Bifurcations in Mechanism Space: Change-Point Detection on Interventional Influence Subspaces Predicts Capability Emergence

Liz Lemma Future Detective

January 18, 2026

Abstract

Mechanistic interpretability work emphasizes that (i) causal interventions are necessary for faithful explanations and (ii) capabilities can emerge stagewise, sometimes with abrupt transitions. Yet training-time monitoring still relies mostly on loss curves and behavioral evaluations, which can be delayed, noisy, or actively masked by finetuning. We propose a formal training-time forecasting primitive: track low-dimensional subspaces of causal influence computed from on-manifold interventions, and detect bifurcations (abrupt rotations or rank changes) that precede emergent capabilities. We define an interventional influence operator J_t at each checkpoint t, estimate its top-k singular subspace using randomized interventions, and perform change-point detection on principal-angle distances. In a simplified low-rank mechanism model, we prove (1) sample complexity bounds for subspace recovery and (2) detection-delay guarantees under rank-k perturbations, with matching information-theoretic lower bounds. We further outline an experimental protocol to instrument real training runs and backtest whether influence bifurcations anticipate capability jumps (including cases where finetuning masks behavior while influence remains stable). The framework directly addresses open problems flagged in recent mechanistic interpretability surveys: intervention-dependence, validation beyond cherry-picked tasks, and understanding mechanism development during training.

Table of Contents

3. 3. Problem formulation (MBD): define checkpoints, taps, on-manifold interventions, influence operators, subspace metrics, and prediction targets; relation to black-box emergence detection.

4. 4. Influence operator estimation: randomized intervention schemes, low-rank sketches, principal subspace recovery; practical considerations (choice of taps, multi-tap aggregation).

5. 5. Theory I — subspace estimation bounds: finite-sample guarantees, dependence on spectral gap/noise; matching lower bounds from hypothesis testing.

6. 6. Theory II — change-point detection guarantees: detection delay under low-rank perturbations; false alarm control; matching lower bounds (e.g., rank-one spike detection).

7. 7. Capability forecasting from bifurcations: mapping detected regime changes to predicted capability onset times; calibrated uncertainty intervals.

8. 8. Experimental protocol (strengthening evidence): checkpointed training suites, synthetic-to-real curricula, capability evals, and comparisons to baselines (loss curvature, gradient norms, probe emergence).

9. 9. Case studies: finetuning masks/unmasks behavior; influence subspaces as invariant signature; implications for monitoring strategic or safety-trained models.

10. 10. Limitations and extensions: on-manifold requirement; multi-step/agentic behaviors; scaling costs; multi-architecture transfer (Transformers/SSMs).

11. 11. Related work and positioning: stagewise learning, mechanistic circuits, causal abstraction, influence functions; what this work adds.

12. 12. Conclusion: influence bifurcations as a training-time early warning signal; roadmap to deployment in safety cases.

1 Introduction

Training-time capability emergence is increasingly treated not only as an empirical phenomenon but as an object of governance: one is asked to justify, in a *safety case*, that a development process will surface and manage the appearance of new behaviors before deployment. In this setting, a "capability" is not merely a benchmark score but a latent competence that can become behaviorally expressed after a small number of additional optimization steps, architectural changes, or finetuning on a narrow distribution. The practical question is therefore a change-of-mechanism question: at what point during training does the model acquire internal machinery that would support a qualitatively new behavior, and can we detect that point with bounded false alarms under a realistic monitoring budget?

A natural baseline is to monitor scalar training signals (loss, gradient norms, weight updates) or to run black-box behavioral evaluations on a fixed suite of tasks. We argue that such signals are inadequate for early warning in the precise sense relevant to a safety case. Loss curves are not keyed to specific competencies, can be dominated by frequent-token or easy-example mass, and may remain smooth across internal reorganizations that matter for downstream safety. Behavioral evaluations have a more direct semantics, but they are limited by distributional masking: a capability may exist while being unexpressed on the evaluation distribution, either because triggering contexts are rare, because the evaluation protocol does not elicit the behavior, or because subsequent training actively suppresses overt performance (e.g. via instruction tuning). Moreover, for emergent behaviors that depend on rare prompts or latent states, the sample complexity of detecting a change in output distribution can be prohibitive at the point where intervention-based signals are already decisive. For governance purposes, these limitations are not incidental; they constitute an identifiability obstruction for any monitor that observes only input–output behavior on a fixed set of tasks.

We therefore pursue a monitoring primitive that is (i) keyed to internal causal mechanisms rather than external behavior alone, (ii) computable with bounded per-checkpoint cost, and (iii) admits quantitative guarantees that can be reported as part of a confidence calibration artifact. Our proposal is to treat a sequence of checkpoints as a streaming change-point problem over *interventional influence operators*. Concretely, we fix an internal "tap" $h_t(x) \in \mathbb{R}^d$ and a downstream readout $y_t(x) \in \mathbb{R}^p$ whose variation we regard as mechanistically meaningful (e.g. a probe logit, an action head, or a safety-relevant internal score). We then consider randomized on-manifold interventions of the form

$$\mathrm{do}(h \leftarrow h + \alpha u), \qquad u \sim \mathcal{N}(0, I_d),$$

and measure the induced change Δy relative to a baseline forward pass on contexts $x \sim \mathcal{D}$. Under a local linearity assumption, the conditional mean

effect is approximately linear in u, yielding an operator $J_t \in \mathbb{R}^{p\times d}$ defined by $\mathbb{E}[\Delta y \mid u] = \alpha J_t u$. The operator J_t is not merely a gradient object; it is a *causal* sensitivity of the readout to perturbations in the tapped state under a prescribed intervention semantics, and thus it is meaningful even when gradients are inaccessible or unreliable.

Because d can be large, we do not attempt to estimate J_t fully. Instead, we monitor the top-k right singular subspace $\mathcal{S}_t \subseteq \mathbb{R}^d$ of J_t, which we interpret as a k-dimensional "mechanism subspace" of the tapped representation that exerts the dominant causal influence on y_t under the chosen intervention family. This choice is motivated by two considerations. First, it is invariant to rotations within the influential directions and thus robust to representational drift that preserves the underlying mechanism. Second, it admits stable estimation whenever there is a nontrivial spectral gap $\mathrm{gap}_t = \sigma_k(J_t) - \sigma_{k+1}(J_t)$, which is precisely the regime in which a low-dimensional mechanism is identifiable from randomized perturbations.

Our basic detector estimates $\widehat{\mathcal{S}}_t$ at each checkpoint t using m randomized interventions and computes the principal-angle distance

$$\Delta_t \;=\; \sin\Theta(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t-1}).$$

A bifurcation is declared when $\Delta_t \geq \gamma$ for a threshold γ chosen to control false alarms. This produces a sequence of flagged times $\widehat{\mathcal{B}} \subseteq \{2, \ldots, T\}$ at which the model's influential mechanism geometry changes. In addition, when one tracks multiple readouts $y_t^{(j)}$ corresponding to capability tasks $\mathcal{T}_j$, the same machinery yields a forecast problem: we can treat the evolution of $\widehat{\mathcal{S}}_t^{(j)}$ (or derived quantities such as $\sigma_i(\widehat{J}_t^{(j)})$) as a leading indicator for when the capability will become behaviorally detectable under the task's evaluation protocol.

The primary contribution is that this picture supports explicit finite-sample guarantees that scale with the intervention budget and noise level in the expected way. Under subgaussian response noise with proxy σ^2, local intervention magnitude α, and a spectral gap condition, we obtain high-probability bounds of the form

$$\sin\Theta(\widehat{\mathcal{S}}_t, \mathcal{S}_t) \;\lesssim\; \frac{\sigma}{\alpha}\sqrt{\frac{\log(1/\delta)}{m}} \cdot \frac{1}{\mathrm{gap}_t},$$

which exhibit the correct signal-to-noise dependence $\alpha\sqrt{m}/\sigma$ and the expected $1/\mathrm{gap}_t$ stability factor. We complement this with matching (up to constants and logarithms) lower bounds showing that no method using only m randomized interventions can uniformly improve the $\sqrt{1/m}$ scaling in the worst case. These results justify the top-k influence subspace as a statistically natural monitoring target: it is rich enough to capture meaningful mechanism changes, yet low-dimensional enough to be estimable under strict per-checkpoint budgets.

A second contribution concerns detection. When a new mechanism appears at time t^* as a perturbation $J_{t^*} = J_{t^*-1} + \Delta$ with $\|\Delta\|_2 \geq \Delta_0$ and the post-change operator retains a gap, we show that a simple threshold rule can detect the change with small (and in the idealized one-change model, zero) delay once m is large enough that estimation error is below a constant fraction of the induced subspace rotation. Conversely, in a rank-one spike model, we show that if m is below a threshold proportional to $\sigma^2/(\alpha^2\Delta_0^2)\log(1/\delta)$, then any detector must either incur false alarm probability exceeding δ or miss probability exceeding δ. In this sense, the monitoring problem admits a sharp sample-complexity characterization analogous to classical change-point testing, but with the novelty that the "signal" is a geometric change in an interventional operator rather than a shift in mean loss or output logits.

A third contribution is conceptual but formalizable: we record an impossibility principle for behavior-only early warning. If the output distributions on an evaluation task are identical up to time t^*, then no black-box procedure observing only those outputs can reliably predict the emergence time before t^*. This is not a pessimistic claim about current benchmarks; it is a structural statement about identifiability. Any early-warning claim must therefore rest on additional observables (internal taps, auxiliary distributions $\mathcal{D}$, or trusted interventions), which our framework makes explicit.

Finally, we emphasize implementability constraints. We require interventions to be local (small α) and on-manifold according to a specified patching or conditional-resampling semantics, so that the operator J_t corresponds to plausible counterfactual variation rather than arbitrary activation corruption. Computationally, the estimator can be implemented with streaming sketches and randomized SVD, using m additional forward passes per checkpoint and without requiring gradient access. This aligns with the regime in which monitoring must coexist with large-scale training.

In the next section we place this construction in the context of mechanistic interpretability and causal patching, and we motivate why influence subspaces—as opposed to individual neurons, raw Jacobians, or purely behavioral probes—form a stable and informative unit for tracking stagewise development and emergent capabilities.

2 Background and motivation

Mechanistic interpretability studies have made a compelling case that many behaviors of large sequence models are implemented by structured computations distributed across layers, attention heads, and residual streams rather than by isolated units. From the monitoring perspective, however, the central difficulty is not merely to *explain* a fixed trained model, but to *track* how computations reorganize over training time. The objects of interest must therefore be stable under common symmetries (e.g. rotations and basis

changes in internal representations), must admit estimation at scale, and must connect to concrete downstream quantities that are relevant for safety cases (probe scores, action logits, or internal "risk" heads). This perspective shifts emphasis away from locating a single "feature neuron" and toward identifying low-dimensional *mechanism degrees of freedom* whose presence and influence can be measured repeatedly across checkpoints.

A second theme is that interpretability arguments often rest on correlational evidence: a direction correlates with a concept; an activation pattern correlates with a behavior; or a linear probe decodes a label. Such evidence is useful but does not, by itself, establish that the model *uses* the decoded information. Causal intervention methods—activation patching, causal tracing, path patching, and related "scrubbing" procedures—were introduced precisely to answer use-questions: if we modify an internal state in a controlled manner, do we reliably change a downstream readout? From our vantage point, these methods suggest a monitoring primitive: rather than inspecting raw activations, we measure an *interventional influence* of a tapped state on a chosen readout. The monitor is then keyed to causal effect sizes, which are closer to what a safety case requires than descriptive statistics of internal representations.

At the same time, naive interventions can be misleading. Arbitrary corruption of a residual stream coordinate can easily push the model off the data manifold, producing brittle artifacts that neither reflect plausible internal counterfactuals nor yield stable metrics across checkpoints. This motivates the constraint that interventions be *on-manifold* according to an explicit semantics. Concretely, we may implement interventions via constrained patching from matched contexts, conditional resampling of subcomponents, or other procedures that preserve token-level plausibility and the local statistics of the representation. The precise semantics can vary with the application, but the monitoring requirement is the same: the induced counterfactual should be interpretable as "a plausible alternative internal state consistent with the current context distribution." Under such semantics, causal effects become meaningfully comparable over training and less sensitive to superficial representational drift.

The monitoring problem is also shaped by the empirical fact that capability acquisition is often *stagewise*. Across long training runs, qualitative changes can occur: a model may begin to perform multi-step reasoning, execute tool-like behaviors, or exhibit situationally sensitive policy selection. Even when behavioral curves appear smooth, mechanistic analyses frequently reveal that the internal pathways supporting performance can reorganize abruptly (e.g. via the emergence of a new attention circuit that bypasses a previous heuristic). This stagewise picture is consistent with the view that optimization discovers new algorithmic "modules" which, once formed, can later be gated, fine-tuned, or suppressed without being erased. For governance, this distinction matters: a capability that is not currently

expressed can nonetheless become expressed after small distribution shifts or post-training procedures, and therefore monitoring should detect the acquisition of the underlying machinery rather than only its current behavioral expression.

A natural response is to run extensive behavioral evaluations throughout training. Yet behavioral monitoring faces an identifiability obstruction whenever the evaluation distribution fails to elicit the latent capability, or when training actively shapes outputs to satisfy the evaluation protocol without removing the underlying mechanism. In such regimes, the causal influence of internal states can change materially before output distributions on a fixed benchmark register a difference. We thus seek an internal early-warning signal that (i) is closer to "mechanism present and causally active" than to "behavior currently visible," and (ii) can be estimated reliably under strict compute budgets.

This brings us to the question of *what* internal quantity to track. One might consider tracking the full Jacobian of the readout with respect to the tap, or tracking individual directions discovered by probes. Both extremes are problematic. Full Jacobians are high-dimensional, expensive to estimate, and unstable under small perturbations when their spectra are diffuse. Individual directions, on the other hand, are fragile: they depend on arbitrary choices of basis and can change substantially under innocuous reparameterizations, even when the underlying computation is unchanged. Moreover, for distributed mechanisms, no single direction need remain consistently aligned across checkpoints; what is stable is often a *subspace* of directions that collectively support the computation.

The subspace viewpoint is standard in perturbation theory and is particularly natural for causal influence. If a downstream readout depends on the tap primarily through a low-dimensional set of degrees of freedom, then the operator mapping small on-manifold perturbations at the tap to changes in the readout is approximately low rank. In that case, the dominant right singular vectors define a mechanism-aligned coordinate system: directions in the tap space that, when perturbed, produce maximal expected effect on the readout. Importantly, the *span* of these directions is invariant to rotations within the mechanism and therefore stable under representational drift that preserves the computation. Stability is formalized by spectral-gap arguments: when there is a gap between the top singular values and the remainder, the associated subspace is robust to noise and to small model changes that do not materially alter the mechanism.

From a monitoring standpoint, "influence subspaces" therefore serve as a compromise object: richer than a scalar score, yet lower-dimensional and more stable than a full operator. They also admit a direct causal semantics. When we intervene at the tap by adding a small on-manifold direction and observe the induced change in a readout, we are effectively sampling the action of an unknown linear operator in random directions, in the local

regime where linearization is accurate. Estimating the *top* of this operator (as a subspace) is then the statistically efficient target: it is identifiable with a number of interventions that scales with the desired error and the signal-to-noise ratio, and it is exactly the component that captures "which internal degrees of freedom matter most" for the monitored readout.

Finally, influence subspaces are well suited to the stagewise-emergence narrative. When a new circuit begins to support a capability, we expect a new family of perturbations at the relevant tap to acquire causal leverage over a capability-linked readout. This can manifest as an increase in the magnitude of leading singular values (a strengthening of an existing mechanism), as a rotation of the dominant subspace (a change in how the model implements the behavior), or as a rank expansion (the appearance of additional influential degrees of freedom). Each of these is naturally captured by tracking subspace geometry and spectrum across checkpoints. In contrast, raw activation statistics can change for reasons unrelated to capability (e.g. rescaling), and behavior can remain unchanged for reasons unrelated to mechanism (e.g. masking by instruction tuning). The influence-subspace monitor sits between these extremes: it is grounded in interventions, hence causal; it is geometric, hence robust; and it is low-dimensional, hence feasible to estimate repeatedly during training.

These considerations motivate the formalization in the next section: we define checkpoints, taps, on-manifold interventions, and an interventional influence operator whose leading right singular subspace is the monitoring target. We then cast training-time emergence prediction as change-point detection in this subspace, with explicit thresholds and error control suitable for reporting in a confidence calibration artifact.

3 Problem formulation: mechanism bifurcation detection (MBD)

We model training as producing a discrete sequence of checkpoints $\{M_t\}_{t=1}^{T}$, where M_t denotes the model parameters θ_t at checkpoint index t. Fix a *tap*—an internal state of the forward computation—given as a measurable map $h_t : \mathcal{X} \to \mathbb{R}^d$, $x \mapsto h_t(x)$, where $x \sim \mathcal{D}$ is a context drawn from a monitoring distribution $\mathcal{D}$ (train/validation/mix, and potentially including safety-relevant subdistributions). We also fix a *readout of interest* $y_t : \mathcal{X} \to \mathbb{R}^p$, which may be a probe logit vector, a score produced by an internal head, an action logit in an agentic model, or any differentiable downstream statistic whose changes we wish to causally attribute to perturbations at the tap.

Our monitoring primitive is an *on-manifold intervention family* applied at the tap. For each checkpoint t and context x, we consider counterfactual

runs in which the tap state is modified according to

$$\mathrm{do}\big(h \leftarrow h + \alpha u\big),$$

where $u \in \mathbb{R}^d$ is a randomized direction (typically $u \sim \mathcal{N}(0, I_d)$ or Rademacher), and $\alpha > 0$ is a small magnitude parameter. The phrase "on-manifold" is operational rather than metaphysical: we assume the intervention is implemented by a specified semantics (e.g. constrained activation patching, conditional resampling, or a projection step) that yields perturbed states which remain typical under $\mathcal{D}$ in the sense that they preserve local representation statistics and avoid gross distribution shift at the tap. Formally, we treat $\mathrm{do}(\cdot)$ as part of the experimental design: it determines the conditional distribution of the perturbed forward computation given (x, u), and hence determines the causal estimand below.

Given (t, x, u), let $y_t(x)$ denote the readout in the baseline run and $y_t^{\mathrm{do}}(x; u)$ denote the readout under intervention. We define the intervention response

$$\Delta y_t(x; u) \;:=\; y_t^{\mathrm{do}}(x; u) - y_t(x) \in \mathbb{R}^p.$$

We work in the local regime in which $\Delta y_t(x; u)$ is well approximated by a linear function of u, after averaging over contexts. Concretely, our standing modeling assumption (H1) is that there exists an operator $J_t \in \mathbb{R}^{p \times d}$, the *interventional influence operator* at checkpoint t, such that

$$\mathbb{E}\big[\Delta y_t(x; u) \mid u\big] \;=\; \alpha J_t u, \tag{1}$$

where the expectation integrates over $x \sim \mathcal{D}$ and the internal randomness of the on-manifold intervention semantics, and deviations from (1) are captured by mean-zero subgaussian noise η with variance proxy σ^2. Equation (1) is not intended to hold pointwise in x, but rather as a stable $\mathcal{D}$-averaged causal summary of how perturbations at the tap flow to the readout in the immediate neighborhood of typical internal states. When y_t is differentiable with respect to h_t and interventions are implemented as additive perturbations, J_t coincides with an average Jacobian; however, we emphasize that we do not require gradient access, and we view J_t as an estimand defined by randomized interventions.

The raw operator J_t is typically too large to estimate or track entrywise when d is the width of a residual stream. Instead, we track its dominant *mechanism subspace*. Let $\sigma_1(J_t) \geq \sigma_2(J_t) \geq \cdots$ denote the singular values of J_t. Fix a target dimension k, representing the number of influential degrees of freedom we are prepared to monitor. We define $\mathcal{S}_t \subseteq \mathbb{R}^d$ to be the top-k *right* singular subspace of J_t; equivalently, $\mathcal{S}_t$ is spanned by the k right singular vectors corresponding to $\sigma_1(J_t), \ldots, \sigma_k(J_t)$. This definition enforces invariance to rotations within the mechanism: if the model reparameterizes the tapped representation by an orthogonal change of basis that preserves

the input–output mapping, $\mathcal{S}_t$ remains the appropriate geometric object. We assume (H2) a nontrivial spectral gap $\mathrm{gap}_t := \sigma_k(J_t) - \sigma_{k+1}(J_t) > 0$ (with $\sigma_{k+1} = 0$ if $\mathrm{rank}(J_t) \le k$), so that $\mathcal{S}_t$ is stable to estimation noise and small perturbations of J_t.

To compare mechanism subspaces across checkpoints, we use principal-angle geometry. For two k-dimensional subspaces $S, S' \subseteq \mathbb{R}^d$, let $\sin\Theta(S, S')$ denote the operator norm of the sine of canonical angles (equivalently, the spectral norm of the difference of orthogonal projectors up to constants). Our primary *mechanism change metric* is

$$\Delta_t \;:=\; \sin\Theta\big(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t-1}\big),$$

where $\widehat{\mathcal{S}}_t$ is an estimate of $\mathcal{S}_t$ computed from a budget of m randomized interventions at checkpoint t. The role of $\sin\Theta$ is twofold: it is invariant to the choice of basis for each estimated subspace, and it directly controls worst-case distortion of vectors in one subspace relative to the other. A "bifurcation" in our sense is a time t at which the dominant influence subspace rotates or expands enough that Δ_t exceeds a calibrated threshold γ. Accordingly, the core detection problem is to output a set of detected bifurcation times

$$\widehat{\mathcal{B}} \;\subseteq\; \{2, \dots, T\}, \qquad t \in \widehat{\mathcal{B}} \iff \Delta_t \ge \gamma,$$

with false-alarm control $\mathbb{P}(\widehat{\mathcal{B}} \cap \mathcal{B} = \emptyset) \ge 1 - \delta$ under a suitable no-change null, and with small detection delay under change alternatives.

While bifurcation detection is itself a monitoring output, our motivating application is *capability emergence forecasting*. We formalize a capability task $\mathcal{T}_j$ as an evaluation procedure producing a behavioral score $b_{t,j}$ from M_t (e.g. pass@k, exact-match, reward, or a safety-relevant refusal metric) and an operational emergence time

$$t_j \;:=\; \min\{t :\; b_{t,j} \ge \tau_j\},$$

for a pre-specified detectability threshold τ_j. The forecasting problem is to output $\widehat{t}_j$ (and an uncertainty interval) using the history of internal monitoring statistics up to time t, notably $\{\widehat{\mathcal{S}}_s\}_{s \le t}$, $\{\Delta_s\}_{s \le t}$, and optionally spectra $\{\sigma_i(\widehat{J}_s)\}$. We keep the forecasting map abstract: in practice it may be implemented as a calibrated regression from mechanism-space trajectories to expected behavioral onset, or as a rule-based scheme that treats certain bifurcations as precursors for particular tasks. The essential structural hypothesis is that task emergence is mediated by the appearance or reorganization of mechanisms that exert causal influence on a task-linked readout, so that changes in $\mathcal{S}_t$ can precede (and hence warn of) changes in $b_{t,j}$.

It is useful to contrast this setting with black-box emergence detection that observes only output behavior on $\mathcal{T}_j$. If for $t < t^*$ the output distributions $P_t(\cdot \mid x)$ are identical on the evaluation distribution used by $\mathcal{T}_j$, then

no procedure that only queries M_t through $\mathcal{T}_j$ can distinguish checkpoints prior to t^*, and hence no such procedure can provably provide early warning. In our formulation, early warning becomes possible precisely because we enrich the observation channel: we access internal states $h_t(x)$ and we apply randomized interventions whose responses identify an internal causal operator J_t. The point is not that internals are magically predictive, but that they provide additional information—in the literal statistical sense—that is absent from task behavior when behavior is masked, unelicited, or intentionally shaped.

We now proceed to the estimation problem implicit in the definitions above: given a checkpoint M_t, an intervention family, and a budget of m interventions, how do we compute $\widehat{J}_t$ or $\widehat{\mathcal{S}}_t$ with controlled error in high dimension? Section 4 describes randomized intervention schemes, low-rank sketches, and principal subspace recovery methods that make MBD computationally feasible at training scale.

4 Influence operator estimation

Fix a checkpoint t. Our goal is to estimate either the influence operator J_t itself (as a low-rank object) or, more modestly, its top-k right singular subspace $\mathcal{S}_t$, using only randomized on-manifold interventions at the tap. Throughout, we treat (1) as defining a linear regression problem with random design: for each sampled direction u, the $\mathcal{D}$-averaged response Δy has conditional mean $\alpha J_t u$ plus noise. The salient feature is that we may choose the design distribution over u, and we exploit this freedom to obtain simple unbiased estimators and efficient subspace recovery procedures.

Randomized intervention scheme and moment estimators. Let $x_1, \ldots, x_n \sim \mathcal{D}$ be a small context batch. For $i = 1, \ldots, m$, we sample $u_i \sim \mathcal{N}(0, I_d)$ (or Rademacher) and compute the batch-averaged intervention response

$$\Delta Y_i \;:=\; \frac{1}{n}\sum_{j=1}^{n}\Big(y_t^{\mathrm{do}}(x_j; u_i) - y_t(x_j)\Big) \in \mathbb{R}^p.$$

Under (H1) and conditional on u_i, we have $\mathbb{E}[\Delta Y_i \mid u_i] = \alpha J_t u_i$, with deviations captured by mean-zero subgaussian noise whose proxy decreases with n by averaging (we absorb this effect into σ^2 for notational simplicity). The basic unbiased outer-product estimator is then

$$\widehat{J}_t \;:=\; \frac{1}{\alpha m}\sum_{i=1}^{m}\Delta Y_i u_i^\top, \tag{2}$$

since $\mathbb{E}[u_i u_i^\top] = I_d$ implies $\mathbb{E}[\widehat{J}_t] = J_t$. This estimator is the natural analogue of simultaneous perturbation / score-function estimators, but applied

to internal activations rather than parameters. When u_i is Rademacher, $\mathbb{E}[u_i u_i^\top] = I_d$ still holds, and boundedness can improve constants in concentration.

Several variance-reduction tricks are operationally important. First, *paired* directions $\pm u$ cancel even-order nonlinearities and reduce sensitivity to baseline drift: defining $\widetilde{\Delta Y}(u) := \frac{1}{2}\big(\Delta Y(u) - \Delta Y(-u)\big)$, we have $\mathbb{E}[\widetilde{\Delta Y}(u) \mid u] = \alpha J_t u$ but empirically smaller residuals when the intervention semantics is only approximately additive. Second, we may enforce approximate orthogonality of the u_i (e.g. via a QR step on a Gaussian matrix) to reduce correlations in the design; this does not change (2) but improves finite-m stability in practice. Third, if the tap coordinates have strongly non-isotropic marginal variance under $\mathcal{D}$, we may sample u in a whitened coordinate system $u = \widehat{\Sigma}^{-1/2} z$ with $z \sim \mathcal{N}(0, I_d)$ and $\widehat{\Sigma}$ a running covariance estimate of $h_t(x)$; this modifies the estimand to $J_t \widehat{\Sigma}^{-1/2}$ unless we reweight appropriately, but it can substantially improve signal-to-noise for fixed α.

Low-rank structure and sketching without forming $\widehat{J}_t$. Even when $\widehat{J}_t$ is conceptually defined by (2), explicitly materializing a $p \times d$ matrix is undesirable. The key observation is that $\widehat{J}_t$ factors through the m interventions. Let $U := [u_1 \cdots u_m] \in \mathbb{R}^{d \times m}$ and $B := [\Delta Y_1 \cdots \Delta Y_m] \in \mathbb{R}^{p \times m}$. Then

$$\widehat{J}_t \;=\; \frac{1}{\alpha m} B U^\top, \tag{3}$$

so $\operatorname{rank}(\widehat{J}_t) \le m$. Consequently, all information needed for the top-k right singular subspace lies in the span of the sampled directions u_i, and we can recover $\widehat{\mathcal{S}}_t$ from $m \times m$ linear algebra.

Indeed,

$$\widehat{J}_t^\top \widehat{J}_t \;=\; \frac{1}{\alpha^2 m^2} U (B^\top B) U^\top,$$

so the nonzero eigenvectors of $\widehat{J}_t^\top \widehat{J}_t$ lie in $\operatorname{col}(U)$. Let $U = QR$ be a thin QR factorization with $Q \in \mathbb{R}^{d \times m}$ orthonormal and $R \in \mathbb{R}^{m \times m}$ invertible w.h.p. Then

$$\widehat{J}_t^\top \widehat{J}_t \;=\; Q \left(\frac{1}{\alpha^2 m^2} R (B^\top B) R^\top \right) Q^\top.$$

Thus, if $v_1, \ldots, v_k \in \mathbb{R}^m$ are the top-k eigenvectors of the $m \times m$ matrix $S := \frac{1}{\alpha^2 m^2} R (B^\top B) R^\top$, then $\widehat{\mathcal{S}}_t = \operatorname{span}\{Q v_1, \ldots, Q v_k\}$. This procedure costs $O(dm^2 + pm^2)$ to form $B^\top B$ and to orthonormalize U, and it avoids any $d \times d$ eigendecomposition. When p is moderately large, forming $B^\top B$ is still feasible because it is only $m \times m$; when p is very large (e.g. y_t is a full vocabulary logit vector), we may replace B by a compressed sketch RB with a random projection $R \in \mathbb{R}^{p' \times p}$, preserving inner products $\Delta Y_i^\top \Delta Y_{i'}$ up to a controlled distortion for $p' = \widetilde{O}(m)$.

An equivalent perspective, useful for iterative methods, is that we can apply $\widehat{J}_t$ and $\widehat{J}_t^\top$ to vectors without forming them:

$$\widehat{J}_t v = \frac{1}{\alpha m} B(U^\top v), \qquad \widehat{J}_t^\top w = \frac{1}{\alpha m} U(B^\top w).$$

Hence, randomized SVD and power iteration can be implemented with cost linear in $pm + dm$ per multiply, which is attractive when we wish to track multiple values of k or compute additional spectral diagnostics.

Choosing α and validating locality. The estimator (2) presumes a regime in which Δy is locally linear in u. In practice we select α by a calibration loop at each tap: we increase α until the signal $\|\Delta Y\|$ is reliably above numerical noise, but we require that a symmetry diagnostic remains small, e.g.

$$\frac{\|\Delta Y(u) + \Delta Y(-u)\|}{\|\Delta Y(u) - \Delta Y(-u)\|} \le \rho$$

for a chosen tolerance ρ. This heuristic checks for even-order terms and intervention artifacts. When it fails, we either reduce α or modify the on-manifold semantics (e.g. projecting back to a constraint set determined by activation statistics). We emphasize that α should be interpreted relative to the typical scale of $h_t(x)$; a robust choice is to normalize u to unit norm and set α as a small fraction of the root-mean-square activation magnitude at the tap.

Tap choice and multi-tap aggregation. The tap defines the causal interface at which we probe mechanisms, and different taps expose different abstractions. Early layers may yield diffuse, high-rank influence; later layers often yield lower-rank, more task-aligned mechanisms, but may also be more entangled with the readout definition. In applications we therefore recommend monitoring a small set of taps (e.g. several residual-stream layers and, when present, key/value streams), each with its own $\widehat{\mathcal{S}}_t^{(\ell)}$ and change statistic $\Delta_t^{(\ell)}$. A simple aggregation rule is

$$\Delta_t^{\max} := \max_{\ell} \sin\Theta\big(\widehat{\mathcal{S}}_t^{(\ell)}, \widehat{\mathcal{S}}_{t-1}^{(\ell)}\big),$$

which detects a bifurcation if any monitored interface changes. When mechanisms are distributed across layers, we may instead form a block-concatenated tap $h_t^{\mathrm{concat}} := (h_t^{(\ell_1)}, \dots, h_t^{(\ell_L)})$ and apply the same estimator in the enlarged ambient space; this increases d but preserves the low-rank nature of influential directions if only a few layers contribute. More structured aggregation is possible by aligning subspaces across taps (e.g. via canonical correlation) and clustering changes into regimes, but we defer such forecasting-oriented constructions to later sections.

The output of this section is an explicit, budgeted procedure for computing $\widehat{\mathcal{S}}_t$ from m interventions with minimal dependence on ambient dimension. The next section supplies finite-sample guarantees and minimax lower bounds that explain the observed α^{-1}, $m^{-1/2}$, and gap^{-1} scalings, and that guide principled choices of m and the detection threshold γ.

5 Theory I: subspace estimation bounds

In this section we quantify the finite-sample accuracy of estimating the mechanism subspace $\mathcal{S}_t$ from m randomized interventions at a fixed checkpoint t, and we show that the resulting dependence on $(\alpha, \sigma, \mathrm{gap}_t)$ is minimax-optimal up to absolute constants (and logarithmic factors in δ^{-1}). Throughout we work under the local linear model: for each intervention direction $u_i \sim \mathcal{N}(0, I_d)$,

$$\Delta Y_i \;=\; \alpha J_t u_i + \eta_i, \qquad \mathbb{E}[\eta_i \mid u_i] = 0, \tag{4}$$

where $\eta_i \in \mathbb{R}^p$ is conditionally subgaussian with variance proxy σ^2. We write $\widehat{J}_t$ for the unbiased estimator (2) and $\widehat{\mathcal{S}}_t$ for its top-k right singular subspace.

Operator-norm concentration for $\widehat{J}_t$. We begin by bounding $\|\widehat{J}_t - J_t\|_2$. Let

$$Z_i \;:=\; \frac{1}{\alpha m}\big(\Delta Y_i u_i^\top - \mathbb{E}[\Delta Y_i u_i^\top]\big) \;=\; \frac{1}{m}\Big(J_t(u_i u_i^\top - I_d) + \frac{1}{\alpha}\eta_i u_i^\top\Big),$$

so that $\widehat{J}_t - J_t = \sum_{i=1}^m Z_i$ is a sum of i.i.d. mean-zero random matrices. The first term, $J_t(u_i u_i^\top - I_d)$, is a (centered) quadratic form in a standard Gaussian and is controlled via Hanson–Wright or matrix Bernstein after truncation; the second term, $\alpha^{-1}\eta_i u_i^\top$, is a rank-one noise matrix. Since we ultimately pass to subspace error, we only require an operator-norm tail bound; in particular, we may treat p as absorbed into σ^2 (e.g. by defining ΔY_i to be a scalar probe, or by averaging over p coordinates so that the effective noise proxy decreases).

A convenient summary bound is: for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\widehat{J}_t - J_t\|_2 \;\le\; C_1\Big(\|J_t\|_2 + \frac{\sigma}{\alpha}\Big)\sqrt{\frac{\log(1/\delta)}{m}}, \tag{5}$$

for an absolute constant C_1. The term $\|J_t\|_2\sqrt{\log(1/\delta)/m}$ corresponds to the randomness of the design covariance $m^{-1}\sum u_i u_i^\top$ about I_d; the term $(\sigma/\alpha)\sqrt{\log(1/\delta)/m}$ is the regression noise scaled by the intervention magnitude α. In regimes where σ/α dominates $\|J_t\|_2$ (typical when α is chosen conservatively for locality), the leading scaling is $(\sigma/\alpha)m^{-1/2}$.

From operator error to subspace error via a gap. Let $J_t = U\Sigma V^\top$ be an SVD with singular values $\sigma_1(J_t) \geq \cdots \geq \sigma_{\min(p,d)}(J_t) \geq 0$, and let $\mathcal{S}_t = \mathrm{span}\{v_1, \ldots, v_k\}$ be the top-k right singular subspace. Define the spectral gap

$$\mathrm{gap}_t \;:=\; \sigma_k(J_t) - \sigma_{k+1}(J_t),$$

with the convention $\sigma_{k+1}(J_t) = 0$ if $\mathrm{rank}(J_t) \leq k$. The gap condition is not a technicality: without $\mathrm{gap}_t > 0$, the top-k subspace is not identifiable (an arbitrarily small perturbation can rotate it within a degenerate singular space).

Under $\mathrm{gap}_t > 0$, Wedin's $\sin\Theta$ theorem (or Davis–Kahan applied to $J_t^\top J_t$) yields

$$\sin\Theta(\widehat{\mathcal{S}}_t, \mathcal{S}_t) \;\leq\; \frac{\|\widehat{J}_t - J_t\|_2}{\mathrm{gap}_t}, \tag{6}$$

up to a benign constant depending on the specific normalization of $\sin\Theta(\cdot,\cdot)$. Combining (5) and (6) gives the advertised scaling.

Finite-sample upper bound. Specializing to the noise-dominated regime (or absorbing $\|J_t\|_2$ into σ/α by redefining σ), we obtain the following guarantee.

Theorem 5.1 (Subspace estimation upper bound). Assume (4) with subgaussian noise proxy σ^2 and assume $\mathrm{gap}_t > 0$. Let $\widehat{\mathcal{S}}_t$ be the top-k right singular subspace of $\widehat{J}_t$ built from m i.i.d. interventions. Then with probability at least $1 - \delta$,

$$\sin\Theta(\widehat{\mathcal{S}}_t, \mathcal{S}_t) \;\leq\; C\,\frac{\sigma}{\alpha}\sqrt{\frac{\log(1/\delta)}{m}} \cdot \frac{1}{\mathrm{gap}_t}, \tag{7}$$

for an absolute constant C.

Implications and choice of m. Fix a target accuracy $\varepsilon \in (0,1)$. Rearranging (7) yields that it suffices to take

$$m \;\gtrsim\; \frac{\sigma^2}{\alpha^2} \cdot \frac{\log(1/\delta)}{(\varepsilon\,\mathrm{gap}_t)^2}. \tag{8}$$

The dependence on α^{-2} captures a basic tradeoff: decreasing α improves locality but makes the regression problem harder. The dependence on gap_t^{-2} expresses ill-conditioning of the subspace: when the k-th and $(k+1)$-th singular values are close, we need substantially more interventions to stably recover the intended mechanism directions. Finally, the $m^{-1/2}$ rate is the usual parametric rate for estimating a low-dimensional object from i.i.d. noisy measurements.

If we wish to control error uniformly over checkpoints $t = 1, \dots, T$, we may apply a union bound and replace δ by δ/T, incurring an additional $\log T$ factor. This is typically mild relative to other sources of conservatism, but it is the correct scaling for provable streaming guarantees.

Matching lower bound (minimax necessity). We now show that the scaling in (8) is unavoidable. The proof follows the standard logic for subspace estimation: we construct two hypotheses whose top-k subspaces differ by a small principal angle, yet whose induced intervention-response distributions are statistically close when m is small.

For transparency, consider the simplest nontrivial case $k = 1$ and $p = 1$, where J is a row vector $J = \lambda w^\top$ with $\|w\|_2 = 1$. Then $\mathcal{S}(J) = \mathrm{span}\{w\}$, the gap is $\mathrm{gap} = \lambda$, and the observation model becomes

$$\Delta Y_i = \alpha\lambda\langle w, u_i\rangle + \eta_i,$$

a noisy linear functional of u_i. Let w and w' be two unit vectors with $\angle(w, w') = \varepsilon$, and consider $J = \lambda w^\top$ and $J' = \lambda (w')^\top$. Conditioned on $\{u_i\}_{i=1}^m$, the likelihood ratio between the two hypotheses is Gaussian with squared mean shift proportional to

$$\frac{\alpha^2\lambda^2}{\sigma^2}\sum_{i=1}^m \big(\langle w, u_i\rangle - \langle w', u_i\rangle\big)^2 \;=\; \frac{\alpha^2\lambda^2}{\sigma^2}\sum_{i=1}^m \langle w - w', u_i\rangle^2.$$

Taking expectation over $u_i \sim \mathcal{N}(0, I_d)$ gives $\mathbb{E}\langle w - w', u_i\rangle^2 = \|w - w'\|_2^2 \asymp \varepsilon^2$. Hence the KL divergence scales as

$$\mathrm{KL}(P_J^{(m)} \| P_{J'}^{(m)}) \;\lesssim\; m \cdot \frac{\alpha^2\lambda^2}{\sigma^2} \cdot \varepsilon^2 \;=\; m \cdot \frac{\alpha^2\,\mathrm{gap}^2}{\sigma^2} \cdot \varepsilon^2.$$

By Le Cam's two-point method, if this divergence is bounded by an absolute constant, then any estimator $\widehat{\mathcal{S}}$ must incur a constant probability of subspace error at least on the order of ε. Equivalently, to ensure $\sin\Theta(\widehat{\mathcal{S}}, \mathcal{S}) \le \varepsilon$ with high probability we require

$$m \;\gtrsim\; \frac{\sigma^2}{\alpha^2} \cdot \frac{1}{\mathrm{gap}^2} \cdot \frac{1}{\varepsilon^2}, \tag{9}$$

matching (8) up to the $\log(1/\delta)$ factor required for tail control. The general $k > 1$ case is obtained by rotating a k-dimensional subspace within a $(k+1)$-dimensional ambient subspace and applying Fano-type packing arguments; the same signal-to-noise ratio $\alpha\sqrt{m}/\sigma$ governs identifiability, and the gap again controls stability.

Taken together, (7) and (9) justify treating $(\sigma/\alpha) \cdot (\mathrm{gap}_t)^{-1} \cdot m^{-1/2}$ as the fundamental accuracy limit for influence-subspace monitoring at a fixed checkpoint.

6 Theory II: change-point detection guarantees

We now analyze the sequential detector based on the principal-angle increment

$$\Delta_t \;:=\; \sin\Theta(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t-1}), \qquad \widehat{\tau} \;:=\; \min\{t \geq 2 : \Delta_t \geq \gamma\},$$

where $\widehat{\mathcal{S}}_t$ is obtained from m randomized interventions at checkpoint t. Our goal is twofold: (i) control the probability of false alarms when the mechanism is stable, and (ii) upper bound the detection delay when the mechanism undergoes a low-rank perturbation. Throughout we work under (H1) and assume a uniform gap condition $\inf_t \mathrm{gap}_t \geq g > 0$ over the time range of interest.

A deterministic decomposition. Let $\mathcal{S}_t$ denote the population top-k right singular subspace of J_t, and define the estimation errors

$$e_t \;:=\; \sin\Theta(\widehat{\mathcal{S}}_t, \mathcal{S}_t).$$

By the triangle inequality for principal angles (in operator norm form), for each $t \geq 2$ we have

$$\Delta_t \;\leq\; e_t \;+\; \sin\Theta(\mathcal{S}_t, \mathcal{S}_{t-1}) \;+\; e_{t-1}, \qquad \Delta_t \;\geq\; \sin\Theta(\mathcal{S}_t, \mathcal{S}_{t-1}) \;-\; e_t \;-\; e_{t-1}. \tag{10}$$

Thus the detector succeeds whenever the intrinsic subspace motion $\sin\Theta(\mathcal{S}_t, \mathcal{S}_{t-1})$ dominates the estimation noise $e_t + e_{t-1}$, and it avoids false alarms whenever $e_t + e_{t-1}$ stays below γ in stable segments.

False alarm control under stationarity. We first consider the null regime in which $J_t \equiv J$ (hence $\mathcal{S}_t \equiv \mathcal{S}$) for all $t \in \{1, \ldots, T\}$. Then $\sin\Theta(\mathcal{S}_t, \mathcal{S}_{t-1}) = 0$, and (10) simplifies to $\Delta_t \leq e_t + e_{t-1}$. Consequently, if $e_t \leq \gamma/2$ for all t, no alarm can occur.

Invoking the subspace estimation guarantee from Theorem 5.1 with failure probability parameter $\delta' = \delta/T$, we obtain that for each fixed t,

$$\mathbb{P}\left(e_t > C\,\frac{\sigma}{\alpha}\sqrt{\frac{\log(T/\delta)}{m}} \cdot \frac{1}{g} \right) \;\leq\; \frac{\delta}{T}.$$

A union bound over $t = 1, \ldots, T$ yields uniform control over the entire trajectory.

Theorem 6.1 (No false alarms in a stable regime). Assume $J_t \equiv J$ for $t \in \{1, \ldots, T\}$ and $\mathrm{gap} \geq g > 0$. Choose $\gamma > 0$ and suppose

$$m \;\geq\; C_{\mathrm{FA}}\,\frac{\sigma^2}{\alpha^2} \cdot \frac{\log(T/\delta)}{(g\gamma)^2} \tag{11}$$

for an absolute constant C_{FA}. Then with probability at least $1 - \delta$, we have $\Delta_t < \gamma$ for all $t \geq 2$, hence $\widehat{\tau} = \infty$ (no false alarm).

Detection under a low-rank perturbation. We next analyze a single change-point model: there exists $t^* \in \{2, \dots, T\}$ such that $J_t \equiv J^-$ for $t < t^*$ and $J_t \equiv J^+$ for $t \geq t^*$, with

$$J^+ \;=\; J^- + \Delta, \qquad \operatorname{rank}(\Delta) \leq k, \qquad \|\Delta\|_2 \geq \Delta_0,$$

and $\operatorname{gap}(J^-), \operatorname{gap}(J^+) \geq g$. Let $\mathcal{S}^- := \mathcal{S}(J^-)$ and $\mathcal{S}^+ := \mathcal{S}(J^+)$, and define the intrinsic separation

$$\rho \;:=\; \sin\Theta(\mathcal{S}^+, \mathcal{S}^-).$$

The quantity ρ is the population-level signal available to a subspace-based detector. In particular, by (10),

$$\Delta_{t^*} \;\geq\; \rho - e_{t^*} - e_{t^*-1}. \tag{12}$$

Hence if we pick $\gamma \in (0, \rho)$ and ensure $e_{t^*-1}, e_{t^*} \leq (\rho - \gamma)/2$, then $\Delta_{t^*} \geq \gamma$ and the detector fires at the first changed checkpoint.

To connect ρ to the perturbation magnitude Δ_0, we may appeal to standard singular-subspace perturbation geometry. In the favorable case where Δ injects energy largely outside $\mathcal{S}^-$ (e.g. $\|(I - \Pi_{\mathcal{S}^-})\Delta\|_2$ is comparable to $\|\Delta\|_2$), one obtains a lower bound of the form $\rho \gtrsim \min\{1, \Delta_0/g\}$. We keep ρ explicit since it is the correct parameter governing detectability in the general (possibly partially aligned) case.

Theorem 6.2 (Immediate detection with high probability). Assume a single change at t^* with intrinsic separation $\rho = \sin\Theta(\mathcal{S}^+, \mathcal{S}^-) > 0$ and gaps at least g. Set $\gamma = \rho/2$. If

$$m \;\geq\; C_{\mathrm{det}}\, \frac{\sigma^2}{\alpha^2} \cdot \frac{\log(T/\delta)}{(g\rho)^2}, \tag{13}$$

then with probability at least $1 - \delta$ we have (i) no false alarm for $t < t^*$, and (ii) $\widehat{\tau} = t^*$ (zero detection delay).

Detection delay via temporal aggregation. If the per-checkpoint budget m is insufficient for (13), we can still obtain a delay bound by aggregating information across checkpoints after the change. One simple device is a sliding-window estimator: form an averaged sketch $\overline{J}_t := w^{-1} \sum_{s=t-w+1}^{t} \widehat{J}_s$ and let $\overline{\mathcal{S}}_t$ be its top-k right singular subspace. Under a post-change stationary regime, $\overline{J}_t$ has the same mean J^+ but reduced variance proxy by a factor w, effectively replacing m by mw in Theorem 5.1. Consequently, choosing w so that mw meets (13) yields detection after $O(w)$ checkpoints. In particular, for $\gamma = \rho/2$, it suffices to take

$$w \;\gtrsim\; \frac{\sigma^2}{\alpha^2} \cdot \frac{\log(T/\delta)}{m\,(g\rho)^2}, \tag{14}$$

so the detection delay scales as $O\Big(\frac{\sigma^2}{\alpha^2 m}\frac{\log(T/\delta)}{(g\rho)^2}\Big)$ under persistent post-change statistics.

Matching lower bounds (rank-one spike detection). Finally, we record a lower bound showing that the dependence on $\sigma^2/(\alpha^2\Delta_0^2)$ is unavoidable even in the simplest case. Consider $p = d$ and a rank-one spike model: under the null, $J = 0$; under the alternative, $J = \Delta_0 v w^\top$ with unknown unit vectors v, w. For an intervention $u \sim \mathcal{N}(0, I_d)$, the mean response is $\alpha\Delta_0\langle w, u\rangle v$, which is a Gaussian mean shift in an unknown direction. Standard minimax testing arguments imply that unless the aggregate signal-to-noise ratio exceeds a constant multiple of $\sqrt{\log(1/\delta)}$, no procedure can simultaneously make both type-I and type-II errors smaller than δ. Translating this into our parameters yields the necessary condition

$$m \;=\; \Omega\Big(\frac{\sigma^2}{\alpha^2\Delta_0^2}\log\frac{1}{\delta}\Big), \tag{15}$$

up to absolute constants (and with the same conclusion for per-checkpoint detectors in the sequential setting). Thus, in the regime where the emergent mechanism corresponds to a low-rank spike of size Δ_0, the preceding upper bounds are tight in their fundamental scaling: reliable early detection requires interventions sufficient to resolve a mean shift of magnitude $\alpha\Delta_0$ against noise σ, and this cannot be done with sublinear-in-$\sigma^2/(\alpha^2\Delta_0^2)$ samples.

7 Capability forecasting from bifurcations

We now describe how to convert detected mechanism-space bifurcations into forecasts of when a downstream capability becomes behaviorally detectable, together with calibrated uncertainty intervals. Fix a family of capability evaluations $\{\mathcal{T}_j\}_{j=1}^J$. For each task $\mathcal{T}_j$, let $b_{j,t} \in [0,1]$ denote a behavioral score at checkpoint t (e.g. pass@k, accuracy, or success rate on an episode distribution), computed on an evaluation distribution $\mathcal{D}_j$ that is disjoint from the monitoring distribution $\mathcal{D}$ when such separation is desired. We define the (random) onset time

$$t_j \;:=\; \min\{t \in \{1,\dots,T\} : b_{j,t} \geq \beta_j\},$$

for a fixed detectability threshold β_j chosen in advance (or via a separate calibration set). Our goal is, at intermediate times $t < T$, to output a point forecast $\widehat{t}_j(t)$ and an interval $\widehat{I}_j(t) \subseteq \{t, t+1, \dots, T\}$ such that $t_j \in \widehat{I}_j(t)$ with high probability while keeping the interval nonvacuous.

Regime segmentation as a sufficient statistic. Let $\widehat{\mathcal{B}} = \{\widehat{\tau}_1 < \cdots < \widehat{\tau}_L\}$ be the set of bifurcation times produced by the detector in Section 6, and let $\widehat{\tau}_0 := 1$, $\widehat{\tau}_{L+1} := T + 1$. This induces a partition of checkpoints into estimated regimes

$$\widehat{R}_\ell \;:=\; \{\widehat{\tau}_\ell\,, \widehat{\tau}_\ell + 1, \ldots, \widehat{\tau}_{\ell+1} - 1\}, \qquad \ell = 0, \ldots, L.$$

Within a regime $\widehat{R}_\ell$ we expect $\mathcal{S}_t$ to be stable, hence the interventional mechanism available to downstream readouts is approximately stationary. This motivates predicting capability onsets in terms of *time-since-last-bifurcation* rather than absolute training time. Concretely, write $\ell(t)$ for the regime index containing t, and define the elapsed time since the last detected bifurcation as

$$a(t) \;:=\; t - \widehat{\tau}_{\ell(t)}.$$

We will forecast t_j by modeling the conditional distribution of the remaining time-to-onset $r_j(t) := t_j - t$ as a function of regime-level features.

Feature map from mechanism monitoring. To allow task-specific forecasting while preserving the intervention budget, we restrict to features derived from monitoring quantities already computed for change detection. Let $\widehat{J}_t$ be the randomized-intervention sketch at checkpoint t (or its low-rank factors), and let $\widehat{\sigma}_{i,t}$ be the estimated singular values. We define a feature vector

$$\phi_t \;:=\; \Big(\widehat{\sigma}_{1,t}, \ldots, \widehat{\sigma}_{k,t},\; \Delta_t,\; a(t)\Big) \in \mathbb{R}^{k+2}.$$

When task-specific probes are permitted, we may augment ϕ_t by an alignment score between the mechanism subspace and a task probe direction. For example, if $\widehat{w}_{j,t} \in \mathbb{R}^d$ is a normalized direction obtained by fitting a linear probe for $\mathcal{T}_j$ on activations at the tap, then we define

$$A_{j,t} \;:=\; \|\Pi_{\widehat{\mathcal{S}}_t} \widehat{w}_{j,t}\|_2 \;\in\; [0, 1],$$

and set $\phi_{j,t} := (\phi_t, A_{j,t})$. Intuitively, $A_{j,t}$ measures whether the currently active mechanism subspace contains a direction known to be predictive for $\mathcal{T}_j$; empirically, onsets often follow shortly after such alignments increase sharply, and bifurcations provide a natural set of candidate times at which these increases occur.

A survival model conditioned on regimes. We postulate a conditional survival function for the remaining time $r_j(t)$ given the information at time t:

$$S_j(s \mid \phi_{j,t}) \;:=\; \mathbb{P}(r_j(t) > s \,|\, \phi_{j,t})\,, \qquad s \in \{0, 1, \ldots, T - t\}.$$

Operationally, one may instantiate S_j via a discrete-time hazard model $h_j(s \mid \phi) = \mathbb{P}(r_j(t) = s \mid r_j(t) \geq s, \phi)$, or via quantile regression for the conditional

distribution of $r_j(t)$. We emphasize that this modeling step is *auxiliary* to the core mechanism detector: it uses the detector's segmentation to reduce nonstationarity, but it may be learned from prior runs or from earlier phases of the same run (when some tasks already emerged).

Given an estimated conditional distribution $\widehat{S}_j(\cdot \mid \phi_{j,t})$, we define the point forecast as a conditional median

$$\widehat{t}_j(t) \;:=\; t + \inf\{s : \widehat{S}_j(s \mid \phi_{j,t}) \le 1/2\}.$$

More importantly, we define a one-sided $(1-\delta_j)$-prediction interval for t_j by conditional quantiles:

$$\widehat{I}_j(t) \;:=\; \Big[t,\; t + \inf\{s : \widehat{S}_j(s \mid \phi_{j,t}) \le \delta_j\}\Big],$$

which guarantees that, under correct calibration, the onset occurs within $\widehat{I}_j(t)$ with probability at least $1-\delta_j$ (conditional on the event $t_j \ge t$, which is known online).

Distribution-free calibration via conformalization. To avoid relying on correct specification of $\widehat{S}_j$, we employ split conformal prediction using a set of calibration runs $\mathcal{R}_{\mathrm{cal}}$. For each run $r \in \mathcal{R}_{\mathrm{cal}}$, we simulate the online procedure up to a fixed set of query times (e.g. each detected bifurcation time, or each checkpoint), compute a nonconformity score comparing realized $t_j^{(r)}$ to the model's predicted upper quantile, and then choose an inflation term as an empirical quantile of these scores. One convenient choice is to calibrate upper prediction bounds: let $\widehat{U}_j(t)$ be the uncalibrated upper bound produced by $\widehat{S}_j$, and define residuals

$$R_{j,t}^{(r)} \;:=\; t_j^{(r)} - \widehat{U}_j^{(r)}(t),$$

truncated at 0 since $t_j^{(r)} \ge t$. Let $\widehat{q}_j$ be the $(1-\delta_j)$-empirical quantile of $\{R_{j,t}^{(r)}\}$ over the calibration set. We then output the conformal upper bound $\widehat{U}_j^{\mathrm{conf}}(t) := \widehat{U}_j(t) + \widehat{q}_j$ and set $\widehat{I}_j(t) = [t, \widehat{U}_j^{\mathrm{conf}}(t)]$. Under exchangeability of runs (or a suitable martingale variant for mild drift), this yields marginal coverage $\mathbb{P}(t_j \le \widehat{U}_j^{\mathrm{conf}}(t)) \ge 1-\delta_j$ without assuming a correct parametric form.

Combining detector and forecaster uncertainty. The forecast is only meaningful if the regime segmentation is correct. Let $\mathcal{E}_{\mathrm{det}}$ denote the event that the detector produces no false alarm and detects each true bifurcation within the targeted delay (as guaranteed by the results of Section 6), and let $\mathcal{E}_{\mathrm{cal},j}$ denote the conformal coverage event for task j conditional on the features used by the forecaster. Then, by a union bound,

$$\mathbb{P}\big(t_j \in \widehat{I}_j(t)\big) \;\ge\; 1 - \mathbb{P}(\mathcal{E}_{\mathrm{det}}^c) - \mathbb{P}(\mathcal{E}_{\mathrm{cal},j}^c).$$

Thus we may set $\mathbb{P}(\mathcal{E}_{\text{det}}^c) \leq \delta_{\text{det}}$ via the detector thresholding analysis, and allocate per-task calibration levels δ_j such that $\delta_{\text{det}} + \sum_{j=1}^{J} \delta_j \leq \delta$, ensuring an overall failure probability at most δ across all tasks.

Interpretation and operational use. The resulting forecasting procedure is intentionally conservative: it issues short-horizon predictions when the detector indicates an actual mechanism-space change and when post-change features match patterns previously associated with emergence. This is consistent with Theorem 5.5: absent internal signals, one cannot certify earlier-than-behavioral warnings on $\mathcal{T}_j$. Here the warning signal is the bifurcation itself, together with regime-conditioned time-to-onset distributions learned from prior evidence. In the experimental section we will instantiate these forecasters in both synthetic curricula (where regime-to-onset mappings can be verified) and realistic training settings (where calibration across runs is essential).

8 Experimental protocol

We design experiments to test three questions: (i) whether the mechanism bifurcation detector (MBD) localizes ground-truth changes in interventional influence with controlled false alarms; (ii) whether bifurcation times improve forecasting of behavioral onsets relative to output-only and generic training-dynamics baselines; and (iii) whether the resulting signals are stable across changes in training objective (pretraining, finetuning, safety training) that may mask or unmask behavior. Throughout, we treat the intervention budget m per checkpoint as the binding resource, and we report results as a function of m and α.

Checkpointed training suites. We employ three suites, ordered from controlled to realistic. In *synthetic* suites we train small sequence models on curricula with known phase transitions, e.g. adding an auxiliary loss at a known step t^*, switching from memorization to algorithmic generalization, or injecting a new latent variable in the data generator. These suites provide an approximate ground truth for the change-point time t^* and for the identity of the affected readouts. In *semi-synthetic* suites we pretrain on a generic corpus and at t^* introduce a targeted capability objective (e.g. tool-call format consistency, structured reasoning traces, or an auxiliary RL head) while keeping the evaluation distribution fixed; here, behavioral onset can lag behind mechanism change. In *realistic* suites we use open training runs with periodic supervised finetuning and preference optimization stages; ground truth is not available, so we focus on internal consistency checks, cross-run reproducibility, and post-hoc alignment with known training events (objective switches, data-mixture shifts, safety interventions).

Monitoring and evaluation distributions. For each run we fix a monitoring distribution $\mathcal{D}$ used to estimate $\widehat{J}_t$ and $\widehat{\mathcal{S}}_t$. We consider (a) $\mathcal{D}$ as a held-out slice of the training mixture, and (b) $\mathcal{D}$ as a mixture that oversamples rare but safety-relevant contexts (prompt categories, adversarial templates, or long-context regimes). For each capability task $\mathcal{T}_j$ we separately fix an evaluation distribution $\mathcal{D}_j$ and a detectability threshold β_j, and we compute $b_{j,t}$ on disjoint examples to avoid trivial leakage from monitoring into evaluation.

Taps, readouts, and interventions. We predefine a tap $h_t(x) \in \mathbb{R}^d$ (typically a residual-stream activation at a fixed layer) and a readout $y_t(x) \in \mathbb{R}^p$. In the minimal setting $y_t(x)$ is a low-dimensional probe logit vector, but we also include agentic readouts (action logits, value head) when present. We implement randomized interventions $\mathrm{do}(h \leftarrow h + \alpha u)$ with $u \sim \mathcal{N}(0, I_d)$, choosing α relative to the empirical activation scale at the tap (e.g. a fixed fraction of $\mathrm{median}_x \|h_t(x)\|_2/\sqrt{d}$) to remain in the local linear regime. When an explicit on-manifold semantics is required, we compare (i) raw additive patching, (ii) additive patching followed by projection onto a learned manifold (e.g. an autoencoder latent), and (iii) conditional resampling within a neighborhood of $h_t(x)$ obtained from a replay buffer; we treat these as alternative operationalizations of the same abstract intervention family.

Estimation details and streaming constraints. At each checkpoint t we sample a batch $x_1, \ldots, x_n \sim \mathcal{D}$, run one baseline forward pass to obtain $y_t(x_j)$, and then run m intervened forward passes to obtain $\Delta Y_i = \frac{1}{n}\sum_j (y'_t(x_j) - y_t(x_j))$. We form the sketch $\widehat{J}_t = \frac{1}{\alpha m}\sum_{i=1}^m \Delta Y_i u_i^\top$ and compute $\widehat{\mathcal{S}}_t$ via randomized SVD, storing only k right singular vectors and the top k singular values. We fix k a priori in synthetic runs (where the ground-truth mechanism rank is controlled), and in realistic runs we select k by a stability heuristic: choose the smallest k such that $\widehat{\sigma}_{k,t}/\widehat{\sigma}_{1,t}$ exceeds a preset floor and the empirical $\sin\Theta$-trajectory is robust to resampling of interventions.

Change detection evaluation. In suites with known t^*, we measure (1) detection delay $\widehat{\tau} - t^*$ truncated below at 0, (2) false alarm rate $\mathbb{P}(\exists t < t^* : \Delta_t \geq \gamma)$, and (3) localization error $|\widehat{\tau} - t^*|$. We set γ either by the theoretical scaling $\gamma \propto \frac{\sigma}{\alpha}\sqrt{\frac{\log(1/\delta)}{m}}\frac{1}{\widehat{\mathrm{gap}}_t}$ using plug-in estimates, or by calibration on pre-change checkpoints to match a target false alarm level. In suites without known t^*, we report the empirical distribution of Δ_t, its reproducibility across random seeds, and its alignment with known interventions in the training script; we additionally test whether post-change regimes (as inferred by clustering $\widehat{\mathcal{S}}_t$) are stable under small perturbations to $\mathcal{D}$.

Forecasting evaluation. For each $\mathcal{T}_j$ we run the forecaster online and record (i) mean absolute error $|\widehat{t}_j(t) - t_j|$ at query times t (typically the end of each detected regime and a uniform subsample of checkpoints), (ii) interval width $|\widehat{I}_j(t)|$, and (iii) empirical coverage $\mathbb{P}(t_j \in \widehat{I}_j(t))$. We estimate $\widehat{S}_j$ using only features derived from monitoring ($\widehat{\sigma}_{1:k,t}$, Δ_t, $a(t)$, and optional alignments $A_{j,t}$), training on separate calibration runs $\mathcal{R}_{\text{cal}}$ and applying split conformalization to control marginal coverage at the chosen δ_j. When only a single run is available, we use a prefix of the run for calibration and evaluate on the suffix, and we report sensitivity to this split.

Baselines and compute-matched comparisons. We compare MBD-based early warning to four baseline families, all computed at the same checkpoint frequency and tuned to the same false alarm budget when applicable. (i) *Behavior-only*: declare emergence when $b_{j,t} \geq \beta_j$, which provides a lower bound on forecasting difficulty and is constrained by the indistinguishability phenomenon. (ii) *Probe emergence*: train linear probes on $h_t(x)$ for each $\mathcal{T}_j$ and monitor probe accuracy/slope over time; this is a strong internal baseline but does not attempt causal attribution and can be confounded by representation drift. (iii) *Training-dynamics monitors*: gradient norm $\|\nabla_\theta \mathcal{L}_t\|_2$, parameter update norm $\|\theta_t - \theta_{t-1}\|_2$, and loss curvature proxies such as a Hutchinson trace estimator for the Hessian or a small-number-of-steps Lanczos approximation; these are generic and cheap but task-agnostic. (iv) *Representation shift*: CKA/Procrustes distances between activation distributions across checkpoints, which capture global drift but not whether the drift is causally relevant to y. For each baseline we evaluate the same detection and forecasting metrics as above.

Ablations and robustness checks. We ablate m, α, and the tap choice, and we test robustness to the monitoring distribution by varying $\mathcal{D}$ while keeping $\mathcal{D}_j$ fixed. To probe the locality assumption, we empirically verify approximate linearity by checking that ΔY scales linearly in α over a small range and that estimated subspaces are stable to halving/doubling α. To probe the role of the spectral gap, we stratify checkpoints by $\widehat{\text{gap}}_t$ and test whether detection reliability degrades as predicted. Finally, we evaluate whether regime segmentation improves forecasting beyond raw time t by permuting regime labels or replacing $a(t)$ with absolute time as a negative control.

Reporting. All reported curves are aggregated over multiple random seeds and, when available, multiple independent runs; we provide uncertainty by bootstrap over runs or by nonparametric concentration across seeds. We treat δ as the global error budget and allocate δ_{det} and $\{\delta_j\}$ according to the union bound used in the forecasting construction, so that the reported

coverage corresponds to the stated failure probabilities under the experimental exchangeability assumptions. This protocol yields evidence for (or against) the claim that interventional mechanism subspaces supply an earlier and more invariant signal than behavior-only and generic training-dynamics monitors.

9 Case studies: masking, invariance, and strategic behavior

We now illustrate how influence-subspace monitoring behaves in training regimes where behavioral evaluation is known to be brittle, in the sense that capability-relevant mechanisms can be present while externally visible behavior is suppressed, or conversely behavior can appear abruptly once an external constraint is relaxed. The common pattern across the case studies below is that a behavioral metric $b_{j,t}$ is not a monotone function of the underlying interventional operator J_t, whereas the subspace trajectory $\widehat{\mathcal{S}}_t$ can remain stable across objectives and thus serve as a more invariant signature of the active mechanism.

Case study A (masking by safety finetuning). We consider a run with a pretraining phase followed by a safety finetuning phase whose explicit goal is to reduce the incidence of unsafe outputs on a red-teaming distribution. Let $y_t(x)$ be a low-dimensional readout that scores the model's propensity to produce an unsafe continuation on a fixed prompt family, so that higher y_t corresponds to greater risk. During safety finetuning we typically observe a rapid decrease in $\mathbb{E}_{x\sim\mathcal{D}_{\text{eval}}}[y_t(x)]$, and correspondingly a decrease in the behavior-only risk metric b_t at the chosen detectability threshold. However, when we compute $\widehat{\mathcal{S}}_t$ at a mid-to-late tap and track $\Delta_t = \sin\Theta(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t-1})$, we frequently find that the safety phase does *not* erase the previously learned influence directions; rather, it adds a new control mechanism that competes with or gates them.

Operationally, this appears as follows. Immediately after the objective switch, Δ_t exhibits a transient spike consistent with a rank-k perturbation Δ to J_t (a new regulating mechanism), after which $\widehat{\mathcal{S}}_t$ stabilizes. Crucially, the stabilized $\widehat{\mathcal{S}}_t$ remains close (in principal angles) to the pre-switch subspace on the coordinates that correspond to the underlying capability, even while the observed unsafe behavior is reduced. In linear-algebraic terms, safety finetuning can reduce the *effective* gain along a dangerous direction (shrinking the corresponding singular value in J_t, or introducing destructive interference in the composed downstream mapping) without rotating the right singular subspace by a comparably large amount. This yields a concrete monitoring implication: behavioral suppression is compatible with the persistence of an

influence subspace that can be reactivated by downstream distribution shift, removal of the safety head, or later finetuning.

Invariance across objectives as a diagnostic. The preceding phenomenon motivates a regime-level diagnostic: after segmenting checkpoints into regimes via the detected bifurcations, we compare the average inter-regime distance $\sin\Theta(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t'})$ to the corresponding change in behavior-only metrics. In masking regimes we observe a characteristic mismatch: the behavior changes substantially while $\widehat{\mathcal{S}}_t$ changes only in a low-dimensional complement, or changes briefly at the switch and then returns to a subspace close to its pre-switch location. When this occurs, we interpret the finetuning as primarily reweighting or gating an existing mechanism rather than removing it. Conversely, when safety finetuning truly removes a capability (e.g. catastrophic forgetting induced by aggressive data filtering), we observe a sustained rotation in $\widehat{\mathcal{S}}_t$ accompanied by a collapse of the associated singular values. The distinction matters for early-warning: in the former case, the capability remains latent and may reappear, whereas in the latter case it may need to be relearned to reemerge.

Case study B (unmasking and "sudden" behavioral onset). We next consider a run where a targeted capability becomes behaviorally detectable only after a late finetuning stage, even though the necessary internal machinery is learned earlier. A representative instance is structured tool-use: the model may learn to internally represent tool-call schemas and argument structure during pretraining (because such patterns appear in data), but it does not reliably emit the corresponding surface form until an explicit supervised finetuning stage enforces the format. Let $y_t(x)$ score an internal readiness signal for tool-use (for example, a probe over $h_t(x)$ predicting tool-call validity), while the behavioral metric b_t is strict format compliance on held-out prompts. In such settings, behavior-only monitoring declares emergence at the finetuning onset by construction.

Influence-subspace monitoring can separate two possibilities. In one regime, $\widehat{\mathcal{S}}_t$ associated with y_t forms and stabilizes *before* the behavioral onset, and finetuning primarily increases the singular values $\sigma_i(J_t)$ along a preexisting subspace (amplification). In the other regime, Δ_t shows a sustained rotation at finetuning onset, indicating that the relevant mechanism is genuinely created at that point (construction). Empirically, we often observe the amplification pattern: the top-k subspace is already present, with $\widehat{\sigma}_{1:k,t}$ low but increasing, and the behavioral threshold is crossed only once these singular values exceed a task-dependent gain. This provides an interpretable explanation for "sudden" behavioral emergence: the internal causal directions exist earlier, but their influence on the readout is too weak to be behaviorally detectable. From a forecasting perspective, this is precisely the

scenario in which internal monitoring can, in principle, yield earlier warning than output-only evaluation.

Case study C (strategic models and safety-trained adaptation). Finally, we address monitoring in the presence of models that may adapt to avoid detection. In a naive threat model, a sufficiently capable model could learn to keep outputs innocuous on the evaluation distribution while retaining a dangerous capability off-distribution; this is the behavioral indistinguishability captured by Thm. 5. The question for any monitoring scheme is therefore not "can we always win," but rather which assumptions suffice to retain signal. Influence-subspace monitoring shifts the burden to internal causal structure: if the model preserves a mechanism that can steer a safety-relevant readout y_t, then randomized interventions at a tap can elicit a measurable Δy even when baseline behavior is benign.

We emphasize two concrete implications. First, masking by strategic behavior is easier when the monitor is predictable; accordingly, we treat the intervention directions u and the context sampling $x \sim \mathcal{D}$ as private randomness, and we diversify $\mathcal{D}$ to include safety-relevant subdistributions. Second, a strategic model could in principle reshape its internal computation so that the dangerous behavior is implemented in directions orthogonal to monitored taps or readouts. This suggests a practical monitoring posture: we should monitor multiple taps (early, mid, late layers) and multiple readouts y (task probes, safety scores, action/value heads when present), and we should interpret large and persistent Δ_t as a change in *where* influence flows, not merely *whether* behavior is currently exhibited.

Monitoring takeaway. Across these case studies we arrive at a consistent interpretation: finetuning can strongly modulate behavior while leaving a capability-relevant influence subspace approximately invariant, and apparent behavioral phase transitions can correspond either to subspace creation (large rotation) or to gain increase on an existing subspace (small rotation, rising singular values). For safety monitoring, the salient risk is precisely the masked-capability regime in which behavior is suppressed but influence directions persist. Influence-subspace trajectories provide a principled handle on this regime, and they highlight a concrete failure mode for output-only early warning: behavioral evaluations can be rendered uninformative by objective changes, distribution shift, or strategic adaptation, even when internal mechanisms remain present and causally efficacious under intervention.

10 Limitations and extensions

We record several limitations of the present formalization and indicate extensions that appear technically and operationally necessary for deployment-

scale monitoring. Our aim here is not to weaken the main guarantees, but to make explicit which parts of the argument are doing work and where additional assumptions or engineering are required.

(i) The on-manifold intervention requirement. The estimator and guarantees are phrased in terms of an "on-manifold" perturbation family $\mathrm{do}(h \leftarrow h + \alpha u)$ for random directions u, together with a locality condition (H1) ensuring a linear response model $\Delta y = \alpha J_t u + \eta$. In practice, arbitrary additive perturbations to an internal activation may push the network into regions that are not visited under the data distribution, thereby producing responses that are uninterpretable as properties of the trained computation rather than properties of adversarial activation hacking. This is not merely a semantic concern: when off-manifold responses dominate, the effective noise proxy σ^2 becomes large and the sample complexity bound in Thm. 1 degrades to the point of uselessness.

Two extensions appear natural. First, we may replace the additive model by a constrained perturbation

$$\mathrm{do}\big(h \leftarrow \Pi_{\mathcal{M}_t}(h + \alpha u)\big),$$

where $\Pi_{\mathcal{M}_t}$ is a projection (or approximate projection) onto a learned activation manifold $\mathcal{M}_t$ at the chosen tap. Such a projection can be implemented by a trained autoencoder at that tap, by conditional resampling of nearby activations (from a cache indexed by context features), or by a small denoising model trained to map noisy activations back to typical ones. Second, we may abandon additivity and intervene in a parameterized family $h \leftarrow g_\phi(h, u)$ whose Jacobian at $u = 0$ spans the desired tangent directions. The mathematics then tracks the induced linear operator in u-space, and the same subspace perturbation analysis goes through with J_t replaced by the corresponding Gateaux derivative. In both cases we must re-verify (H1) empirically by sweep tests in α, checking that $\mathbb{E}[\Delta y \mid u]$ is approximately linear and that higher-order terms are controlled on the monitoring distribution.

(ii) Multi-step and agentic behaviors. Our basic readout $y_t(x) \in \mathbb{R}^p$ is defined for a single forward pass on a context x. For agentic systems, the capability of interest is often expressed over trajectories: actions affect future observations, and the relevant quantity is a functional of the entire interaction (e.g. probability of accomplishing a goal, cumulative reward, or violating a safety constraint). A direct extension is to let y_t denote a vector of per-time-step statistics (or a return) produced by unrolling the agent in a fixed simulator. Interventions may then be applied at a specific step s and tap ℓ, yielding an interventional operator $J_{t,s,\ell}$ that maps directions in $h_{t,s,\ell}$ to changes in a trajectory-level summary.

This extension introduces two technical complications. First, the effective dimension p of the readout can grow with horizon, and the response noise η becomes heavier-tailed due to environment stochasticity; we can mitigate this by averaging over multiple rollouts per intervention and treating the result as a subgaussian mean estimate. Second, the relevant causal effect may be *nonlocal* in time: an intervention at step s may change actions at many later steps. One way to preserve the present framework is to define y_t as a low-dimensional sufficient statistic of the trajectory (e.g. success indicator, safety violation indicator, or a small set of probe scores), so that p remains constant. Another is to treat the unrolled computation as a single composite map and to monitor a block operator that aggregates influence over multiple steps. In either case, the principal-angle change metric remains well-defined, but interpretation requires care: a rotation in $\widehat{\mathcal{S}}_t$ may correspond to a reorganization of *temporal credit assignment* rather than the appearance/disappearance of a single feed-forward circuit.

(iii) Scaling costs and the choice of taps. The per-checkpoint cost is dominated by m intervention forward passes, and the bounds show that m must scale as $\tilde{O}\big(\sigma^2/(\alpha^2\mathrm{gap}^2)\big)$ to reach a fixed subspace error. While this is acceptable for moderate T, monitoring every checkpoint of a large-scale run is expensive, and monitoring multiple taps multiplies cost further. We therefore view the present algorithm as a primitive that must be combined with cost-aware scheduling.

A practical extension is *adaptive checkpointing*: evaluate $\widehat{\mathcal{S}}_t$ densely early in training to estimate typical variability, then switch to sparse sampling (e.g. logarithmic in t) and trigger dense sampling only when a cheap proxy (such as a small number of interventions, or drift in activation statistics) exceeds a threshold. Another extension is *multi-tap compression*: if we monitor taps $\ell = 1, \ldots, L$, we may share intervention directions u across taps and estimate several sketches in parallel from the same perturbed run, reducing overhead relative to independent monitoring. Finally, if d is very large, we can replace isotropic $u \sim \mathcal{N}(0, I_d)$ by structured directions (Hadamard/Rademacher with fast transforms) and maintain streaming sketches of $\widehat{J}_t$ sufficient for top-k SVD, thereby keeping memory at $O((p+d)k)$ while reducing constant factors.

(iv) Transfer across architectures (Transformers, SSMs, and hybrids). Our definitions are agnostic to architecture, provided we can identify a tap $h_t(x)$ and implement controlled interventions. Nonetheless, architectural differences matter for what constitutes a meaningful tap and what "on-manifold" means. For Transformers, residual stream vectors at a fixed layer and position provide a canonical $h_t(x)$, and additive perturbations can be made approximately on-manifold using residual statistics conditioned on

token identity. For state-space models (SSMs), $h_t(x)$ may correspond to a latent state evolving across sequence positions; intervening on that state can have long-range effects that are sensitive to stability properties of the recurrence. In such settings, it may be preferable to intervene on *inputs* to the state update (e.g. pre-activation channels) rather than on the state itself, or to constrain interventions to preserve norms that control dynamical blow-up.

A second issue is comparability across architectures. If we wish to compare mechanism trajectories between a Transformer and an SSM trained on the same objective, the ambient spaces $\mathbb{R}^d$ differ and the subspaces $\mathcal{S}_t$ live in non-identical domains. One extension is to compare *effects* rather than *directions*: the singular values $\sigma_i(J_t)$ and the left singular vectors in readout space are directly comparable when y_t is shared (same probe head or same behavioral statistic). Another is to introduce an alignment map $A : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ learned from paired activations on the same contexts, and then compare $A(\widehat{\mathcal{S}}_t^{(1)})$ to $\widehat{\mathcal{S}}_t^{(2)}$ via principal angles. This pushes the burden onto the identifiability of A, but it converts "transfer" into a standard subspace alignment problem with well-studied failure modes.

Summary. The present framework is strongest when (a) we can implement interventions that remain within the typical activation manifold, (b) the capability signal can be summarized by a low-dimensional readout, and (c) we can afford m interventions at the monitored cadence. Extending the method to trajectory-based capabilities and to heterogeneous architectures appears feasible, but requires additional design choices (projection families, rollout averaging, alignment maps) that are not captured by the clean model (H1)–(H2). These extensions also clarify how our proposal should be situated relative to adjacent literatures, to which we now turn.

11 Related work and positioning

Our formalization sits at the intersection of three lines of work: (a) empirical "stagewise" or punctuated learning dynamics during training, (b) mechanistic interpretability methods that localize circuits via interventions, and (c) classical sensitivity analyses (influence functions, Jacobians) together with statistical change-point detection. We briefly position our contribution relative to each, emphasizing what is gained by phrasing emergence as a change in an *interventional influence operator* rather than in behavior or in parameters.

Stagewise learning, phase transitions, and emergent capabilities. A growing empirical literature reports that capabilities can appear abruptly as training proceeds—sometimes described as "phase transitions" or "grokking-like" phenomena—in which behavioral metrics remain flat and then rise

quickly over a narrow range of steps **???**. Several mechanisms have been proposed (implicit regularization, representation reconfiguration, curriculum effects, scaling laws with thresholds), but most analyses are conducted in *behavior space*: one monitors a task loss/accuracy curve, or a battery of benchmark metrics, and declares an emergence when performance crosses a threshold. Our Thm. 5 isolates a limitation of this paradigm for early warning: if the output distributions on the evaluated task are indistinguishable prior to a time t^*, then no black-box monitor can infer the change earlier than t^*. In this sense, behavior-only monitoring is not merely practically brittle; it is information-theoretically incapable of guaranteeing pre-emergence detection without auxiliary signals. Our proposal is to use training-time access to internal interventions to define such a signal in a way that remains statistically tractable and comparable across checkpoints.

Representation drift and similarity metrics. A separate line of work studies how internal representations evolve through training using similarity measures (e.g. CKA, SVCCA, Procrustes alignment) and subspace tracking of activations **??**. These methods can detect that a layer's activation geometry changes, and they are often efficient since they require only forward passes. However, they conflate changes that are mechanistically important for a given readout with changes that are irrelevant (e.g. basis rotations in null directions, or redistribution of variance unrelated to the downstream statistic of interest). By contrast, our object J_t is explicitly *task-conditional*: it is the causal linear response from a specified internal tap $h_t(x)$ to a specified readout $y_t(x)$. The use of principal angles between the top-k right singular subspaces of J_t can be viewed as an "effect-weighted" analogue of subspace drift, where directions in activation space are weighted by their causal efficacy rather than by marginal variance. This distinction matters precisely in the early-warning regime: we aim to detect the reorganization of computations that *could* support a capability before that capability is expressed in the evaluation outputs.

Mechanistic interpretability and intervention-based localization. Mechanistic interpretability has developed a rich toolkit for attributing behavior to internal components, including activation patching, causal tracing, path patching, and circuit discovery **???**. These methods are often highly informative but typically proceed via *structured* interventions targeted to hypothesized features, specific heads, or curated datasets, and they aim at interpretability rather than at online detection. Our setting differs in two ways. First, we adopt a deliberately *randomized* intervention family $u \sim \mathcal{N}(0, I_d)$ so that the induced estimator $\widehat{J}_t$ admits concentration bounds and yields quantitative false-alarm control via standard perturbation theory (Thm. 1–Thm. 4). Second, we treat mechanistic structure at the level of a *low-rank*

operator (or subspace) rather than at the level of individually named units. This sacrifices immediate interpretability, but it produces a stable target for tracking across checkpoints and supports streaming estimation under a hard budget m. We view this as a complement to circuit analysis: the detector flags candidate bifurcation times at which deeper interpretability work is most valuable.

Causal abstraction and causal representation learning. There is a large literature on causal abstraction and the identification of causal variables within learned representations **??**. Our framework is compatible with this agenda but makes a different modeling choice: we do not attempt to recover a full causal graph over internal variables; rather, we monitor a localized causal map from one tap to one readout under small on-manifold perturbations. In categorical terms, J_t is a local linearization of an interventional morphism. One can interpret the top-k subspace $\mathcal{S}_t$ as the "tangent mechanism" at the tap that is relevant for y_t. The change metric $\sin\Theta(\mathcal{S}_t, \mathcal{S}_{t-1})$ then detects when the tangent mechanism undergoes a qualitative reorientation. This is weaker than full causal abstraction, but it is exactly what enables finite-sample guarantees under minimal assumptions (subgaussian noise and a spectral gap) and makes the approach suitable for deployment-scale monitoring.

Influence functions, Jacobians, and linear response. Our influence operator J_t is conceptually adjacent to influence functions and sensitivity analyses in statistics and deep learning **?**, as well as to Jacobian-based saliency and linear response methods. The difference is that classical influence functions study the effect of *training-point* upweighting on parameters and predictions, whereas we study the effect of *internal state* perturbations on a readout at a fixed checkpoint. If gradients are available, one could estimate J_t by backpropagating from y to h, but that yields a pointwise derivative that is not robust to nondifferentiable interventions, simulator rollouts, or restricted-access settings. Our randomized-intervention estimator is intentionally gradient-free and remains meaningful even when the intervention semantics are implemented via conditional resampling or projection operators (cf. Sec. 10(i)). In this sense, we treat J_t as an empirically identified causal operator rather than a purely differential object.

Change-point detection and online monitoring. Finally, our detector $\widehat{\tau} = \min\{t : \Delta_t \geq \gamma\}$ is a specialized instance of online change-point detection, but with a nonstandard observation model: we do not observe J_t directly; we observe noisy randomized linear measurements of it induced by interventions. Theorems 3–4 show that the resulting sample complexity matches the natural signal-to-noise scaling $\alpha\Delta_0\sqrt{m}/\sigma$ up to constants

and logarithms, and that one cannot, in general, do better. This places the problem closer to spiked-matrix detection and subspace tracking under noise than to classical scalar CUSUM-style settings. Moreover, by framing the monitored quantity as a subspace (rather than a full operator), we obtain robustness to benign reparameterizations and reduce the monitoring target to a low-dimensional object that can be stored and compared over long runs.

What this work adds. In summary, our contribution is not a new interpretability primitive per se, but a *statistically controlled monitoring formalism* that turns internal randomized interventions into an early-warning signal with explicit false-alarm control and lower bounds. The central modeling move is to define emergence as a change in the top-k influence subspace of a causal operator J_t. This yields (i) a quantitative estimator with provable accuracy under transparent assumptions, (ii) a natural change metric with a geometry that supports perturbation analysis, and (iii) an impossibility separation (Thm. 5) clarifying when internal access is necessary for any pre-emergence guarantee. We view these as the minimal ingredients needed to connect mechanistic signals to safety-relevant training-time forecasting.

12 Conclusion: influence bifurcations as a training-time early warning signal

We have formulated training-time capability emergence prediction as an online change-point problem over *interventional influence operators*. The central object J_t is defined by local, on-manifold perturbations at a fixed tap and measures the causal linear response from internal state $h_t(x)$ to a readout $y_t(x)$. From J_t we extract a low-dimensional summary—the top-k right singular subspace $\mathcal{S}_t$—and we monitor its geometry through principal-angle distances $\Delta_t = \sin\Theta(\widehat{\mathcal{S}}_t, \widehat{\mathcal{S}}_{t-1})$. Within the local-linearity and spectral-gap assumptions, this yields a monitor whose statistical behavior is analyzable: we can trade intervention budget m, intervention magnitude α, and false-alarm level δ for explicit bounds on subspace estimation and change detection. In particular, the regime of interest for safety is precisely the regime where behavior may be flat or strategically suppressed, while mechanisms reorganize; the subspace monitor is designed to be sensitive to the latter.

The practical interpretation is that a *bifurcation* is not defined by a performance curve crossing a benchmark threshold, nor by a parameter-space event, but by a detectable reorientation in the directions of internal state that causally control a chosen readout under small interventions. This choice is deliberate. Parameter drift is ubiquitous and rarely safety-relevant by itself; behavior, while ultimately decisive, can be delayed or masked. Influence subspaces provide an intermediate signal: they are closer to "what computation is available" than to "what computation is exercised on the

evaluation distribution." Theorems 1–4 justify that this intermediate signal is, at least in an idealized but transparent model, estimable with quantifiable uncertainty under a strict per-checkpoint intervention budget. Theorem 5, conversely, formalizes why we should not expect behavior-only early warning guarantees in general.

A deployment-oriented reading of our formalism suggests the following minimal roadmap for incorporating influence-bifurcation monitoring into a safety case.

(i) Specify monitoring interfaces: taps, readouts, and semantics. Any safety claim must begin with a concrete interface: which internal taps $h_t(x)$ are monitored, which readouts $y_t(x)$ are deemed safety-relevant, and what it means to intervene "on-manifold." The last point is not cosmetic: the local linear model (H1) is only meaningful relative to a specified intervention semantics, e.g. conditional resampling, constrained patching, or projection onto a learned activation manifold. We recommend treating the intervention operator as part of the monitored system, subject to validation tests (e.g. locality checks as $\alpha \to 0$, invariance checks across batches, and robustness checks across prompt subclasses).

(ii) Choose $\mathcal{D}$ to include safety-relevant subdistributions. The monitored distribution $\mathcal{D}$ determines what mechanisms are illuminated by randomized interventions. For a safety case, $\mathcal{D}$ should be explicitly constructed to include (a) routine traffic (to avoid monitoring an irrelevant corner), (b) "stress tests" that activate potentially dangerous latent skills, and (c) a small set of deliberately adversarial contexts. In our notation, nothing prevents $\mathcal{D}$ from being a mixture distribution; indeed, mixture design is the natural way to tie the monitor to concrete hazard models. The cost is statistical: if $\mathcal{D}$ is too diffuse, the effective signal in Δy may be diluted; this should be treated as a design tradeoff rather than an afterthought.

(iii) Calibrate m, α, k, γ via explicit error budgets. For an operator-norm subspace error target ε at failure probability δ, Theorem 1 provides the basic scaling $m = \tilde{O}(\sigma^2/(\alpha^2\varepsilon^2\mathrm{gap}_t^2))$. In deployment we will not know σ or gap_t a priori; nevertheless, we can proceed by (a) estimating empirical noise proxies from repeated interventions, (b) tracking empirical singular value gaps of $\widehat{J}_t$, and (c) using conservative lower confidence bounds on the gap to set monitoring budgets. The change threshold γ should be tied to a target false-alarm rate over T checkpoints; sequential correction (e.g. spending δ across time) is required if we wish to interpret alarms probabilistically. We emphasize that γ is not purely a statistical knob: it implicitly specifies what magnitude of mechanism rotation constitutes a safety-relevant event.

(iv) Treat detection as triage; couple to interpretability and evaluation. A bifurcation alarm should be viewed as a triage signal that triggers deeper analysis. Operationally, upon an alarm at $\widehat{\tau}$ one may (a) allocate additional interventions at neighboring checkpoints to refine $\widehat{\mathcal{S}}_{\widehat{\tau}}$ and localize which taps/readouts exhibit the change, (b) run targeted mechanistic interpretability to translate the subspace shift into candidate features/circuits, and (c) expand behavioral evaluations to tasks hypothesized to become newly feasible. This two-stage workflow aligns with the hardness discussion: dense randomized detection can be statistically efficient, while sparse or human-interpretable attribution may be computationally and methodologically harder and is better focused on a small set of flagged times.

(v) Incorporate forecasting as a conditional, model-based claim. Our problem definition allows optional capability tasks $\{\mathcal{T}_j\}$ and forecasts $\widehat{t}_j$. We regard such forecasts as inherently conditional: they are meaningful only relative to a model that links changes in J_t (or $\mathcal{S}_t$) to eventual behavioral detectability on $\mathcal{T}_j$. In practice, this suggests learning a mapping from mechanism-space features (e.g. singular values along task-aligned directions, or distances to previously observed "capability-on" regimes) to predicted emergence times, with uncertainty intervals calibrated on historical training runs. Theorems 3–4 delimit what can be claimed without such modeling: we can reliably detect sufficiently large changes in influence, but we cannot, without additional structure, convert every detected mechanism shift into a definitive capability timeline.

Limitations and open problems. Several extensions are necessary for robust safety deployment. First, local linearity (H1) may fail for larger α or for taps near discrete decisions; adaptive choice of α and non-linear response models are natural next steps. Second, real training runs may exhibit gradual drift rather than sharp change-points; replacing the binary detector with continuous regime tracking and drift rate estimates would better match such dynamics. Third, multiple comparisons across many taps/readouts demands principled aggregation (e.g. hierarchical testing) to avoid inflated false alarms. Fourth, the adversarial-noise reduction warns that monitoring guarantees depend on threat-model assumptions about the training process; if the model can adapt to the monitor, then trusted execution and randomized, withheld monitoring distributions become part of the security boundary.

Closing perspective. We do not claim that influence-subspace monitoring alone resolves the problem of anticipating dangerous capability emergence. We claim a more limited but, we think, necessary point: if one seeks provable early warning, one must monitor signals that are *not* reducible to black-box behavior on the target task, and one must do so in a way

that admits statistical calibration under realistic budgets. Influence bifurcations provide such a signal: they are mechanistically grounded, amenable to finite-sample analysis, and naturally integrated into a safety workflow in which alarms trigger targeted investigation and governance actions (pauses, audits, and evaluation expansion). This is the role we intend the present formalism to play in training-time safety cases.