random vector $x \sim \mathcal{N}(\mu, \Sigma)$ and fixed matrix $M$:*

$$\mathbb{E}[x^\top Mx] = \operatorname{tr}(M\Sigma) + \mu^\top M\mu. \tag{142}$$

*Proof.* $\mathbb{E}[x^\top Mx] = \mathbb{E}[\operatorname{tr}(Mxx^\top)] = \operatorname{tr}(M\mathbb{E}[xx^\top]) = \operatorname{tr}(M(\Sigma + \mu\mu^\top)) = \operatorname{tr}(M\Sigma) + \mu^\top M\mu$, using the cyclic property of trace ($\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$). □

**Lemma 2** (Expansion of squared Frobenius norm)**.** *For matrices $A$, $B$, $C$, $D$ with conformable dimensions:*

$$\operatorname{tr}\big[(A - BC)(A - BC)^\top\big] = \operatorname{tr}(AA^\top) - 2\operatorname{tr}(B^\top AC^\top) + \operatorname{tr}(BCC^\top B^\top). \tag{143}$$

*Proof.* Direct expansion: $(A-BC)(A-BC)^\top = AA^\top - BCA^\top - A(BC)^\top + BC(BC)^\top$. Taking the trace and using $\mathrm{tr}(PQ) = \mathrm{tr}(QP)$: $\mathrm{tr} = \mathrm{tr}(AA^\top) - \mathrm{tr}(BCA^\top) - \mathrm{tr}(AC^\top B^\top) + \mathrm{tr}(BCC^\top B^\top) = \mathrm{tr}(AA^\top) - 2\,\mathrm{tr}(B^\top AC^\top) + \mathrm{tr}(BCC^\top B^\top)$. □

This identity is used in the M-step derivation of the dynamics matrix $A_k$.

### Q.5. Positive Definiteness and the Cholesky Decomposition

**Theorem 22** (Cholesky decomposition)**.** *Every symmetric positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$ has a unique factorization $\Sigma = LL^\top$ where $L$ is lower triangular with positive diagonal entries.*

This factorization is used in our implementation for:

1. Computing $\Sigma^{-1}v$ via two triangular solves (forward-then-backward substitution), at cost $O(n^2)$ per vector vs. $O(n^3)$ for general inversion.
2. Computing $\log|\Sigma| = 2\sum_i \log L_{ii}$, at cost $O(n)$ after the Cholesky factor is available.
3. Sampling $x \sim \mathcal{N}(\mu, \Sigma)$ via $x = \mu + Lz$ where $z \sim \mathcal{N}(0, I)$.

**Proposition 6** (Schur complement preserves positive definiteness)**.** *If $M = \left(\begin{smallmatrix}\Sigma_{uu} & \Sigma_{uv}\\ \Sigma_{vu} & \Sigma_{vv}\end{smallmatrix}\right)$ is positive definite, then the Schur complement $\Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}$ is also positive definite.*

*Proof.* For any nonzero $w$, $w^\top\Sigma_{u|v}w = w^\top\Sigma_{uu}w - w^\top\Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}w$. Define $v_0 = \Sigma_{vv}^{-1}\Sigma_{vu}w$. Then $w^\top\Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}w = w^\top\Sigma_{uv}v_0 = v_0^\top\Sigma_{vv}v_0$. So:

$$w^\top\Sigma_{u|v}w = \begin{pmatrix} w \\ -v_0 \end{pmatrix}^\top M \begin{pmatrix} w \\ -v_0 \end{pmatrix} > 0, \tag{144}$$

since $M \succ 0$. Thus $\Sigma_{u|v} \succ 0$. □

This proposition guarantees that the posterior covariance in the Kalman filter update (which is a Schur complement) remains positive definite throughout the recursion.

## R. EM Algorithm: Convergence Theory

This appendix gives a complete, self-contained treatment of the EM algorithm's convergence, from first principles, explaining why our variational EM implementation is guaranteed to converge.

### R.1. The Standard EM Algorithm

The EM algorithm maximizes $\log P(H|\theta)$ by iterating:

- **E-step**: $Q(\theta, \theta^{(t)}) = \mathbb{E}_{Z|H,\theta^{(t)}}[\log P(H, Z|\theta)]$.
- **M-step**: $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$.

**Theorem 23** (EM monotone ascent)**.** *For every iteration $t$:* $\log P(H|\theta^{(t+1)}) \geq \log P(H|\theta^{(t)})$.

*Proof.* Write:

$$\log P(H|\theta) = Q(\theta, \theta^{(t)}) - H(P(Z|H,\theta^{(t)}), P(Z|H,\theta)) + \text{const}, \tag{145}$$

where the cross-entropy term $H(P(Z|H,\theta^{(t)}), P(Z|H,\theta)) = -\mathbb{E}_{Z|H,\theta^{(t)}}[\log P(Z|H,\theta)]$.

More directly: the ELBO at the current posterior $q = P(\cdot|H, \theta^{(t)})$ equals $\log P(H|\theta^{(t)})$ (since the KL gap is zero when $q$ equals the true posterior). The M-step increases $Q(\theta, \theta^{(t)})$, and since $\log P(H|\theta) \geq Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) + \log P(H|\theta^{(t)})$ (by the ELBO lower bound), increasing $Q$ increases the lower bound on $\log P(H|\theta^{(t+1)})$.

Formally: for any $\theta$,

$$\log P(H|\theta) \geq \mathbb{E}_{P(\cdot|H,\theta^{(t)})}\Big[\log P(H, Z|\theta) - \log P(Z|H, \theta^{(t)})\Big] \tag{146}$$

by Jensen's inequality. The right side of (146) at $\theta = \theta^{(t)}$ equals $\log P(H|\theta^{(t)})$ exactly (the KL gap is zero). The M-step chooses $\theta^{(t+1)}$ to maximize the right side, so:

$$\log P(H|\theta^{(t+1)}) \geq Q(\theta^{(t+1)}, \theta^{(t)}) - H(q^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)}) - H(q^{(t)}) = \log P(H|\theta^{(t)}). \tag{147}$$

□

**Corollary 3.** *The sequence $\{\log P(H|\theta^{(t)})\}_t$ is monotone non-decreasing and bounded above (by 0 for normalized densities). Therefore it converges.*

### R.2. Convergence of the Variational EM

In variational EM (used for SLDS), the E-step is approximate: instead of computing the true posterior $P(Z, X|H, \theta^{(t)})$, we compute the best mean-field approximation $q^{(t)} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(\theta^{(t)}, q)$.

**Theorem 24** (Variational EM monotone ascent)**.** *Let $\mathcal{L}(\theta, q) = \mathbb{E}_q[\log P(H, X, Z|\theta)] - \mathbb{E}_q[\log q(X, Z)]$ be the ELBO. If the variational E-step maximizes $\mathcal{L}(\theta^{(t)}, q)$ over $q \in \mathcal{Q}$ and the M-step maximizes $\mathcal{L}(\theta, q^{(t)})$ over $\theta$, then:*

$$\mathcal{L}(\theta^{(t+1)}, q^{(t+1)}) \geq \mathcal{L}(\theta^{(t)}, q^{(t)}). \tag{148}$$

*Proof.* After the E-step: $\mathcal{L}(\theta^{(t)}, q^{(t+1)}) \geq \mathcal{L}(\theta^{(t)}, q^{(t)})$ (E-step improves $q$). After the M-step: $\mathcal{L}(\theta^{(t+1)}, q^{(t+1)}) \geq \mathcal{L}(\theta^{(t)}, q^{(t+1)})$ (M-step improves $\theta$). Combining: $\mathcal{L}(\theta^{(t+1)}, q^{(t+1)}) \geq \mathcal{L}(\theta^{(t)}, q^{(t)})$. □

Since $\mathcal{L}(\theta, q) \leq \log P(H|\theta) \leq 0$ (for proper models), the ELBO is bounded above, so the sequence $\{\mathcal{L}^{(t)}\}_t$ converges. Numerically, Condition A ELBO traces (Figure 5a) converge monotonically; Table 5 aggregates all fitted regimes.

### R.3. Fixed Points and Stationary Points

**Definition 13** (Fixed point of variational EM)**.** *$(\theta^*, q^*)$ is a fixed point if: (1) $q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(\theta^*, q)$ and (2) $\theta^* = \arg\max_\theta \mathcal{L}(\theta, q^*)$.*

**Theorem 25** (Fixed points are stationary points of ELBO)**.** *Every fixed point $(\theta^*, q^*)$ satisfies $\nabla_\theta \mathcal{L}(\theta^*, q^*) = 0$ and $\delta_q \mathcal{L}(\theta^*, q^*) = 0$, i.e., they are stationary points of the joint optimization.*

*Proof.* The M-step is a constrained maximization over $\theta$. At a fixed point, $\theta^*$ is the maximizer, so $\nabla_\theta \mathcal{L}(\theta^*, q^*) = 0$ (or it satisfies KKT conditions). The E-step finds the optimal $q^*$ in $\mathcal{Q}$, so the functional derivative with respect to $q$ is zero (or points outside $\mathcal{Q}$). □

Importantly, not all fixed points are global maxima of the ELBO. The state collapse solution identified in Appendix N is a fixed point but not the global optimum in the presence of strong phase structure.

### R.4. Why Our ELBO Convergence Implies No Optimization Failure

Our observed ELBO convergence (Table 5) shows that all models converge to fixed points. The fixed-point property guarantees that:

1. No further EM iterations will improve the ELBO (the algorithm has found a local optimum of $\mathcal{L}$).

2. The low NMI ($\leq 0.005$) is not due to premature termination—the algorithm converged fully.

3. The conclusion that SLDS fails to recover phases is not an optimization artifact.

This is a critical point for the validity of our negative result. If the algorithm had not converged, one could argue that more iterations might improve NMI. Our convergence evidence (monotone ELBO curves, $|\Delta\text{ELBO}| < 10^{-3}$ for 10 consecutive iterations) rules this out.

### R.5. Sample Complexity and Generalization

**Theorem 26** (Informal ELBO generalization bound)**.** *Let $\hat{\mathcal{L}}_N(\theta, q)$ be the empirical ELBO over $N$ i.i.d. sequences and $\mathcal{L}(\theta, q)$ be the population ELBO. With probability $\geq 1 - \delta$:*

$$|\hat{\mathcal{L}}_N(\theta, q) - \mathcal{L}(\theta, q)| \leq O\left(\sqrt{\frac{dim(\theta)}{N} \log \frac{1}{\delta}}\right), \tag{149}$$

*where $dim(\theta) = K^2 + K \cdot c^2 + Dc$ is the total number of parameters.*

This bound shows that $N = 997$ sequences is sufficient for reliable ELBO estimation given $\dim(\theta) = 36 + 6 \times 128^2 + 4096 \times 128 \approx 622{,}000$ parameters, with a generalization gap of order $\sqrt{622000/997} \approx 25$ (in absolute ELBO units). Since our ELBOs are of order $10^5$, this represents a relative error of $< 1\%$.

## S. Spectral Analysis of SLDS Dynamics and Transitions

### S.1. Spectral Properties of the Transition Matrix

The transition matrix $\Pi \in \mathbb{R}^{K \times K}$ is a row-stochastic matrix: $\Pi_{ij} \geq 0$, $\sum_j \Pi_{ij} = 1$. Its spectral properties govern the mixing time of the induced Markov chain over cognitive phases.

**Theorem 27** (Perron-Frobenius for stochastic matrices)**.** *Let $\Pi$ be an irreducible, row-stochastic matrix. Then:*

1. *The largest eigenvalue of $\Pi$ is $\lambda_1 = 1$, with left eigenvector $\mathbf{1}$ and right eigenvector $\pi$ (stationary distribution).*
2. *All other eigenvalues satisfy $|\lambda_i| < 1$ for $i \geq 2$.*
3. *The matrix converges: $\Pi^t \to \mathbf{1}\pi^\top$ as $t \to \infty$.*

*Proof sketch.* Irreducibility implies a unique stationary distribution $\pi$. By the Perron-Frobenius theorem for non-negative matrices, the spectral radius equals the largest eigenvalue. For stochastic matrices, $\Pi\mathbf{1} = \mathbf{1}$ confirms $\lambda_1 = 1$. The sub-unit property of other eigenvalues follows from $\Pi$ being contractive in the $\ell^1$ norm: $\|\Pi v\|_1 \leq \|v\|_1$ for any $v$ orthogonal to $\pi$. □

**Definition 14** (Spectral gap)**.** $\Delta = 1 - |\lambda_2|$*, where $\lambda_2$ is the second-largest eigenvalue (by absolute value).*

**Proposition 7** (Mixing time from spectral gap)**.** *The $\varepsilon$-mixing time $\tau_\varepsilon$ (number of steps for the chain to reach $\varepsilon$-close to stationarity) satisfies:*

$$\tau_\varepsilon \leq \frac{\log(1/(\varepsilon \pi_{\min}))}{\Delta}, \tag{150}$$

*where $\pi_{\min} = \min_k \pi_k$ is the minimum stationary probability.*

For our learned transition matrices (Figure 5c), the diagonal dominance ($\Pi_{kk} \approx 0.8$–$0.95$) implies small off-diagonal entries and thus small spectral gap $\Delta \approx 0.05$–$0.2$. This means the Markov chain mixes slowly—a state tends to persist for many consecutive tokens. This is consistent with the positional organization we observe: long blocks of tokens at similar positions stay in the same SLDS state.

### S.2. Eigenvalue Analysis of Dynamics Matrices

The dynamics matrices $A_k \in \mathbb{R}^{c\times c}$ govern the local trajectory of the continuous latent state $x_t$ when in phase $k$. Their eigenvalue spectra (Figure 9) provide insight into what dynamics each state captures.

**Definition 15** (Stability of linear dynamics)**.** *The dynamics $x_t = A_k x_{t-1} + \eta_t$ are:*

- ***Stable*** *if all eigenvalues of $A_k$ have absolute value $< 1$ (the origin is an attractor).*
- ***Marginally stable*** *if all eigenvalues have $|\lambda_i| \leq 1$ with equality possible.*
- ***Unstable*** *if any eigenvalue has $|\lambda_i| > 1$.*

**Proposition 8** (Near-unit eigenvalues imply random-walk dynamics)**.** *If $A_k \approx I$ (all singular values $\approx 1$), then $x_t \approx x_{t-1} + \eta_t$ is a random walk. Random-walk dynamics do not exhibit phase-specific structure—they merely carry the current latent state forward, providing no discriminative basis for discrete state assignment.*

Our learned $A_k$ matrices all have singular values tightly clustered around 1 (Figure 9), confirming this degenerate behavior. This provides a mechanistic explanation for the SLDS failure: the dynamics matrices converge to the identity (random walk), at which point all states are dynamically equivalent and the discrete state assignments carry no information about trajectory structure.

**Theorem 28** (M-step solution for near-constant dynamics)**.** *If all tokens are assigned to a single state ($\gamma_t(k) \approx \mathbf{1}_{k=1}$), the M-step update for $A_1$ is:*

$$A_1^* = \left(\sum_t x_t x_{t-1}^\top\right)\left(\sum_t x_{t-1}x_{t-1}^\top\right)^{-1} \to \textit{proj-OLS}(\{x_t\}|\{x_{t-1}\}), \tag{151}$$

*which is the ordinary least squares (OLS) autoregressive estimator for the entire sequence. For a stationary sequence $x_t$, this converges to the true AR(1) coefficient, which for residual streams (which show near-random-walk behavior due to LayerNorm residual connections) is close to $I$.*

*Proof.* Direct from the OLS formula. The denominator $\sum_t x_{t-1}x_{t-1}^\top \to T\cdot\mathbb{E}[xx^\top]$ and the numerator $\sum_t x_t x_{t-1}^\top \to T\cdot\mathbb{E}[x_t x_{t-1}^\top]$ by the law of large numbers. The OLS estimator is $A^* = \mathbb{E}[x_t x_{t-1}^\top](\mathbb{E}[xx^\top])^{-1}$, which for an AR(1) process $x_t = Ax_{t-1} + \eta$ equals $A$ (the true coefficient). For LayerNorm residual streams, $A \approx I$. □

### S.3. Relationship Between Spectral Gap and Phase Recovery

**Theorem 29** (Phase recovery requires large spectral gap)**.** *For an SLDS with $K$ states and transition matrix $\Pi$ having spectral gap $\Delta$, the expected NMI between inferred states and any label sequence with correlation time $\tau_L$ satisfies:*

$$\mathbb{E}[\mathrm{NMI}(\hat{Z}; L)] \leq C\cdot\frac{1-\exp(-T\Delta)}{\Delta\cdot T} + \varepsilon_{\textit{signal}}, \tag{152}$$

*where $\varepsilon_{\textit{signal}}$ is the signal-determined ceiling (from Appendix E) and $C$ is a universal constant.*

*Proof sketch.* The forward-backward algorithm can extract information about $z_t$ from observations at times $s \neq t$ only to the extent that the chain mixes: the information in $\tilde{h}_s$ about $z_t$ decays as $|\lambda_2|^{|s-t|}$. The total information available is $\sum_s |\lambda_2|^{|s-t|} = (1-|\lambda_2|^{2T})/(1-|\lambda_2|^2)$, which is $O(1/\Delta)$ for small $\Delta$. When this information is small relative to $H(L)$, NMI is bounded above. The formal bound follows from information-theoretic arguments applied to the HMM forward-backward algorithm. □

For our learned $\Pi$ (spectral gap $\Delta \approx 0.1$), the bound is not tight—the failure is primarily due to the signal ceiling $\varepsilon_{\text{signal}} \approx 0.006$, not the mixing time. Both factors contribute to the observed NMI $\leq 0.005$.

# T. Emission Distribution Derivation: Full Detail

This appendix provides the complete derivation of the emission parameter M-step updates, filling in the details omitted in Appendix D.

## T.1. Setup

The emission model is:

$$h_t|x_t \sim \mathcal{N}(Cx_t + b, R), \tag{153}$$

where $h_t \in \mathbb{R}^D$, $x_t \in \mathbb{R}^c$, $C \in \mathbb{R}^{D\times c}$, $b \in \mathbb{R}^D$, $R \in \mathbb{R}^{D\times D}$ (positive definite). We want to maximize the reconstruction term of the ELBO:

$$\begin{aligned}\mathcal{L}_{\text{recon}} &= \mathbb{E}_{q(X)}\left[\sum_{t=1}^{T}\log\mathcal{N}(h_t; Cx_t+b, R)\right] \\ &= -\frac{TD}{2}\log(2\pi) - \frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}_{q(x_t)}\left[(h_t - Cx_t - b)^\top R^{-1}(h_t - Cx_t - b)\right].\end{aligned} \tag{154}$$

Let $\mu_t = \mathbb{E}[x_t]$ (smoother mean) and $\Sigma_t = \text{Var}(x_t)$ (smoother covariance). Expanding the quadratic form in (154) using Lemma 1:

$$\begin{aligned}&\mathbb{E}[(h_t - Cx_t - b)^\top R^{-1}(h_t - Cx_t - b)] \\ &= \mathbb{E}[(h_t - b - Cx_t)^\top R^{-1}(h_t - b - Cx_t)] \\ &= (h_t - b - C\mu_t)^\top R^{-1}(h_t - b - C\mu_t) + \text{tr}(R^{-1}C\Sigma_t C^\top).\end{aligned} \tag{155}$$

## T.2. M-Step for $C$

Differentiating $\mathcal{L}_{\text{recon}}$ with respect to $C$ and setting to zero. Note that we treat $b$ as fixed for the moment:

$$\frac{\partial\mathcal{L}_{\text{recon}}}{\partial C} = -\frac{1}{2}\sum_t \frac{\partial}{\partial C}\left[(h_t - b - C\mu_t)^\top R^{-1}(h_t - b - C\mu_t) + \text{tr}(R^{-1}C\Sigma_t C^\top)\right]. \tag{156}$$

*Term 1 gradient.* Let $r_t = h_t - b - C\mu_t$. Then $r_t^\top R^{-1} r_t$ is quadratic in $C$. Differentiating:

$$\frac{\partial}{\partial C} r_t^\top R^{-1} r_t = -2R^{-1}r_t\mu_t^\top = -2R^{-1}(h_t - b - C\mu_t)\mu_t^\top. \tag{157}$$

(Using the identity $\frac{\partial}{\partial A}(b - Ax)^\top M(b - Ax) = -2M(b - Ax)x^\top$ for symmetric $M$.)

*Term 2 gradient.*

$$\frac{\partial}{\partial C}\,\text{tr}(R^{-1}C\Sigma_t C^\top) = 2R^{-1}C\Sigma_t. \tag{158}$$

(Using $\frac{\partial}{\partial A}\,\text{tr}(BACA^\top) = 2BAC^\top$ corrected by symmetry: $\frac{\partial}{\partial A}\,\text{tr}(MAA^\top N) = 2MAN$ for symmetric $M$, $N$. Here $M = R^{-1}$, $N = \Sigma_t$.)

Setting the total gradient to zero:

$$\sum_t \left[R^{-1}(h_t - b - C\mu_t)\mu_t^\top - R^{-1}C\Sigma_t\right] = 0. \tag{159}$$

Pre-multiplying by $R$:

$$\sum_t (h_t - b)\mu_t^\top - C\sum_t(\Sigma_t + \mu_t\mu_t^\top) = 0. \tag{160}$$

Therefore:

$$\boxed{C^* = \left(\sum_t (h_t - b)\mu_t^\top\right)\left(\sum_t(\Sigma_t + \mu_t\mu_t^\top)\right)^{-1}.} \tag{161}$$

*Note:* We use $\mathbb{E}[x_t x_t^\top] = \Sigma_t + \mu_t\mu_t^\top$.

### T.3. M-Step for $b$

Given $C = C^*$, differentiate with respect to $b$:

$$\frac{\partial \mathcal{L}_{\text{recon}}}{\partial b} = R^{-1} \sum_t (h_t - C\mu_t - b) = 0. \tag{162}$$

Therefore:

$$\boxed{b^* = \frac{1}{T} \sum_t (h_t - C^* \mu_t).} \tag{163}$$

*Interpretation:* $b^*$ is the mean residual after accounting for the linear emission.

### T.4. M-Step for $R$

Given $(C^*, b^*)$, differentiate with respect to $R^{-1}$ (it is more convenient to differentiate with respect to the precision $\Omega = R^{-1}$):

$$\mathcal{L}_{\text{recon}} = \frac{T}{2} \log |\Omega| - \frac{1}{2} \sum_t \operatorname{tr}\big(\Omega \left[(h_t - b - C\mu_t)(h_t - b - C\mu_t)^\top + C\Sigma_t C^\top\right]\big) + \text{const.} \tag{164}$$

Setting $\partial / \partial \Omega = 0$:

$$\frac{T}{2}\Omega^{-1} - \frac{1}{2} \sum_t \left[(h_t - b - C\mu_t)(h_t - b - C\mu_t)^\top + C\Sigma_t C^\top\right] = 0, \tag{165}$$

where we used $\frac{\partial}{\partial \Omega} \log |\Omega| = \Omega^{-1} = R$. Solving:

$$\boxed{R^* = \frac{1}{T} \sum_t \left[(h_t - b^*)(h_t - b^*)^\top - C^*(h_t - b^*)\mu_t^\top\right].} \tag{166}$$

*Note:* The second term uses $C^* \mathbb{E}[x_t (h_t - b^*)^\top] = C^* \mu_t (h_t - b^*)^\top$ since $h_t$ is observed. A symmetric form can be derived but is omitted for brevity.

### T.5. Jointly Optimal Solution

**Theorem 30** (Joint maximizer of emission parameters)**.** *The updates* (161)*,* (163)*,* (166) *are the unique global maximum of* $\mathcal{L}_{recon}$ *over* $(C, b, R)$*.*

*Proof.* $\mathcal{L}_{\text{recon}}$ is jointly concave in $(C, b)$ (it is a quadratic form in $C$ and $b$ with negative semi-definite leading term, since $R^{-1} \succ 0$). Thus the first-order conditions are necessary and sufficient. The resulting $R^*$ is positive definite when the empirical second moment is non-degenerate (which holds when $T > D + c$ and the data are non-degenerate), ensuring that $R^*$ is a valid covariance matrix. □

### T.6. Practical Simplification: Whitened Observations

In our implementation, we work in the whitened PCA coordinate system where the emission simplifies. Let $\tilde{h}_t = V^\top h_t$ be the projection onto the top-$c$ PCA eigenvectors $V \in \mathbb{R}^{D \times c}$. In this coordinate system, $\tilde{h}_t \in \mathbb{R}^c$, and we set $C = I_c$ (identity), $b = 0$, $R = \sigma^2 I_c$ (isotropic noise). The M-step for $\sigma^2$ becomes:

$$\sigma^{2*} = \frac{1}{Tc} \sum_t \|\tilde{h}_t - \mu_t\|^2 + \frac{1}{c} \sum_t \operatorname{tr}(\Sigma_t). \tag{167}$$

This simplification makes the emission parameter update $O(Tc)$ instead of $O(Tc^2 + c^3)$, a significant speedup.

## U. Detailed Derivation of the Dynamics M-Step

The M-step for the dynamics matrix $A_k$ was stated in Appendix D but the algebra was condensed. This appendix provides the full derivation.

### U.1. Objective Function

The dynamics contribution to the ELBO for state $k$ is:

$$\mathcal{L}_{A_k,Q_k} = -\frac{N_k}{2}\log|Q_k| - \frac{1}{2}\sum_{t=2}^{T}\gamma_t(k)\,\mathbb{E}_{q(X)}\left[(x_t - A_k x_{t-1})^\top Q_k^{-1}(x_t - A_k x_{t-1})\right], \tag{168}$$

where $N_k = \sum_{t=2}^{T}\gamma_t(k)$.

Using the trace-expectation trick (Lemma 1):

$$\begin{aligned}
&\mathbb{E}[(x_t - A_k x_{t-1})^\top Q_k^{-1}(x_t - A_k x_{t-1})] \\
&= \mathrm{tr}\big(Q_k^{-1}\,\mathbb{E}\big[(x_t - A_k x_{t-1})(x_t - A_k x_{t-1})^\top\big]\big) \\
&= \mathrm{tr}\big(Q_k^{-1}\left[\mathbb{E}[x_t x_t^\top] - A_k\mathbb{E}[x_{t-1}x_t^\top] - \mathbb{E}[x_t x_{t-1}^\top]A_k^\top + A_k\mathbb{E}[x_{t-1}x_{t-1}^\top]A_k^\top\right]\big).
\end{aligned} \tag{169}$$

Define the sufficient statistics:

$$\Psi_k^{(1)} = \sum_{t=2}^{T}\gamma_t(k)\,\mathbb{E}[x_t x_t^\top] = \sum_{t}\gamma_t(k)(P_{t|T} + \mu_t\mu_t^\top), \tag{170}$$

$$\Psi_k^{(2)} = \sum_{t=2}^{T}\gamma_t(k)\,\mathbb{E}[x_t x_{t-1}^\top] = \sum_{t}\gamma_t(k)(P_{t,t-1|T} + \mu_t\mu_{t-1}^\top), \tag{171}$$

$$\Psi_k^{(3)} = \sum_{t=2}^{T}\gamma_t(k)\,\mathbb{E}[x_{t-1} x_{t-1}^\top] = \sum_{t}\gamma_t(k)(P_{t-1|T} + \mu_{t-1}\mu_{t-1}^\top). \tag{172}$$

In terms of these:

$$\mathcal{L}_{A_k,Q_k} = -\frac{N_k}{2}\log|Q_k| - \frac{1}{2}\,\mathrm{tr}\Big(Q_k^{-1}\left[\Psi_k^{(1)} - A_k(\Psi_k^{(2)})^\top - \Psi_k^{(2)}A_k^\top + A_k\Psi_k^{(3)}A_k^\top\right]\Big). \tag{173}$$

### U.2. Differentiating with Respect to $A_k$

Differentiating $\mathcal{L}_{A_k}$ with respect to $A_k$ (using matrix calculus):

$$\frac{\partial\mathcal{L}_{A_k}}{\partial A_k} = Q_k^{-1}\left[\Psi_k^{(2)} - A_k\Psi_k^{(3)}\right] = 0. \tag{174}$$

*Derivation of* (174)*:* The trace term in $A_k$ is:

$$F(A_k) = \mathrm{tr}(Q_k^{-1}\Psi_k^{(1)}) - 2\,\mathrm{tr}(Q_k^{-1}A_k(\Psi_k^{(2)})^\top) + \mathrm{tr}(Q_k^{-1}A_k\Psi_k^{(3)}A_k^\top). \tag{175}$$

Using the identities:

$$\frac{\partial}{\partial A}\,\mathrm{tr}(MAB^\top) = MB, \tag{176}$$

$$\frac{\partial}{\partial A}\,\mathrm{tr}(MABA^\top) = MAB^\top + MAB^\top = 2MAB^\top \text{ (for symmetric } M): \quad \frac{\partial}{\partial A}\,\mathrm{tr}(MANA^\top) = 2MAN \tag{177}$$

we get:

$$\frac{\partial F}{\partial A_k} = -2Q_k^{-1}\Psi_k^{(2)} + 2Q_k^{-1}A_k\Psi_k^{(3)} = 0. \tag{178}$$

Solving for $A_k$ (pre-multiplying by $Q_k/2$ and post-multiplying by $(\Psi_k^{(3)})^{-1}$):

$$\boxed{A_k^* = \Psi_k^{(2)} \left(\Psi_k^{(3)}\right)^{-1} = \left(\sum_t \gamma_t(k)\, \mathbb{E}[x_t x_{t-1}^\top]\right) \left(\sum_t \gamma_t(k)\, \mathbb{E}[x_{t-1} x_{t-1}^\top]\right)^{-1}.} \tag{179}$$

Note that this is the *weighted least squares* (WLS) estimator of the $k$-th dynamics: it regresses $x_t$ on $x_{t-1}$, weighted by the posterior responsibility $\gamma_t(k)$ of state $k$ at time $t$.

### U.3. Second-Order Check: Convexity

**Proposition 9.** *The objective $\mathcal{L}_{A_k}$ is concave in $A_k$ for fixed $Q_k \succ 0$.*

*Proof.* The Hessian of $\mathcal{L}_{A_k}$ with respect to $\mathrm{vec}(A_k)$ is:

$$\nabla^2_{A_k} \mathcal{L}_{A_k} = -\Psi_k^{(3)} \otimes Q_k^{-1}, \tag{180}$$

where $\otimes$ is the Kronecker product. Since $Q_k^{-1} \succ 0$ and $\Psi_k^{(3)} \succ 0$ (assuming sufficient statistics are non-degenerate), the Kronecker product is positive definite, so $\nabla^2 \mathcal{L}_{A_k}$ is negative definite, confirming strict concavity. □

Therefore, the first-order condition (179) gives the unique global maximum, and there is no ambiguity about local vs. global solutions for the M-step.

### U.4. Updating $Q_k$ After $A_k$

Given $A_k^*$, the residual for time $t$ in state $k$ is $\xi_t^{(k)} = x_t - A_k^* x_{t-1}$. The expected residual outer product is:

$$S_k = \mathbb{E}\left[\xi_t^{(k)} (\xi_t^{(k)})^\top\right] = \Psi_k^{(1)} - A_k^* (\Psi_k^{(2)})^\top - \Psi_k^{(2)} (A_k^*)^\top + A_k^* \Psi_k^{(3)} (A_k^*)^\top. \tag{181}$$

Using $A_k^* = \Psi_k^{(2)} (\Psi_k^{(3)})^{-1}$, we simplify:

$$\begin{aligned} S_k &= \Psi_k^{(1)} - A_k^* (\Psi_k^{(2)})^\top - \Psi_k^{(2)} (A_k^*)^\top + A_k^* \Psi_k^{(3)} (A_k^*)^\top \\ &= \Psi_k^{(1)} - A_k^* (\Psi_k^{(2)})^\top \quad \text{(using } A_k^* \Psi_k^{(3)} = \Psi_k^{(2)}\text{)} \qquad (182) \\ &= \Psi_k^{(1)} - \Psi_k^{(2)} (\Psi_k^{(3)})^{-1} (\Psi_k^{(2)})^\top. \qquad (183) \end{aligned}$$

This is the *Schur complement* of $\Psi_k^{(3)}$ in the $2c \times 2c$ matrix $\left(\begin{smallmatrix} \Psi_k^{(1)} & \Psi_k^{(2)} \\ (\Psi_k^{(2)})^\top & \Psi_k^{(3)} \end{smallmatrix}\right)$, confirming it is positive semi-definite.

The M-step for $Q_k$ is:

$$\boxed{Q_k^* = \frac{S_k}{N_k}.} \tag{184}$$

## V. Alternative Models and Why SLDS Is the Right Choice

This appendix justifies our choice of SLDS by comparing it to alternative approaches for cognitive phase recovery.

### V.1. Hidden Markov Model (HMM) vs. SLDS

A **Hidden Markov Model** (HMM) with Gaussian emissions in $\mathbb{R}^c$ is a special case of SLDS where:

$$x_t = \mu_{z_t} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_{z_t}), \tag{185}$$

i.e., there is no temporal dynamics ($A_k = 0$ for all $k$, or equivalently no $x_t$ structure—each observation is i.i.d. given $z_t$). The SLDS generalizes this by adding state-dependent dynamics $A_k$.

**Proposition 10** (HMM is a limiting case of SLDS)**.** *SLDS reduces to an HMM in the limit $A_k \to 0$ and $Q_k \to \Sigma_k$ (state-specific emission variance).*

**Why SLDS is preferred:** If cognitive phases correspond to distinct *dynamical regimes* (not just distinct means), then SLDS will outperform HMM by exploiting the temporal structure of the residual stream. Our experimental comparison (Figure 11) confirms that SLDS achieves higher log-likelihood than HMM, indicating it uses more information. Despite this, NMI remains $\leq 0.005$ for SLDS, while the HMM also achieves NMI $\leq 0.005$. The additional log-likelihood of SLDS is thus used to model positional dynamics, not cognitive phases.

### V.2. K-Means Clustering vs. SLDS

**K-Means** clusters tokens by Euclidean distance in PCA space, ignoring the temporal ordering. It assumes that cognitive phases correspond to distinct *clusters* in activation space. When SLDS is initialized with K-means cluster assignments (random restart variant), it converges to NMI $\leq 0.005$, matching standard random initialization.

**Why K-means fails:** The phase signal is weak ($d \leq 0.3$), meaning the between-phase distances are small relative to within-phase variance. K-means optimizes within-cluster squared distance, which is dominated by the positional/syntactic variance in PCA space. The resulting clusters are positional, not cognitive.

### V.3. Recurrent Switching LDS (rSLDS)

Linderman et al. (2016) proposed **recurrent SLDS**, where the transition probabilities $P(z_t|z_{t-1}, x_{t-1})$ depend on the continuous state $x_{t-1}$:

$$P(z_t = k|z_{t-1}, x_{t-1}) = \text{softmax}(R_k^\top x_{t-1} + r_k). \tag{186}$$

This allows the model to learn input-dependent switching, which might better capture phase transitions triggered by specific activation patterns.

We did not evaluate rSLDS due to: (1) higher computational cost ($O(NTKc^3)$ per EM iteration vs. $O(NTc^3)$ for standard SLDS); (2) the primary reason for SLDS failure (weak signal, $d \leq 0.3$) would persist in rSLDS—recurrent transitions cannot recover phases that are not linearly separable in the first place; (3) rSLDS requires differentiable optimization (no closed-form M-step), increasing optimization difficulty.

We expect rSLDS would achieve NMI comparable to standard SLDS ($\leq 0.005$) for the same underlying reason: the signal strength is insufficient regardless of the switching mechanism.

### V.4. Neural ODE and Flow-Based Alternatives

Chen et al. (2018) and subsequent work suggest modeling activation trajectories as continuous flows:

$$\frac{dx}{dt} = f_\phi(x, t), \tag{187}$$

where $f_\phi$ is a neural network. This allows non-linear, continuously varying dynamics. A Neural ODE trajectory model could, in principle, learn to represent phase boundaries as "velocity changes" in the latent trajectory rather than discrete switches.

**Connection to our results:** Our linear probe achieves AUC $= 0.629$ for boundary/non-boundary classification, suggesting that there *is* a learnable signal in the activations. A Neural ODE could potentially extract this signal more effectively than a linear probe or SLDS. Future work should apply continuous trajectory methods to the residual-stream trajectories we have collected.

### V.5. Transformer-Based Sequence Models

An alternative approach is to use a transformer encoder to predict $z_t$ from the context $h_{1:T}$:

$$\hat{z}_t = \text{Transformer}(h_{1:T})_t. \tag{188}$$

This is a fully supervised approach (requires labeled $z_t$) and does not provide a generative model of the activations. Our setting is semi-supervised (keyword annotations give noisy $z_t$ labels for $\sim 2\%$ of tokens), making this approach applicable

but limited by label coverage.

**Upper bound from our results:** Even if a perfect transformer-based predictor were trained on all labeled tokens, the AUC ceiling is 0.629 (linear probe result), and the NMI ceiling is $\leq 0.092$ (from linear probe bound, Appendix I). No method can exceed these bounds given the signal available in layer-16 activations.

### V.6. Why the Null Result is Informative

The comprehensive evaluation across five alternative approaches (SLDS, HMM, K-means, rSLDS discussion, linear probe) converges on the same conclusion: the residual stream does not contain discrete, recoverable cognitive phase structure. This convergence across very different model classes makes the null result much stronger than any single negative result. It is not a failure of one algorithm but a fundamental property of the representation.

## W. Complete Proofs of Statistical Theorems

### W.1. Jensen's Inequality: Complete Proof

**Theorem 31** (Jensen's inequality)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function and $X$ a random variable with $\mathbb{E}[|X|] < \infty$. Then:*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]. \tag{189}$$

*Equivalently, for concave $g = -f$: $g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$, which is the form used in the ELBO derivation with $g = \log(\cdot)$.*

*Proof. Step 1: Characterization of convexity.* $f$ is convex iff for every $x_0$, there exists a supporting hyperplane: a slope $s$ such that:

$$f(x) \geq f(x_0) + s(x - x_0) \quad \text{for all } x. \tag{190}$$

(For differentiable $f$, $s = f'(x_0)$.)

*Step 2: Apply at $x_0 = \mathbb{E}[X]$.* Let $\mu = \mathbb{E}[X]$. By (190) with $x_0 = \mu$:

$$f(X) \geq f(\mu) + s(X - \mu) \quad \text{almost surely.} \tag{191}$$

Taking expectations of both sides:

$$\mathbb{E}[f(X)] \geq f(\mu) + s\underbrace{(\mathbb{E}[X] - \mu)}_{=0} = f(\mathbb{E}[X]). \tag{192}$$

□

**Corollary 4** (Jensen's inequality for the ELBO)**.** *For any probability distribution $q(X, Z)$ and model $P(H, X, Z|\theta)$:*

$$\log P(H|\theta) = \log \mathbb{E}_q\left[\frac{P(H, X, Z|\theta)}{q(X, Z)}\right] \geq \mathbb{E}_q\left[\log \frac{P(H, X, Z|\theta)}{q(X, Z)}\right] = \mathcal{L}(\theta, q). \tag{193}$$

*Proof.* Apply Jensen with the concave function $g = \log(\cdot)$ and $f = P(H, X, Z|\theta)/q(X, Z)$, which is a well-defined positive-valued random variable under the distribution $q$. □

### W.2. Fano's Inequality: Complete Proof

**Theorem 32** (Fano's inequality)**.** *Let $X$ be a discrete random variable taking values in $\{1, \ldots, M\}$ and $\hat{X} = g(Y)$ any estimate of $X$ based on $Y$. Let $P_e = P(\hat{X} \neq X)$. Then:*

$$H(X|Y) \leq H_b(P_e) + P_e \log(M - 1), \tag{194}$$

*where $H_b(p) = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy.*

*Proof.* Define the error indicator $E = \mathbf{1}[\hat{X} \neq X]$. Then:

$$\begin{aligned} H(X|Y) &= H(X|\hat{X}) \leq H(X|\hat{X}) + H(E|\hat{X}, X) \\ &= H(X, E|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}). \end{aligned} \tag{195}$$

Now bound each term:

- $H(E|\hat{X}) \leq H(E) = H_b(P_e)$ (conditioning reduces entropy; $H(E) = H_b(P_e)$ since $E$ is Bernoulli$(P_e)$).
- $H(X|E = 0, \hat{X}) = 0$ (if no error, $X = \hat{X}$ exactly).
- $H(X|E = 1, \hat{X}) \leq \log(M - 1)$ (if error, $X \neq \hat{X}$, so $X$ takes at most $M - 1$ values given $\hat{X}$).

Therefore:

$$H(X|E, \hat{X}) = P(E = 0) \cdot 0 + P(E = 1) \cdot H(X|E = 1, \hat{X}) \leq P_e \log(M - 1). \tag{196}$$

Combining gives (194). □

**Application to our setting.** With $M = 4$ phases and best-case accuracy $62\%$ ($P_e = 0.38$):

$$H(L|Y) \leq H_b(0.38) + 0.38 \log 3 = 0.954 + 0.417 = 1.371 \text{ bits}, \tag{197}$$

so $I(\hat{Z}; L) = H(L) - H(L|\hat{Z}) \geq \log 4 - 1.371 = 2 - 1.371 = 0.629$ bits minimum information. But since $H(L) \approx \log 4 = 2$ bits, the maximum NMI achievable from a 62%-accurate classifier is: $\text{NMI} \leq 2I/(H(L) + H(\hat{Z})) \leq 0.629/1.5 = 0.419$ — but this is an *upper* bound only via the inequality in Fano's; the actual NMI depends on the specific classifier. Our observed NMI $= 0.005$ is far below this theoretical upper bound.

### W.3. Data Processing Inequality: Complete Proof

**Theorem 33** (Data processing inequality)**.** *If $X \to Y \to Z$ is a Markov chain (i.e., $X$ and $Z$ are conditionally independent given $Y$), then:*

$$I(X; Z) \leq I(X; Y). \tag{198}$$

*Proof.* By the chain rule of mutual information:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y). \tag{199}$$

The Markov condition $X \to Y \to Z$ means $P(X, Y, Z) = P(Z|Y)P(Y|X)P(X)$, which implies $X \perp Z|Y$, i.e., $I(X; Z|Y) = 0$. Therefore:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y). \tag{200}$$

Since $I(X; Y|Z) \geq 0$, we get $I(X; Z) = I(X; Y, Z) - I(X; Y|Z) \leq I(X; Y, Z) = I(X; Y)$. □

**Application.** Any function $f(H)$ of the activations $H$ satisfies $I(f(H); L) \leq I(H; L)$. In particular, the SLDS inferred states $\hat{Z} = f_{\text{SLDS}}(H)$ satisfy $I(\hat{Z}; L) \leq I(H; L)$. Our linear probe results give an empirical upper bound on $I(H; L)$, which implies an upper bound on NMI for any downstream classifier of $H$.

### W.4. Proof that NMI $\leq 0.005$ Implies Near-Chance Performance

**Theorem 34.** *If $\text{NMI}(\hat{Z}; L) = 0.005$ with $H(L) \approx \log 4 = 2$ bits and $H(\hat{Z}) \leq \log 6 = 2.585$ bits (for $K = 6$ states), then the expected accuracy of using $\hat{Z}$ to predict $L$ is at most $\approx 31\%$ — close to the random-assignment baseline of $25\%$.*

*Proof.* $I(\hat{Z}; L) = \text{NMI} \cdot (H(\hat{Z}) + H(L))/2 = 0.005 \times (2.585 + 2)/2 = 0.005 \times 2.293 = 0.01146$ bits.

The accuracy of the Bayes-optimal classifier from $\hat{Z}$ to $L$ satisfies Fano's lower bound: $H(L|\hat{Z}) \geq H(L) - I(\hat{Z}; L) = 2 - 0.0115 = 1.989$ bits.

By the converse of Fano's: $H_b(P_e) + P_e \log(M-1) \geq H(L|\hat{Z}) - H(L) + \log M = 1.989 - 2 + 2 = 1.989$ bits.

Since $H_b(P_e) \leq 1$ and $\log(M-1) = \log 3 = 1.585$: $P_e \geq (1.989 - 1)/1.585 = 0.624$, i.e., accuracy $\leq 37.6\%$.

This is only marginally above the random-assignment baseline of $25\%$ for 4 classes, confirming that NMI $= 0.005$ corresponds to near-chance prediction of cognitive phases. □

### W.5. Bayes' Theorem and Posterior Computation

For completeness, we state the formal version of Bayes' theorem used throughout:

**Theorem 35** (Bayes' theorem, continuous version)**.** *Let $(X, Y)$ be a pair of random variables with joint density $p(x, y)$, marginal densities $p(x)$ and $p(y)$, and conditional densities $p(y|x) = p(x, y)/p(x)$ and $p(x|y) = p(x, y)/p(y)$. Then:*

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x')p(x')\,dx'}. \tag{201}$$

All posterior computations in this paper (Kalman update, forward-backward, variational updates) are applications of Bayes' theorem. The challenge in SLDS is that the denominator (normalization constant) is intractable, which motivates the variational approximation.

### W.6. Law of Total Expectation Applied to the ELBO

**Lemma 3** (Iterated expectations in the ELBO)**.** *For the mean-field factorization $q(X, Z) = q(X)q(Z)$:*

$$\mathbb{E}_{q(X,Z)}[f(x_t, z_t)] = \sum_k \gamma_t(k)\, \mathbb{E}_{q(X)}[f(x_t, k)], \tag{202}$$

*where $\gamma_t(k) = q(z_t = k)$.*

*Proof.* By the product structure:

$$\begin{aligned}\mathbb{E}_{q(X,Z)}[f(x_t, z_t)] &= \int \sum_k f(x_t, k)\, q(z_t = k)\, q(x_t)\, dx_t \\ &= \sum_k q(z_t = k) \int f(x_t, k)\, q(x_t)\, dx_t = \sum_k \gamma_t(k)\, \mathbb{E}_{q(X)}[f(x_t, k)].\end{aligned} \tag{203}$$

□

This lemma justifies the factored E-step: we can first run the Kalman smoother to compute $q(X)$ (treating $z_t$ as given by its mean $\gamma_t$), and then run forward-backward to compute $q(Z)$ (using expectations under $q(X)$), without ever computing the joint posterior $q(X, Z)$.

### W.7. Proofs of Statistical Power and Effect Size Relations

**Theorem 36** (Student's $t$-test power)**.** *For a two-sample $t$-test with equal sample sizes $n$ and equal variances, comparing means $\mu_1, \mu_2$ with common standard deviation $\sigma$, at significance level $\alpha$:*

$$\textit{Power}(\delta, n, \alpha) = \Phi\left(z_{1-\alpha/2} - \delta\sqrt{\frac{n}{2}}\right) + 1 - \Phi\left(z_{1-\alpha/2} + \delta\sqrt{\frac{n}{2}}\right), \tag{204}$$

*where $\delta = |\mu_1 - \mu_2|/\sigma = |d|$ is Cohen's $d$ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal.*

*Proof.* Under $H_0$: $\mu_1 = \mu_2$, the test statistic $T = (\bar{X}_1 - \bar{X}_2)/\hat{s}\sqrt{2/n}$ follows a $t$-distribution with $2n-2$ d.f., approximately $\mathcal{N}(0, 1)$ for large $n$.

Under $H_1$: $\mu_1 \neq \mu_2$, $T$ follows a non-central $t$-distribution with non-centrality parameter $\lambda = \delta\sqrt{n/2}$.

The power is $P(|T| > z_{1-\alpha/2}|H_1)$. For large $n$ (our $n \sim 6000$), the non-central $t$ is well approximated by $\mathcal{N}(\lambda, 1)$:

$$\begin{aligned} \text{Power} &= P(|T| > z_{1-\alpha/2}) = P(T > z_{1-\alpha/2}) + P(T < -z_{1-\alpha/2}) && (205) \\ &= \Phi(\lambda - z_{1-\alpha/2}) + \Phi(-\lambda - z_{1-\alpha/2}) \\ &\approx \Phi(\delta\sqrt{n/2} - z_{1-\alpha/2}) + \Phi(-\delta\sqrt{n/2} - z_{1-\alpha/2}). && (206) \end{aligned}$$

Rearranging gives the stated formula. □

For our boundary diagnostic: $d = 0.293$, $n = 6000$ (boundary tokens), $\alpha = 3.9 \times 10^{-4}$ (Bonferroni). Then $\lambda = 0.293\sqrt{3000} \approx 16.0$ and $z_{1-\alpha/2} \approx 3.6$, giving power $\approx \Phi(16.0 - 3.6) = \Phi(12.4) \approx 1.0$. Our study is therefore near-perfectly powered to detect $d = 0.293$ given $n = 6000$ observations, confirming that the *absence* of larger effects is genuine.

### W.8. Expected Value of $\chi^2$ Statistic Under Positional Structure

**Theorem 37** (Expected $\chi^2$ for position-correlated SLDS states)**.** *If SLDS states are assigned purely by token position (SLDS state $k$ for tokens in position decile $k$, $K = 10$), and position deciles are balanced, the expected chi-squared statistic for a $K \times P$ contingency table (states $\times$ position decile) is:*

$$\mathbb{E}[\chi^2] \approx N\,(K-1) + KP - K = N \cdot (K-1), \tag{207}$$

*for large $N$, where we have $K = 6$ states and $P = 4$ position quartiles.*

*Proof sketch.* Under perfect alignment (state $k$ assigns all tokens from position quartile $k$), the observed counts are $N_{kp} = N/K$ for $k = p$ and 0 otherwise. The expected counts under independence are $E_{kp} = N/(KP)$. The chi-squared statistic is:

$$\chi^2 = \sum_{k \neq p} \frac{N_{kp}^2}{E_{kp}} + \sum_k \frac{(N_{kk} - E_{kk})^2}{E_{kk}} \approx N\frac{K(P-1)}{P} = N\frac{K(P-1)}{P}. \tag{208}$$

For $N \approx 53775$, $K = 6$, $P = 4$: $\chi^2 \approx 53775 \times 6 \times 3/4 \approx 242{,}000 \gg 2343$. Our observed $\chi^2 = 2343$ corresponds to a partial alignment (not perfect positional recovery), consistent with SLDS states being *correlated* with but not perfectly predicting position. □