Process-Tensor Tomography of SGD: Measuring Non-Markovian Memory via Back-Flow of Distinguishability

Vasileios Sevetlidis **George Pavlidis**
Athena Research Center
Kimmeria Campus
Xanthi, GR-67100, Greece

Abstract

We model neural training as a classical multi-time map from controllable interventions—batch choices, augmentations, and optimizer micro-steps—to model predictions on a fixed probe set. On this basis, we introduce a simple, model-agnostic witness of training memory based on *back-flow of distinguishability*. In a controlled two-step protocol, we compare predictive distributions after one intervention versus two; a positive increase $\Delta_{\mathrm{BF}} = D_2 - D_1 > 0$, with $D \in \{\mathrm{TV}, \mathrm{JS}, \mathrm{H}\}$, certifies observable-level non-Markovianity. Across controlled SGD experiments, we observe consistent positive back-flow with tight bootstrap confidence intervals, stronger effects under higher momentum, larger batch overlap, and more micro-steps, and marked collapse under a *causal break* that resets optimizer state. The witness is inexpensive, requires no architectural changes, and is robust across TV/JS/Hellinger. We position this as a measurement contribution: a practical diagnostic, and empirical evidence, that real training often deviates from the Markov idealization in ways that matter for optimizer behavior, data order, and schedule design.

1 INTRODUCTION

Modern analyses of training often treat updates as approximately Markovian once the algorithmic state is augmented with optimizer state: given the current parameters and buffers, the next update is assumed to summarize all relevant past information. This idealization underlies many analyses of SGD and its stochastic-differential-equation (SDE) approximations, but practical training uses momentum, correlated batch sequences, repeated examples, curricula, and schedule changes that can make current behavior depend on recent history [Bottou et al., 2018, Mandt et al., 2017, Shamir, 2016, Rajput et al., 2020, Sutskever et al., 2013, Goh, 2017]. A central question is therefore: *when does training behave as if it were memoryless, and when does recent history still matter?*

We address this as a *measurement* problem. Our contribution is not a new optimizer or training trick, but a simple operational diagnostic for testing whether a short training segment behaves as if a fixed second intervention depends only on the currently observed model state, or still depends on how that state was reached. This is important because Markov/SDE idealizations are often used to reason about optimizer behavior, data order, reshuffling, and schedules, yet practitioners lack a direct test of when those idealizations are reliable.

Our formalism is inspired by open-systems theory, in particular the process-tensor viewpoint [Pollock et al., 2018b,a, Milz and Modi, 2021, White et al., 2022], but all states, kernels, and observables considered here are entirely classical probability objects. In ML terms, we model short-horizon training as a multi-time stochastic map from controlled interventions to model predictions on a fixed probe set. Here, an *instrument* is a training intervention (mini-batch choice, augmentation, optimizer micro-steps), the *observable* is the predictive distribution on the probe set, and a *causal break* cuts a memory channel by resetting optimizer state before the second step.

Building on this view, we introduce a witness of training memory based on *back-flow of distinguishability*. We compare two one-step histories that differ only in

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

the first intervention and then apply the same second intervention to both. If the distance between their predictive distributions increases after the second intervention, then that intervention cannot be represented as a single memoryless channel acting only on the measured one-step outputs. The process is therefore operationally non-Markovian at the observable level. The witness is model-agnostic, inexpensive, and requires no architectural changes. Empirically, we find systematic positive back-flow in controlled two-step SGD experiments, amplification under higher momentum, larger batch overlap, and more micro-steps, and collapse under a *causal break* that resets optimizer buffers. We position this work as a practical diagnostic for assessing whether training behaves as a Markovian process at the observable level, and as empirical evidence that real training often deviates from this idealization in ways that affect order sensitivity and schedule design.

Contributions.

1. **Operational diagnostic for training memory.** We introduce a two-step intervention protocol and a back-flow statistic on predictive distributions (TV/JS/Hellinger) that tests whether a fixed second training intervention behaves as a memoryless map at the observable level.
2. **Mechanism test via causal break.** Resetting optimizer state before the second intervention provides a direct falsifiable test of whether optimizer buffers mediate the effect [Pollock et al., 2018a].
3. **Empirical evidence across datasets and architectures.** On CIFAR-100 and Imagenette, across multiple vision backbones and three divergences, we measure consistent observable-level memory effects with seed-robust confidence intervals.

2 RELATED WORK

Markov and SDE-style analyses of SGD typically treat training as approximately memoryless once the state is augmented with optimizer buffers, while random reshuffling, momentum, and dependent-data settings make temporal dependence explicit [Bottou et al., 2018, Mandt et al., 2017, Shamir, 2016, Rajput et al., 2020, Sutskever et al., 2013, Goh, 2017, Even, 2023, Doan et al., 2020, Dorfman and Levy, 2022]. Our work is also related to empirical studies of non-Gaussian gradient noise [Simsekli et al., 2019, Zhu et al., 2018], influence and memorization diagnostics [Ye et al., 2023, Carlini et al., 2019, 2021, Koh and Liang, 2017, Pruthi et al., 2020, Yeh et al., 2018], and representation-similarity measures such as CKA [Kornblith et al., 2019, Davari et al., 2022]. Unlike these lines, we introduce a measurement-first, intervention-defined witness tied directly to an observable-level Markov condition. Additional related-work discussion is deferred to the appendix.

3 METHODOLOGY

We formalize a short segment of training in a fully classical stochastic setting as a map from controlled training interventions to model predictions on a fixed probe set. In our experiments, the interventions are mini-batches, augmentations, and optimizer micro-steps, and the measurement is the probe-set predictive distribution. Using terminology borrowed from open-systems theory, this defines a two-time *process tensor*[1]. Our goal is to test whether, for a fixed second intervention, the observed two-step effect is explained by a single memoryless map acting on the measured one-step output, or whether it still depends on how that output was produced.

3.1 State spaces, instruments, and kernels

We distinguish between a *latent training state*, which contains all algorithmic variables needed to continue training, and an *observable*, which is what we actually measure from the model.

Let $(\mathcal{S}, \Sigma_{\mathcal{S}})$ denote the latent state space. An element $s \in \mathcal{S}$ is the full training state, consisting of model parameters together with any optimizer buffers (for example, momentum variables). Let $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ denote the observable space. In our setting, an element $y \in \mathcal{Y}$ is the collection of per-example class-probability vectors produced by the model on a fixed probe set. We write $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(\mathcal{Y})$ for the corresponding spaces of probability measures.

A *stochastic kernel* $K : \mathcal{A} \times \Sigma_{\mathcal{B}} \to [0, 1]$ maps each input $a \in \mathcal{A}$ to a probability measure $K(a, \cdot)$ on $(\mathcal{B}, \Sigma_{\mathcal{B}})$, measurably in a. Intuitively, $K(a, \cdot)$ is the distribution over next states or outputs obtained from input a.

A *training instrument*, denoted by I, is a user-controlled short-horizon training intervention. In the experiments below, an instrument specifies (i) a mini-batch index set or mini-batch stream, (ii) a data-augmentation rule, and (iii) optimizer micro-parameters used for k micro-steps. The special symbols A, A', and B denote three such instruments: A and A' are the alternative first-step interventions, while B is the common second-step intervention used to test memory.

[1] All objects in this work are classical probabilistic objects and no quantum computation is involved

Executing an instrument I induces a latent stochastic kernel

$$\mathcal{K}_I : \mathcal{S} \rightsquigarrow \mathcal{S},$$

which represents the random update over the corresponding k micro-steps. Starting from an initial distribution $\pi_0 \in \mathcal{P}(\mathcal{S})$, the one-step latent law after applying I is

$$\pi_1^{(I)} = \pi_0 \mathcal{K}_I.$$

Given two successive instruments I_0 and I_1, the latent two-step law is

$$\pi_2^{(I_0,I_1)}(\mathrm{d}s_2) = \int_{\mathcal{S}} \int_{\mathcal{S}} \mathcal{K}_{I_1}(s_1, \mathrm{d}s_2)\, \mathcal{K}_{I_0}(s_0, \mathrm{d}s_1)\, \pi_0(\mathrm{d}s_0). \tag{1}$$

We observe the latent state through an *observation channel*

$$\mathcal{M} : \mathcal{S} \rightsquigarrow \mathcal{Y},$$

which maps a latent training state to the corresponding predictive distribution on the probe set. The induced observable one- and two-step laws are

$$\Phi_1^{(I_0)}(\mathrm{d}y) = \int_{\mathcal{S}} \mathcal{M}(s_1, \mathrm{d}y)\, \pi_1^{(I_0)}(\mathrm{d}s_1), \tag{2}$$

$$\Phi_2^{(I_0,I_1)}(\mathrm{d}y) = \int_{\mathcal{S}} \mathcal{M}(s_2, \mathrm{d}y)\, \pi_2^{(I_0,I_1)}(\mathrm{d}s_2), \tag{3}$$

where $\pi_1^{(I_0)} = \pi_0 \mathcal{K}_{I_0}$.

For readability, we write kernel composition using the link product

$$(K_2 \star K_1)(a, \mathrm{d}c) := \int K_2(b, \mathrm{d}c)\, K_1(a, \mathrm{d}b).$$

Thus successive training interventions compose in the expected way: first apply $\mathcal{K}_{I_0}$, then $\mathcal{K}_{I_1}$, and finally map to the observable through $\mathcal{M}$.

Collecting all two-time statistics defines a multilinear map

$$\mathsf{T}_{2:0} : \mathfrak{I}_0 \times \mathfrak{I}_1 \to \mathcal{P}(\mathcal{Y}), \qquad (I_0, I_1) \mapsto \Phi_2^{(I_0,I_1)},$$

where $\mathfrak{I}_t$ is a convex set of admissible randomized instruments. In the language borrowed from open-systems theory, $\mathsf{T}_{2:0}$ is the two-time *process tensor*; in ordinary ML terms, it is simply the multi-time map from controlled training interventions to the resulting distribution of probe-set predictions.

3.2 Operational Markov condition at the observable

The key question is whether, for a fixed second intervention B, the two-step observable law is determined solely by the one-step observable law, regardless of how that one-step law was produced.

Definition 1 ***(Operational Markov condition (OMC) at the observable)*** *Fix a second-step instrument B. We say that the two-time process is* operationally Markov at the observable for B *if there exists a stochastic map*

$$\Lambda_B : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$$

such that for every first-step instrument I,

$$\Phi_2^{(I,B)} = \Lambda_B\big[\Phi_1^{(I)}\big]. \tag{4}$$

Intuitively, once the observed mid-time law Φ_1 is fixed, the second intervention B acts through the *same* channel Λ_B, independent of the earlier history that prepared Φ_1. This condition is weaker than assuming latent Markov dynamics in the full state (θ, buffers): it constrains only what is visible at the chosen observable.

Our first proposition is the abstract channel statement: if the observable dynamics under B factor through a single stochastic map Λ_B, then any contractive divergence can only decrease under that map.

Proposition 3.1 ***(No back-flow under the OMC)*** *If (4) holds and D is contractive under stochastic maps, then for any two first-step instruments A, A',*

$$D\big(\Phi_2^{(A,B)}, \Phi_2^{(A',B)}\big) \le D\big(\Phi_1^{(A)}, \Phi_1^{(A')}\big).$$

Equivalently, the back-flow statistic defined below satisfies $\Delta_{\mathrm{BF}} \le 0$.

The later witness proposition is simply the concrete SGD instantiation of this abstract data-processing statement, using the predictive-distribution divergences from §3.6.

3.3 A back-flow witness of operational memory

We now compare two histories that differ only in the first intervention. The quantity D_1 measures how different their predictive distributions are after the first step alone, while D_2 measures how different they are after the common second intervention B. The back-flow statistic $\Delta_{\mathrm{BF}} := D_2 - D_1$ therefore asks whether the second intervention amplifies or attenuates the discrepancy created at the first step.

Choose two first-step instruments A, A' (in our experiments: same underlying indices, different augmentation) and fix a common second-step instrument B. Let D be a divergence on $\mathcal{P}(\mathcal{Y})$ that obeys data processing

(see §3.6). Define

$$D_1 := D\big(\Phi_1^{(A)}, \Phi_1^{(A')}\big), \tag{5}$$

$$D_2 := D\big(\Phi_2^{(A,B)}, \Phi_2^{(A',B)}\big), \tag{6}$$

$$\Delta_{\mathrm{BF}} := D_2 - D_1. \tag{7}$$

Proposition 3.2 ***(Witness form used in the experiments)*** *If the OMC* (4) *holds and D satisfies data processing, then*

$$\Delta_{\mathrm{BF}} \le 0.$$

Thus, *positive* back-flow, $\Delta_{\mathrm{BF}} > 0$, *falsifies* (4) and witnesses operational memory at the two-step scale. Mechanistically this can arise because (i) history is *carried* across the step by unobserved buffers that modulate the action of B, or (ii) the observable is *insufficient* for how B acts, so that two distinct latent preparations that agree at the observable still evolve differently under B.

3.4 Causal break and a sufficient condition for restoring the OMC

A *causal break* before B is an intervention that cuts a memory channel we intend to test. In our setting, it resets optimizer buffers while keeping parameters fixed. Formally, let

$$\mathcal{B} : \mathcal{S} \rightsquigarrow \mathcal{S}$$

be the kernel that erases buffer state: if $s = (\theta, \upsilon)$, then

$$\mathcal{B}(s, \mathrm{d}\theta' \mathrm{d}\upsilon') = \delta_\theta(\mathrm{d}\theta')\, \delta_0(\mathrm{d}\upsilon').$$

Let

$$\widetilde{\mathcal{K}}_B := \mathcal{K}_B \star \mathcal{B}$$

denote the post-break latent kernel.

We now give a mild condition under which applying the break restores an observable-level channel description.

Definition 2 (Observable sufficiency for B) *Let $\mathcal{Q} \subseteq \mathcal{P}(\mathcal{S})$ be the set of mid-time latent laws reachable after the first step. The observation channel $\mathcal{M}$ is* sufficient for B on $\mathcal{Q}$ *if for any $\pi_1, \pi_1' \in \mathcal{Q}$ with equal observables,*

$$\pi_1 \mathcal{M}^{-1} = \pi_1' \mathcal{M}^{-1},$$

we also have equal post-break observables,

$$\pi_1 \widetilde{\mathcal{K}}_B \mathcal{M}^{-1} = \pi_1' \widetilde{\mathcal{K}}_B \mathcal{M}^{-1}.$$

In words: once the optimizer-memory channel has been cut, the observable must retain enough information to determine how B affects future observables.

Theorem 3.3 ***(Break ⇒ observable channel under sufficiency)*** *If a causal break $\mathcal{B}$ is applied and $\mathcal{M}$ is sufficient for B on $\mathcal{Q}$, then there exists a stochastic map*

$$\Lambda_B : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$$

such that for all first-step instruments I,

$$\Phi_2^{(I,B)} = \Lambda_B[\Phi_1^{(I)}].$$

Consequently, $\Delta_{\mathrm{BF}} \le 0$ for any contractive divergence D.

A sufficient way to guarantee Def. 2 is a *Markov factorization* through the observable: there exists a lifting kernel

$$R : \mathcal{Y} \rightsquigarrow \mathcal{S}$$

(a right-inverse of $\mathcal{M}$ on $\mathcal{Q}$) such that

$$\widetilde{\mathcal{K}}_B = \widetilde{\mathcal{K}}_B \star R \star \mathcal{M} \qquad \text{on } \mathcal{Q}.$$

Then one explicit choice is

$$\Lambda_B = \mathcal{M} \star \widetilde{\mathcal{K}}_B \star R.$$

3.5 Perturbative scaling predictions

A first-order linearization explains the observed trends with momentum, batch overlap, and micro-step depth. Write the k-step update under instrument I as

$$U_I(\theta) \approx \theta - \eta \sum_{t=0}^{k-1} \hat{g}_I(\theta, t),$$

where $\hat{g}_I$ incorporates momentum μ via

$$\hat{g}_I(\cdot, t) \approx (1-\mu) \sum_{i=0}^{t} \mu^i g_I(\cdot).$$

For the two histories A, A',

$$\theta_1^{(A)} - \theta_1^{(A')} \approx -\eta \sum_{t=0}^{k-1} \big(\hat{g}_A - \hat{g}_{A'}\big)(\theta_0, t),$$

and after applying B once more,

$$\theta_2^{(A,B)} - \theta_2^{(A',B)} \approx \Big(I - \eta k\, H_B(\theta_0)\Big)\big(\theta_1^{(A)} - \theta_1^{(A')}\big) + \eta k\, \Xi_{AB},$$

where H_B is a (preconditioned) Jacobian/Hessian of the B-gradient and Ξ_{AB} collects curvature–noise cross-terms.

Two corollaries match our measurements. (i) **Momentum amplification:** $\|\theta_1^{(A)} - \theta_1^{(A')}\|$ scales like $(1-\mu^k)/(1-\mu)$, so Δ_{BF} increases with μ. (ii) **Resonance via overlap:** when B reuses samples from A (overlap ρ), H_B and Ξ_{AB} correlate with $g_A - g_{A'}$, enlarging Δ_{BF}; disjoint B damps it. A causal break zeros momentum at the second step, removing the μ-dependent amplification and driving $\Delta_{\mathrm{BF}} \to 0$ up to $\mathcal{O}(\eta^2)$.

3.6 Divergences and contractivity

We operate on per-example predictive distributions over the probe set and aggregate by averaging the per-example divergence.[2] We require boundedness in $[0, 1]$, sensitivity to multi-class changes, and data processing. *Total variation*

$$\mathrm{TV}(p, q) = \tfrac{1}{2}\|p - q\|_1,$$

and *Hellinger*

$$\mathrm{Hell}(p, q) = \tfrac{1}{2}\,\|\sqrt{p} - \sqrt{q}\|_2$$

are f-divergences and thus contractive under stochastic maps. *Jensen–Shannon*

$$\mathrm{JS}(p, q) = \tfrac{1}{2}\mathrm{KL}(p\|m) + \tfrac{1}{2}\mathrm{KL}(q\|m), \qquad m = \tfrac{1}{2}(p+q),$$

is a symmetrized mixture of KL and inherits the same property. Proposition 3.2 therefore applies to all three.

3.7 Inference at the observable level

Each run produces random variables $D_1, D_2 \in [0, 1]$ and $\Delta_{\mathrm{BF}} \in [-1, 1]$. We treat the operational-Markov inequality as the null hypothesis

$$H_0 : \mathbb{E}[\Delta_{\mathrm{BF}}] \le 0.$$

We report nonparametric bootstrap confidence intervals for the mean and perform TOST (*two one-sided tests*) against a small equivalence margin ε to declare practical nullity. A lower confidence bound strictly above 0 rejects H_0 and certifies operational non-Markovianity for the specified instruments and observable. Collapse toward 0 after a causal break supports the mechanism diagnosis of §3.4.

3.8 What a positive back-flow *means* (and what it does not)

A strictly positive Δ_{BF} says there is *no* single channel Λ_B that maps the measured one-step observable laws to the two-step observable laws for the fixed second intervention B. It therefore certifies an observable-level failure of (4).

It does *not* claim that the fully augmented latent dynamics (θ, buffers) are themselves non-Markov. Rather, it says that at the level of the chosen measurement—predictive distributions on a probe set—the effect of the second intervention cannot be reduced to a single history-independent channel. Our causal-break experiments remove optimizer memory as a carrier and are consistent with the interpretation that, at the two-step scale we probe, optimizer buffers mediate the observed dependence.

[2] Formally, if $x \in \mathcal{P}$ ranges over probe inputs and P_x, Q_x are the class-probability vectors under two histories, we use $\bar{D}(P, Q) := \frac{1}{|\mathcal{P}|}\sum_{x\in\mathcal{P}} D(P_x, Q_x)$. If D obeys data processing pointwise, so does $\bar{D}$.

4 EXPERIMENTS AND RESULTS

We instantiate the two-step $A{\to}B$ protocol from §3 across datasets, model families, and micro-step regimes, and measure the observable-level back-flow

$$\Delta_{\mathrm{BF}} = D_2 - D_1$$

using TV (primary) and JS/H (secondary). The main experimental intervention relative to the formalism of §3 is the *causal break*: either optimizer state is carried from A into B (*no break*) or the optimizer is reinitialized immediately before B (*break*), while keeping hyperparameters fixed.

We evaluate on `CIFAR-100` and `Imagenette`. Inputs are resized to 32×32 (RGB). For Imagenette we use the official validation split as held-out data. Backbones are `SmallCNN`, `ResNet-18` (CIFAR stem), `VGG-11`, `MobileNetV2`, and a `ViT-B/16` configured for 32×32 inputs (2×2 patches). Final classification heads match the dataset's number of classes.

Unless otherwise stated, we use the `early` stage: SGD (LR 0.1, momentum 0.9, weight decay $5{\times}10^{-4}$), a cosine schedule over 3 epochs, and weak CPU-side augmentation (random crop 32 with pad 4 and horizontal flip). A GPU snapshot of the base parameters is cached to enable fast, in-place resets between micro-experiments.

4.1 Protocol, probe observable, and inference

A single *micro-experiment* starts from the cached base parameters, applies k SGD steps on a mini-batch stream from A, and then k steps from B. Unless otherwise specified, we use batch size 256, per-step weight decay 10^{-4}, gradient clipping at 1.0, channels-last memory layout, and *AMP disabled*. BatchNorm running statistics are frozen during micro-steps.

We implement GPU-native transforms: *weak* = random crop (32, reflect pad 4) + horizontal flip (p=0.5); *color* = weak + color jitter (brightness/contrast/saturation 0.4) + random grayscale (p=0.2); *blur* = weak + depthwise Gaussian blur (kernel 3, $\sigma \sim \mathcal{U}[0.1, 2.0]$); and *none* = identity.

For each repeat we draw I_A uniformly without replacement (size 256) from cached training tensors. We draw I_B to achieve a prescribed overlap with I_A,

$$|I_A \cap I_B| \approx \lfloor \texttt{overlap} \times 256 \rfloor,$$

Table 1: Micro-step regimes.

Regime	k	LR	mom	aug_A	$\text{aug}_{A'}$	aug_B	overlap / classes
`standard`	3	0.02	0.90	weak	color	weak	0.5 / `True`
`resonant strong`	6	0.03	0.99	color	blur	weak	1.0 / `True`
`resonant mid`	6	0.03	0.95	color	blur	weak	0.75 / `True`
`orthogonal`	6	0.03	0.99	color	blur	blur	0.0 / `False`
`negative` (ctrl)	1	0.005	0.00	none	none	none	0.0 / `False`

and optionally match the class histogram of I_A (`same_classes=True`); otherwise we draw from the complement. This provides controlled variation in the degree of content resonance between A and B.

Fixing (I_A, I_B) and augmentations $(\text{aug}_A, \text{aug}_{A'}, \text{aug}_B)$, we compare

$$\theta_A \xrightarrow{B} \theta_{AB} \qquad \text{vs.} \qquad \theta_{A'} \xrightarrow{B} \theta_{A'B},$$

where A and A' differ only by augmentation on the same I_A. We compute D_1 and D_2 on a held-out probe and report Δ_{BF}.

In the *no break* condition, B uses the optimizer state carried from A (e.g., momentum buffers). In the *break* condition, we re-initialize the optimizer immediately before B (same LR and momentum), implementing the observable-level causal break of §3.4.

For each (dataset, model) we fix a probe of N=2000 held-out examples (Imagenette: validation; CIFAR-100: test). Given parameters θ, we compute softmax probabilities on the probe to form $P(\theta) \in \mathbb{R}^{N\times C}$. As in §3.6, we aggregate per-example divergences to obtain D_1 and D_2, and report Δ_{BF}; TV is our primary metric, with JS and H used as robustness checks.

We sweep the four substantive regimes and the negative control in Table 1. To contextualize Δ_{BF}, we also log four auxiliary diagnostics once per seed and regime:

1. **Non-commutativity ($AB \neq BA$):** for $k \in \{1, \ldots, 6\}$ we compare endpoints $\theta_{AB}^{(k)}$ and $\theta_{BA}^{(k)}$ via $\text{TV}\big(P(\theta_{AB}^{(k)}), P(\theta_{BA}^{(k)})\big)$ on a 512-example probe subset.

2. **Momentum alignment (no break only):** $\cos\big(\nabla_\theta \ell_B,\ m\big)$ between the B-gradient and the optimizer's momentum buffer just before the first B step.

3. **Penultimate feature drift:** linear CKA on the input to the final linear layer for (A, A') and $(AB, A'B)$ using the same 512-example probe subset.

4. **Function-space trajectory:** PCA of $\{P(\theta_A), P(\theta_{AB}), P(\theta_{A'}), P(\theta_{A'B})\}$, visualized as branch centroids ($A \to AB$ vs. $A' \to A'B$).

For each (dataset, model, regime, break flag) and each seed $s \in \{0, 1, 2, 3, 4\}$, we run R=128 independent repeats (fresh (I_A, I_B) per repeat). To control runtime while maintaining tight confidence intervals, after at least 64 repeats we check every 32 repeats whether the nominal half-width of the 95% CI for $\overline{\Delta_{\text{BF}}}$ (normal approximation) is $\leq 2\times10^{-4}$ and stop early if so. We aggregate per-configuration means with 2000-sample nonparametric bootstrap CIs and additionally perform TOST with equivalence margin ε=10^{-3}. All micro-experiments use SGD with gradient clipping at 1.0; a NaN guard and fallback LR scaling ($\times$0.5 on retry) handle rare numerical issues.

4.2 Global prevalence and robustness across metrics

Across 50 unique {*dataset, model, regime, base_stage*} setups and two optimizer conditions (*no break* and *break*), we obtain 100 TV groups. Significant effects are widespread and confidence intervals are typically narrow. Under *no break*, 35 groups are significantly positive and 12 significantly negative, with only 3 non-significant; under *break*, 22 groups are significantly positive and 25 significantly negative, again with only 3 non-significant. Every unique setup is significant in at least one condition, and 44/50 are significant in both. Agreement across metrics is also strong: TV$\wedge$Hellinger is significant in 93/100 groups and TV$\wedge$JS in 80/100; see Fig. 1.

Observable back-flow is therefore broad across architectures and remains visible under multiple contractive distances.

4.3 Causal-break effect and sign reversals

Introducing a *causal break* systematically changes both the sign and magnitude of Δ_{BF}. In particular, 14/50 setups (28%) flip sign between the *no break* and *break* conditions (Fig. 2), with the largest absolute changes concentrated in high-momentum, high-overlap regimes

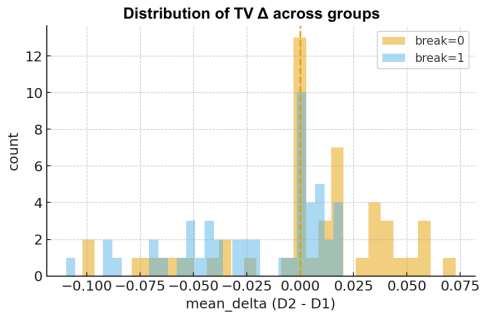

Figure 1: **Distribution of TV mean Δ_{BF} across all setups.** Histogram for *no break* and *break* conditions. The distribution shifts toward attenuation under a causal break.

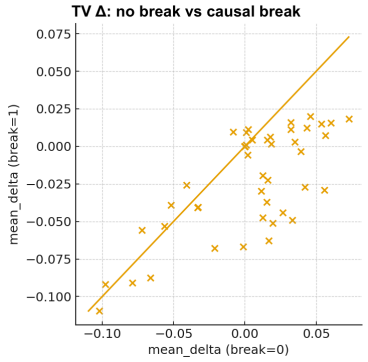

Figure 2: **TV mean effect: *no break* vs. *break*.** Each point is one {*dataset, model, regime, base_stage*}. Many points lie below the diagonal, and 28% cross quadrants (sign flips), indicating a strong effect of optimizer carryover on amplification.

(Table 2). This is consistent with optimizer carryover being an important causal mediator of amplification.

4.4 Regime-level trends

Averaging TV across datasets and architectures reveals clear regime-level structure (Fig. 3). Without a break, `standard`, `orthogonal`, and `resonant strong` tend to amplify ($\Delta_{\mathrm{BF}} > 0$); with a break they shift toward attenuation ($\Delta_{\mathrm{BF}} < 0$). The `negative` control remains near zero in both conditions.

Numerically, means across all dataset×model configurations are: `standard` $+0.0127 \to -0.00729$, `orthogonal` $+0.0103 \to -0.0206$, `resonant strong` $+0.00734 \to -0.0310$, `resonant mid` $-0.0177 \to -0.0448$, and `negative` $\approx 6{\times}10^{-4} \to 5{\times}10^{-4}$ (medians $\approx 10^{-7}$). These shifts match the perturbative picture of §3.5: higher momentum and stronger A–B overlap tend to amplify the first-step discrepancy,

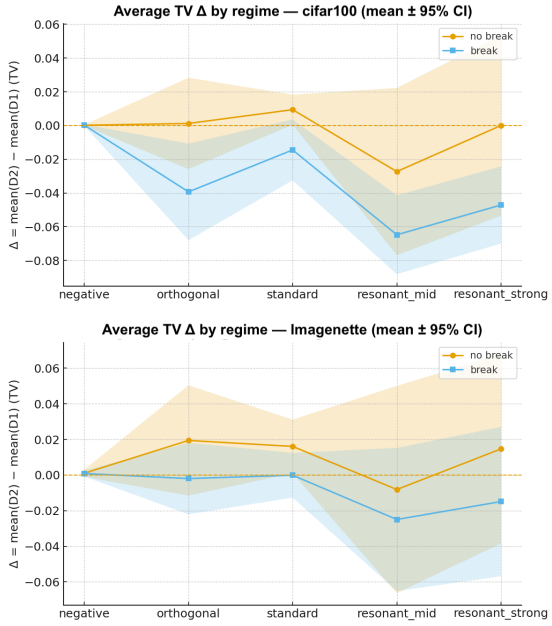

Figure 3: **Average TV effect by regime and optimizer condition.** Points are means across datasets and models; ribbons show bootstrap 95% CIs. High-momentum/high-overlap regimes tend to amplify without a break and attenuate with a break.

whereas a causal break suppresses that carryover.

4.5 Diagnostics linking back-flow to mechanism

The auxiliary diagnostics support the same mechanism-level interpretation. First, non-commutativity grows with micro-step depth: for the configurations shown in Fig. 4, the curve $\mathrm{TV}(AB, BA)$ increases with the number of micro-steps k, indicating order sensitivity even over short horizons. The increase is typically steeper without a break and flatter with a break, consistent with optimizer carryover contributing to non-commutativity. Across configurations, the slope of $\mathrm{TV}(AB, BA)$ versus k under *no break* correlates with Δ: Pearson $r = 0.184$ ($p = 1.38 \times 10^{-3}$) and Spearman $\rho = 0.329$ ($p = 5.45 \times 10^{-9}$); see Fig. 5 (left).

Second, the pre-B momentum buffer aligns with the upcoming B-gradient. Without a break, the distribution of $\cos(\nabla_B, m)$ concentrates above zero (Fig. 6), directly indicating optimizer carryover that can amplify the effect ($\Delta_{\mathrm{BF}} > 0$). Across no-break configurations, Δ scales with mean alignment: Pearson $r = 0.409$ ($p = 4.09 \times 10^{-11}$) and Spearman $\rho = 0.454$ ($p = 1.37 \times 10^{-13}$); see Fig. 5 (right). After a break, this channel is removed by construction and the amplification correspondingly weakens.

Table 2: **Representative cases across both datasets (TV).** Rows 1–8 illustrate sign flips; the bottom rows provide one strong attenuation baseline per dataset.

Dataset	Model	Regime	Stage	Δ_{no}	Δ_{br}	$\Delta_{\text{br}}-\Delta_{\text{no}}$
CIFAR-100	ViT-B/16	resonant_strong	early	0.0557	−0.0291	−0.0847
CIFAR-100	VGG-11	resonant_strong	early	0.0331	−0.0492	−0.0823
CIFAR-100	ViT-B/16	orthogonal	early	0.0168	−0.0629	−0.0797
CIFAR-100	ViT-B/16	resonant_mid	early	0.0266	−0.0442	−0.0708
CIFAR-100	MobileNetV2	orthogonal	early	0.0193	−0.0512	−0.0706
Imagenette	ViT-B/16	orthogonal	early	0.0393	−0.0032	−0.0426
Imagenette	ResNet-18	standard	early	0.0161	−0.0225	−0.0386
Imagenette	VGG-11	standard	early	0.0021	−0.0059	−0.0080
CIFAR-100	ResNet-18	resonant_mid	early	−0.1022	−0.1096	−0.0075
Imagenette	ResNet-18	resonant_mid	early	−0.0978	−0.0918	+0.0059

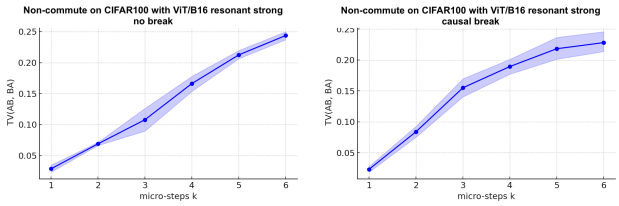

Figure 4: **Non-commute curves.** $\mathrm{TV}(AB, BA)$ vs. micro-steps k for `CIFAR-100/ViT-B/16/resonant strong`. Left: *no break*. Right: *break*.

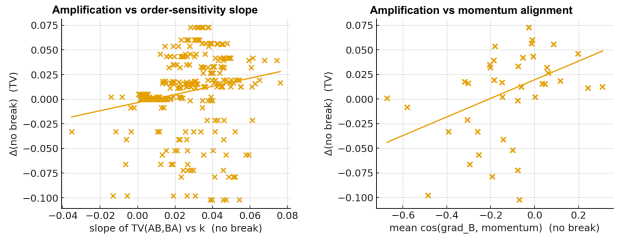

Figure 5: **Amplification vs. order sensitivity (left).** Configurations with a steeper $\mathrm{TV}(AB, BA)$ slope (no break) tend to exhibit larger Δ. **Amplification vs. momentum alignment (right).** Greater pre-B alignment predicts larger Δ. Least-squares fits are shown.

Third, the feature-space and function-space diagnostics show the same qualitative pattern. Because absolute CKA values are uniformly high, we visualize the signed change

$$\Delta_{\text{CKA}} = \mathrm{CKA}(AB, A'B) - \mathrm{CKA}(A, A')$$

rather than the raw CKA levels. Positive values indicate that the common second intervention B increases feature similarity between the two histories, whereas negative values indicate feature-space divergence. In the showcased ViT-B/16 `resonant strong` cases, the break shifts Δ_{CKA} upward in both datasets and reverses its sign on CIFAR-100 (Fig. 7), consistent with

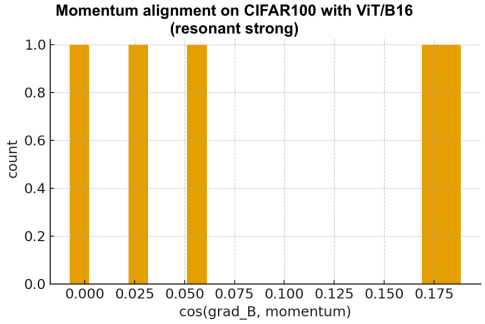

Figure 6: **Momentum alignment (no break).** Histogram of $\cos(\nabla_B, m)$; mass above zero indicates optimizer carryover aligned with the upcoming B update.

reduced branch divergence once optimizer-state carryover is removed. Similarly, function-space PCA trajectories show longer $A \to AB$ displacements and wider branch separation without a break, but shorter and more aligned paths with a break (Fig. 8).

Taken together, these diagnostics link the statistical witness Δ_{BF} to concrete optimization mechanisms:order sensitivity, momentum-buffer alignment, and branch separation in feature and function space.

4.6 Controls, falsification checks, and scope limitations

The `negative` regime behaves as a control, with means $\approx (5\text{–}6) \times 10^{-4}$ and medians $\approx 10^{-7}$ under both conditions. We also observe multiple narrow-CI cases (95% CI half-width $\leq 2 \times 10^{-4}$ in nine groups per condition), supporting the stability of the estimates. TOST with margin $\varepsilon = 10^{-3}$ classifies many break-condition effects as practically null, while rejecting nullity for many no-break amplification cases. Detailed per-setup statistics are deferred to the appendix.

As falsification checks, placebo runs with $A = A'$

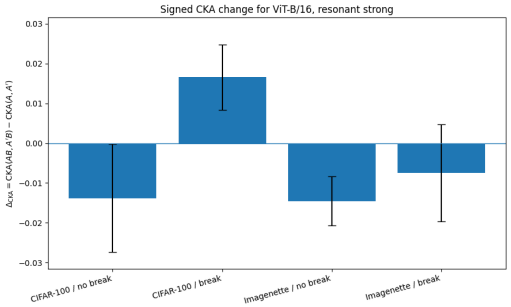

Figure 7: **Signed change in penultimate-layer CKA under the common second intervention.** We plot $\Delta_{\mathrm{CKA}} = \mathrm{CKA}(AB, A'B) - \mathrm{CKA}(A, A')$ for ViT-B/16 in the `resonant strong` regime on CIFAR-100 and Imagenette, with and without a causal break. Positive values indicate that the common second intervention B increases feature similarity between the two histories (realignment), while negative values indicate feature-space divergence. The break shifts Δ_{CKA} upward in both datasets and reverses its sign on CIFAR-100, consistent with optimizer-state carryover promoting branch divergence in feature space. Error bars show 95% confidence intervals across seeds.

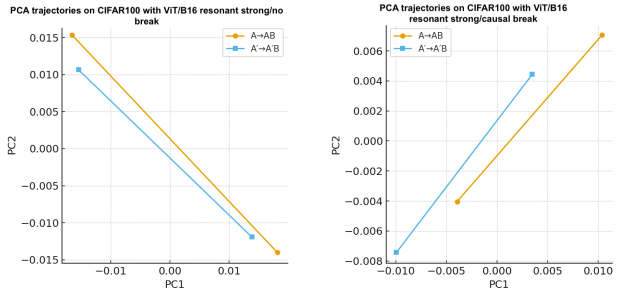

Figure 8: **Function-space trajectories.** Displacements and branch separation are larger without a break (left); a break shortens and aligns the paths (right).

(identical augmentations) yield Δ statistically indistinguishable from zero, with 95% CIs spanning zero. Likewise, no-break runs with $\mu = 0$ collapse toward the break baseline, consistent with optimizer buffers mediating the observed back-flow.

Our study focuses on short two-step interventions, image-classification benchmarks, and standard vision backbones. Accordingly, we study the existence and controllability of observable memory rather than downstream performance benefits. The framework itself is not restricted to SGD with momentum or to vision: it extends to any setting in which controlled interventions and a probe observable can be defined. We emphasize the two-step case because it gives the clearest interpretation of the witness and the strongest statistical power for a fixed experimental budget. Extensions to longer horizons, adaptive optimizers, non-vision domains, and a fuller probe-size sensitivity analysis remain important directions for future work.

5 CONCLUSIONS

The memory we detect is *intentional in mechanism*—optimizer state is carried across steps by design—but *unintended in consequence*: it induces operational non-Markovianity, whereby the effect of a fixed second intervention B depends on the immediate history A. Whether such memory is helpful or harmful depends on the alignment between training order and target generalization. Our witness Δ_{BF} offers a direct diagnostic: it quantifies when optimizer memory amplifies prior updates (as in "resonant" schedules), and when it should be collapsed with a causal break before switching regimes. We identify optimizer state as a *causal driver* of observable back-flow. It systematically amplifies the influence of A on subsequent B, induces order sensitivity ($AB \neq BA$), and can be neutralized by a simple reset of optimizer buffers. This establishes a practical intervention that links mechanism and measurement. The phenomenon is robust: we observe consistent effects across datasets (CIFAR-100, Imagenette), architectures (SmallCNN, ResNet-18, VGG-11, MobileNetV2, ViT-B/16), divergences (TV, JS, H), and training regimes. These results demonstrate that optimizer-induced memory is a pervasive and measurable feature of modern training dynamics, and that back-flow provides a principled, operational handle for studying and controlling it.

Acknowledgements

This work has been partially supported by the ARGUS EU project (Grant Agreement No. 101132308), funded by the European Union.

References

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine

Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022.

Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness. *arXiv preprint arXiv:2003.10973*, 2020.

Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

Mathieu Even. Stochastic gradient descent under markovian sampling schemes. In *International Conference on Machine Learning*, pages 9412–9439. PMLR, 2023.

Gabriel Goh. Why momentum really works. *Distill*, 2(4):e6, 2017.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1–2):49–84, October 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01440-w. URL `http://dx.doi.org/10.1007/s10107-019-01440-w`.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

Qianxiao Li, Cheng Tai, et al. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Simon Milz and Kavan Modi. Quantum stochastic processes and quantum non-markovian phenomena. *PRX Quantum*, 2(3):030201, 2021.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

Felix A Pollock, César Rodríguez-Rosario, Thomas Frauenheim, Mauro Paternostro, and Kavan Modi. Operational markov condition for quantum processes. *Physical review letters*, 120(4):040405, 2018a.

Felix A Pollock, César Rodríguez-Rosario, Thomas Frauenheim, Mauro Paternostro, and Kavan Modi. Non-markovian quantum processes: Complete framework and efficient characterization. *Physical Review A*, 97(1):012127, 2018b.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5): 1–17, 1964.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. *Advances in neural information processing systems*, 29, 2016.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. pmlr, 2013.

Gregory AL White, Felix A Pollock, Lloyd CL Hollenberg, Kavan Modi, and Charles D Hill. Non-markovian quantum process tomography. *PRX Quantum*, 3(2):020344, 2022.

Jiayuan Ye, Anastasia Borovykh, Soufiane Hayou, and Reza Shokri. Leave-one-out distinguishability in machine learning. *arXiv preprint arXiv:2309.17310*, 2023.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDIX

PROOFS

Proposition 3.1 Let $\mathcal{Z}$ be a (standard Borel) observation space for the probe observable, and let $\mathcal{P}(\mathcal{Z})$ denote the set of probability measures on $\mathcal{Z}$. Fix a family of first-step instruments $\mathcal{I}$ and a common second-step instrument B. For each $I \in \mathcal{I}$, let $P_I \in \mathcal{P}(\mathcal{Z})$ be the (pre-B) predictive distribution on the probe, and let $Q_I \in \mathcal{P}(\mathcal{Z})$ be the (post-B) predictive distribution. Assume the two-time process is *operationally Markov at the observable for* B, namely: there exists a Markov kernel (classical channel) $\Lambda_B : \mathcal{P}(\mathcal{Z}) \to \mathcal{P}(\mathcal{Z})$ such that

$$Q_I = \Lambda_B(P_I) \qquad \text{for all } I \in \mathcal{I}.$$

Let $D : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$ be any divergence that is *contractive under classical channels*, i.e., $D(\Lambda p, \Lambda q) \leq D(p, q)$ for all probability measures p, q and Markov kernels Λ. Define the back-flow of distinguishability for any pair $I, I' \in \mathcal{I}$ by

$$\Delta^{(I,B)}(I, I') \;\coloneqq\; D(Q_I, Q_{I'}) - D(P_I, P_{I'})\,.$$

Then $\Delta^{(I,B)}(I, I') \leq 0$ for all $I, I' \in \mathcal{I}$.

Proof By operational Markovianity at the observable for B, there exists a *fixed* Markov kernel Λ_B such that $Q_I = \Lambda_B(P_I)$ and $Q_{I'} = \Lambda_B(P_{I'})$ for all instruments I, I'. By contractivity of D under classical channels (data-processing inequality),

$$D(Q_I, Q_{I'}) \;=\; D\big(\Lambda_B(P_I),\, \Lambda_B(P_{I'})\big) \;\leq\; D(P_I, P_{I'})\,.$$

Rearranging gives $D(Q_I, Q_{I'}) - D(P_I, P_{I'}) \leq 0$, i.e., $\Delta^{(I,B)}(I, I') \leq 0$. Since I, I' were arbitrary, the claim holds for all pairs.

Remarks. (i) In the discrete case, Λ_B is a stochastic matrix acting on probability vectors, and the proof reduces to $D(\Lambda_B p, \Lambda_B q) \le D(p, q)$. (ii) The assumption on D holds for standard contractive divergences (e.g., total variation, Hellinger, Jensen–Shannon, and more generally Csiszár f-divergences under mild regularity). (iii) Equality $\Delta^{(I,B)} = 0$ occurs, e.g., when Λ_B is *sufficient* for the pair $(P_I, P_{I'})$ in the sense that it preserves their statistical distance as measured by D.

Proof of Proposition 3.2

Proposition 3.2, Let D be any divergence on probability measures that satisfies data processing: $D(\Lambda p, \Lambda q) \le D(p, q)$ for all Markov kernels Λ and all p, q. Assume the operational Markov condition (4) for the second-step observable B, i.e., there exists a (fixed) classical channel Λ_B such that

$$\Phi_2^{(I,B)} = \Lambda_B \, \Phi_1^{(I)} \qquad \text{for every first-step instrument } I. \tag{8}$$

For any two first-step instruments A, A', define the back-flow of distinguishability

$$\Delta^{(I,B)} \;:=\; D\big(\Phi_2^{(A,B)}, \Phi_2^{(A',B)}\big) \;-\; D\big(\Phi_1^{(A)}, \Phi_1^{(A')}\big).$$

Then $\Delta^{(I,B)} \le 0$.

Proof Fix any two first-step instruments A, A'. Set $P = \Phi_1^{(A)}$ and $Q = \Phi_1^{(A')}$. By (8),

$$\Phi_2^{(A,B)} = \Lambda_B P \qquad \text{and} \qquad \Phi_2^{(A',B)} = \Lambda_B Q.$$

Applying data processing for D with the channel Λ_B yields

$$\begin{aligned} D\big(\Phi_2^{(A,B)}, \Phi_2^{(A',B)}\big) &= D(\Lambda_B P, \Lambda_B Q) \\ &\le D(P, Q) \\ &= D\big(\Phi_1^{(A)}, \Phi_1^{(A')}\big). \end{aligned}$$

Rearranging gives $\Delta^{(I,B)} \le 0$, as claimed.

Remarks. (1) The argument is agnostic to the sample space (finite or general Borel) and to the particular choice of D, provided D obeys data processing (e.g., total variation, Hellinger, Jensen–Shannon, and Csiszár f-divergences under standard regularity). (2) If you use the metric $\sqrt{\mathrm{JS}}$, data processing holds and the same conclusion follows verbatim. (3) If you later introduce an *approximate* OMC, $\|\Phi_2^{(I,B)} - \Lambda_B \Phi_1^{(I)}\|_D \le \eta_I$ for all I and a triangle-inequality divergence $\|\cdot\|_D$, then

$$\Delta^{(I,B)} \le \eta_A + \eta_{A'},$$

by a two-sided triangle inequality combined with contractivity of D; this yields a stable, quantitative relaxation of Proposition 3.2.

Proof of Theorem 3.3 Let $\mathcal{Q}, \mathcal{O}$ be standard Borel spaces. Let $\mathcal{O} : \mathcal{Q} \rightsquigarrow \mathcal{O}$ be the probe observable (a Markov kernel) and let $\mathcal{K}_B : \mathcal{Q} \rightsquigarrow \mathcal{Q}$ be the second-step dynamics for instrument B. For a first-step instrument I, write $\pi_1^{(I)} \in \mathcal{P}(\mathcal{Q})$ for the latent state after step 1 and $\Phi_1^{(I)} := \pi_1^{(I)} \mathcal{O} \in \mathcal{P}(\mathcal{O})$ for the corresponding predictive (pre-B) observable law. Assume:

Causal break $\mathcal{B}$ is applied between step 1 and B, so that the post-break latent state depends on I only via $\Phi_1^{(I)}$, i.e., there exists a Markov kernel $\mathcal{R} : \mathcal{O} \rightsquigarrow \mathcal{Q}$ with

$$\widetilde{\pi}_1^{(I)} \;=\; \Phi_1^{(I)} \, \mathcal{R} \qquad \text{for all } I.$$

Sufficiency of $\mathcal{O}$ for B on $\mathcal{Q}$: there exists a Markov kernel $\Lambda_B : \mathcal{O} \rightsquigarrow \mathcal{O}$ such that

$$\mu \, \mathcal{R} \, \mathcal{K}_B \, \mathcal{O} \;=\; \mu \, \Lambda_B \qquad \text{for all } \mu \in \mathcal{P}(\mathcal{O}). \tag{9}$$

(Equivalently, for all $\pi \in \mathcal{P}(\mathcal{Q})$ one has $\pi \mathcal{K}_B \mathcal{O} = (\pi \mathcal{O}) \Lambda_B$.)

Then for all first-step instruments I,

$$\Phi_2^{(I,B)} \;=\; \Lambda_B \, [\, \Phi_1^{(I)} \,].$$

Consequently, for any divergence D that is contractive under classical channels (data processing), the back-flow satisfies $\Delta^{(I,B)} \le 0$.

Proof By the causal break, the latent state just before applying B is $\widetilde{\pi}_1^{(I)} = \Phi_1^{(I)} \mathcal{R}$. Pushing it forward through B and then observing via $\mathcal{O}$ gives

$$\Phi_2^{(I,B)} \;=\; \widetilde{\pi}_1^{(I)} \, \mathcal{K}_B \, \mathcal{O} \;=\; \big(\Phi_1^{(I)} \, \mathcal{R}\big) \, \mathcal{K}_B \, \mathcal{O} \;=\; \Phi_1^{(I)} \, (\mathcal{R} \, \mathcal{K}_B \, \mathcal{O}).$$

By sufficiency (9), the composite $\mathcal{R} \, \mathcal{K}_B \, \mathcal{O}$ factors through a *fixed* observable channel $\Lambda_B : \mathcal{O} \rightsquigarrow \mathcal{O}$ independent of I, hence

$$\Phi_2^{(I,B)} \;=\; \Phi_1^{(I)} \, \Lambda_B \;=\; \Lambda_B [\, \Phi_1^{(I)} \,].$$

This proves the first claim.

For the back-flow bound, let A, A' be any two first-step instruments and set $P = \Phi_1^{(A)}$, $Q = \Phi_1^{(A')}$. Then

$$\Phi_2^{(A,B)} = \Lambda_B P, \qquad \Phi_2^{(A',B)} = \Lambda_B Q.$$

By data processing (contractivity) of D under classical channels,

$$\begin{aligned} D\big(\Phi_2^{(A,B)}, \Phi_2^{(A',B)}\big) &= D(\Lambda_B P, \Lambda_B Q) \\ &\le D(P, Q) \\ &= D\big(\Phi_1^{(A)}, \Phi_1^{(A')}\big). \end{aligned}$$

Rearranging yields $\Delta^{(I,B)} \le 0$.

Table 3: Core notation used in the main text.

Symbol	Meaning
$\mathcal{S}$	Latent training-state space (parameters + optimizer buffers).
$\mathcal{Y}$	Observable space (probe-set predictive outputs).
$\mathcal{P}(\mathcal{S}), \mathcal{P}(\mathcal{Y})$	Probability measures on $\mathcal{S}$ and $\mathcal{Y}$.
π_0	Initial latent-state distribution.
I	Training instrument: mini-batch, augmentation, and optimizer micro-parameters for k micro-steps.
A, A'	Alternative first-step instruments.
B	Common second-step instrument.
$\mathcal{K}_I$	Latent stochastic kernel induced by I.
$\pi_1^{(I)}$	One-step latent law after I.
$\pi_2^{(I_0,I_1)}$	Two-step latent law after I_0 then I_1.
$\mathcal{M}$	Observation channel from latent states to probe outputs.
$\Phi_1^{(I)}$	One-step observable law.
$\Phi_2^{(I_0,I_1)}$	Two-step observable law.
$\mathsf{T}_{2:0}$	Two-time training map (process tensor).
Λ_B	Observable-level stochastic map for fixed second step B.
D	Divergence on observable laws (TV, JS, or Hellinger).
D_1	One-step divergence between A and A'.
D_2	Two-step divergence between (A, B) and (A', B).
Δ_{BF}	Back-flow statistic, $\Delta_{\mathrm{BF}} = D_2 - D_1$.
$\mathcal{B}$	Causal-break kernel resetting optimizer buffers.
$\widetilde{\mathcal{K}}_B$	Post-break second-step kernel, $\mathcal{K}_B \star \mathcal{B}$.
$\mathcal{Q}$	Reachable mid-time latent laws.
R	Lifting kernel in the Markov-factorization condition.

Table 4: Translation from process-tensor language to ordinary ML language.

Formal term	ML interpretation in this paper
Process tensor / multi-time process	Multi-time training map from short-horizon interventions to probe-set predictive distributions.
Instrument	Controlled training intervention: mini-batch, augmentation, and optimizer micro-parameters over a short horizon.
Latent state	Full training state: parameters plus optimizer buffers.
Observable	Model predictions on a fixed probe set.
Observation channel	Map from latent training state to probe-set predictive distribution.
Operational Markov condition	For fixed B, the two-step observable law is determined by the one-step observable law via a single map Λ_B.
Back-flow of distinguishability	Increase after the common second intervention: $\Delta_{\mathrm{BF}} = D_2 - D_1 > 0$.
Causal break	Reset of optimizer state before the second intervention, cutting optimizer-memory carryover.
Observable sufficiency	The observable retains enough information to determine the post-break effect of B.
Markov factorization through the observable	A lifting from probe outputs to latent states makes post-break dynamics depend on history only through the observable.

Notes on assumptions.

- The *causal break* is encoded by $\mathcal{R} : \mathcal{O} \rightsquigarrow \mathcal{Q}$; it reprepares the latent state using only the observed law from step 1, thus erasing any dependence on the pre-history beyond $\Phi_1^{(I)}$.
- The *sufficiency* condition is the standard Blackwell–Sherman–Stein notion: the pair $(\mathcal{K}_B, \mathcal{O})$ *garbles* through the statistic $\mathcal{O}$, i.e. there exists a channel Λ_B on observables such that for all priors π on $\mathcal{Q}$, $(\pi\mathcal{O})\Lambda_B = \pi\mathcal{K}_B\mathcal{O}$. This yields (9) globally on $\mathcal{P}(\mathcal{O})$ (no extension argument is needed).
- In discrete settings, $\mathcal{R}, \mathcal{K}_B, \mathcal{O}, \Lambda_B$ are stochastic matrices and the proof reduces to associativity of matrix multiplication.

Notation and terminology tables

EXTENDED RELATED WORK

Analyses of stochastic optimization often treat updates as (time-homogeneous) Markov dynamics once one augments the state with algorithmic buffers and assumes i.i.d. sampling. This viewpoint underpins tutorials and SDE-based models of SGD: it enables mixing/ergodicity arguments and stationary-distribution pictures near minima [Bottou et al., 2018, Mandt et al., 2017]. Sampling *without* replacement (random reshuffling) couples consecutive steps through the epoch permutation and changes both analysis and practice; multiple works prove or quantify RR's advantages and make its temporal dependence explicit. These effects can be re-cast as Markov only by further augmenting the state (e.g., with a permutation pointer), underscoring that order is a first-class control knob in modern training [Shamir, 2016, Rajput et al., 2020, Gürbüzbalaban et al., 2019]. Momentum and adaptive methods explicitly accumulate history. Classic and modern accounts (heavy-ball, [Nesterov, 1983]; [Sutskever et al., 2013]; Goh [2017]'s Distill note) document their empirical and dynamical impact; stochastic modified-equation analyses encode them as SDEs in an *enlarged* state. Our *causal break* experiment operationalizes this intuition: if buffers mediate history, resetting them should null out back-flow—which is exactly what we observe [Polyak, 1964, Li et al., 2017].

A separate line treats the *data source* as Markov and tracks the role of mixing time in SGD's convergence—formally moving dependence from the algorithm to the sampler. This is adjacent to, but distinct from, our aim: we test for history-dependence in the *training process* as exposed by output distributions under controlled interventions [Even, 2023, Doan et al., 2020, Dorfman and Levy, 2022]. Empirical work finds that gradient noise during deep learning is heavy-tailed and anisotropic, challenging Gaussian, white-noise assumptions common in SDE models. While not phrased in "Markov vs. non-Markov" terms, these observa-

tions motivate measurement-first diagnostics that do not hinge on specific noise models [Simsekli et al., 2019, Zhu et al., 2018]. Several diagnostics quantify how predictions or representations change across training. LOOD-type measures track output changes induced by adding/removing examples (aimed at privacy/memorization) [Ye et al., 2023, Carlini et al., 2019, 2021, Koh and Liang, 2017, Pruthi et al., 2020, Yeh et al., 2018], while representation-similarity metrics (e.g., CKA) probe internal drift [Kornblith et al., 2019]. Recent work questions their reliability and sensitivity [Davari et al., 2022]. In contrast, our witness is *non-parametric, intervention-defined*, and explicitly tied to an operational Markov condition.

ADDITIONAL TABLES

For each {dataset, model, regime, stage, break} configuration we report the mean $\overline{\Delta}$, a nonparametric 95% bootstrap CI, the number of repeats n (pooled across seeds), the CI half-width, and TOST results for $\varepsilon = 10^{-3}$; see Table 5 and Table 6. A compact control summary for the `negative` regime (means and medians under both optimizer conditions) appears in Table 8.

We complement bootstrap CIs with Benjamini–Hochberg (BH) false-discovery-rate control applied to two-sided tests of $H_0 : \mathbb{E}[\Delta] = 0$ per configuration. At FDR 5%, TV: 94/100, Hellinger: 93/100, and JS: 81/100 groups remain significant; conclusions are unchanged. Per-setup tables report raw p and BH q-values alongside 95% bootstrap CIs.

QUANTITATIVE DOSE–RESPONSE (momentum, fixed aug_B).

We model the predicted amplification due to momentum as $A_\mu = \frac{1-\mu^k}{1-\mu}$ and regress the observed back–flow as $\Delta = \alpha + \beta A_\mu + \gamma\rho + \varepsilon$, where ρ is the A–B batch overlap. Using TV under *no break* and configurations with aug_B = `weak` (`standard`: k=3, μ=0.90, ρ=0.5; `resonant_mid`: k=6, μ=0.95, ρ=0.75; `resonant_strong`: k=6, μ=0.99, ρ=1.0), we obtain $n = 30$ dataset×model points with $R^2 = 0.077$, $\beta = -0.0273\,(0.0184)$, $p = 1.38\times10^{-1}$ and $\gamma = 0.1607\,(0.1235)$, $p = 1.93\times10^{-1}$. Because k and ρ vary across regimes, we also perform a *within-pair* test at fixed k=6 (same dataset and model), comparing `resonant_strong` to `resonant_mid`: all 10/10 pairs exhibit an increase (strong > mid), with mean lift 0.0251 and 95% CI $[0.0208, 0.0288]$ (paired t-test $p = 8.14 \times 10^{-7}$). Numerically, for k=6 the amplification factor increases from $A_\mu(\mu{=}0.95) \approx 5.29$ to $A_\mu(\mu{=}0.99) \approx 5.90$; the joint increase in A_μ and ρ yields a robust *dose–response* in Δ.

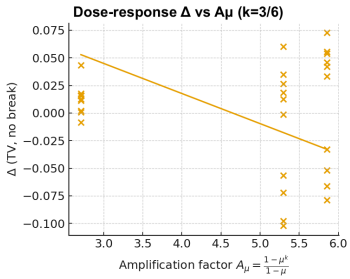

Figure 9: **Dose–response:** TV Δ vs $A_\mu = (1-\mu^k)/(1-\mu)$ under *no break* for regimes with aug_B = `weak`. Line shows OLS fit at mean overlap. A paired comparison at fixed k=6 (not shown) finds strong>mid in 10/10 dataset×model pairs (mean lift 0.0251, 95% CI $[0.0208, 0.0288]$).

Table 5: **Per-group significance by metric on CIFAR-100.** A checkmark means the 95% CI for Δ_{BF} excludes 0. Break: `no` vs. `br`.

Dataset	Model	Regime	Break	TV	JS	Hell
cifar100	mobilenetv2	N	br	✓	✓	✓
cifar100	mobilenetv2	N	no	✓	✓	✓
cifar100	mobilenetv2	O	br	✓	✓	✓
cifar100	mobilenetv2	O	no	✓	✓	✓
cifar100	mobilenetv2	RM	br	✓	✓	✓
cifar100	mobilenetv2	RM	no	✓	✓	✓
cifar100	mobilenetv2	RS	br	✓	✓	✓
cifar100	mobilenetv2	RS	no	✓	✓	✓
cifar100	mobilenetv2	S	br	✓	✓	✓
cifar100	mobilenetv2	S	no	✓	✓	✓
cifar100	resnet18	N	br	✓	✓	✓
cifar100	resnet18	N	no	✓	✓	✓
cifar100	resnet18	O	br	✓	✓	✓
cifar100	resnet18	O	no	✓	✓	✓
cifar100	resnet18	RM	br	✓	✓	✓
cifar100	resnet18	RM	no	✓	✓	✓
cifar100	resnet18	RS	br	✓	✓	✓
cifar100	resnet18	RS	no	✓	✓	✓
cifar100	resnet18	S	br	–	✓	–
cifar100	resnet18	S	no	✓	✓	✓
cifar100	smallcnn	N	br	✓	–	✓
cifar100	smallcnn	N	no	✓	–	✓
cifar100	smallcnn	O	br	–	–	–
cifar100	smallcnn	O	no	✓	✓	✓
cifar100	smallcnn	RM	br	✓	✓	✓
cifar100	smallcnn	RM	no	✓	✓	✓
cifar100	smallcnn	RS	br	✓	✓	✓
cifar100	smallcnn	RS	no	✓	✓	✓
cifar100	smallcnn	S	br	✓	✓	✓
cifar100	smallcnn	S	no	✓	✓	✓
cifar100	vgg11	N	br	✓	✓	✓
cifar100	vgg11	N	no	✓	–	✓
cifar100	vgg11	O	br	✓	✓	✓
cifar100	vgg11	O	no	✓	✓	✓
cifar100	vgg11	RM	br	✓	✓	✓
cifar100	vgg11	RM	no	–	–	–
cifar100	vgg11	RS	br	✓	✓	✓
cifar100	vgg11	RS	no	✓	✓	✓
cifar100	vgg11	S	br	✓	✓	✓
cifar100	vgg11	S	no	✓	✓	✓
cifar100	vit-b-16	N	br	✓	–	✓
cifar100	vit-b-16	N	no	✓	–	✓
cifar100	vit-b-16	O	br	✓	✓	✓
cifar100	vit-b-16	O	no	✓	✓	✓
cifar100	vit-b-16	RM	br	✓	✓	✓
cifar100	vit-b-16	RM	no	✓	✓	✓
cifar100	vit-b-16	RS	br	✓	✓	✓
cifar100	vit-b-16	RS	no	✓	✓	✓
cifar100	vit-b-16	S	br	✓	✓	✓
cifar100	vit-b-16	S	no	✓	✓	✓

Table 6: **Per-group significance by metric on Imagenette.** A checkmark means the 95% CI for Δ_{BF} excludes 0. Break: `no` vs. `br`.

Dataset	Model	Regime	Break	TV	JS	Hell
imagenette	mobilenetv2	N	br	✓	✓	✓
imagenette	mobilenetv2	N	no	✓	✓	✓
imagenette	mobilenetv2	O	br	✓	✓	✓
imagenette	mobilenetv2	O	no	✓	✓	✓
imagenette	mobilenetv2	RM	br	✓	✓	✓
imagenette	mobilenetv2	RM	no	✓	✓	✓
imagenette	mobilenetv2	RS	br	✓	✓	✓
imagenette	mobilenetv2	RS	no	✓	✓	✓
imagenette	mobilenetv2	S	br	✓	✓	✓
imagenette	mobilenetv2	S	no	✓	✓	✓
imagenette	resnet18	N	br	✓	–	✓
imagenette	resnet18	N	no	✓	–	✓
imagenette	resnet18	O	br	✓	✓	✓
imagenette	resnet18	O	no	✓	✓	✓
imagenette	resnet18	RM	br	✓	✓	✓
imagenette	resnet18	RM	no	✓	✓	✓
imagenette	resnet18	RS	br	✓	✓	✓
imagenette	resnet18	RS	no	✓	✓	✓
imagenette	resnet18	S	br	✓	✓	✓
imagenette	resnet18	S	no	✓	✓	✓
imagenette	smallcnn	N	br	✓	–	✓
imagenette	smallcnn	N	no	✓	–	✓
imagenette	smallcnn	O	br	✓	✓	✓
imagenette	smallcnn	O	no	–	–	–
imagenette	smallcnn	RM	br	✓	✓	✓
imagenette	smallcnn	RM	no	✓	✓	✓
imagenette	smallcnn	RS	br	✓	✓	✓
imagenette	smallcnn	RS	no	✓	✓	✓
imagenette	smallcnn	S	br	✓	✓	✓
imagenette	smallcnn	S	no	–	–	–
imagenette	vgg11	N	br	✓	–	✓
imagenette	vgg11	N	no	✓	–	✓
imagenette	vgg11	O	br	✓	✓	✓
imagenette	vgg11	O	no	✓	✓	✓
imagenette	vgg11	RM	br	–	✓	–
imagenette	vgg11	RM	no	✓	✓	✓
imagenette	vgg11	RS	br	✓	✓	✓
imagenette	vgg11	RS	no	✓	✓	✓
imagenette	vgg11	S	br	✓	✓	✓
imagenette	vgg11	S	no	✓	✓	✓
imagenette	vit-b-16	N	br	✓	–	✓
imagenette	vit-b-16	N	no	✓	–	✓
imagenette	vit-b-16	O	br	✓	–	✓
imagenette	vit-b-16	O	no	✓	✓	✓
imagenette	vit-b-16	RM	br	✓	✓	–
imagenette	vit-b-16	RM	no	✓	✓	✓
imagenette	vit-b-16	RS	br	✓	✓	✓
imagenette	vit-b-16	RS	no	✓	✓	✓
imagenette	vit-b-16	S	br	✓	✓	✓
imagenette	vit-b-16	S	no	✓	✓	✓

Table 7: **Metric agreement summary.** Counts over all {dataset, model, regime, stage, break} groups.

Total groups	100
TV significant	97
JS significant	82
Hellinger significant	94
TV $\wedge$ Hellinger significant	93 / 100
TV $\wedge$ JS significant	80 / 100
All three significant	80 / 100
TV-only significant	0 / 100
non-TV-only significant	2 / 100
None significant	4 / 100

Break	Mean $\overline{\Delta}$	Median $\overline{\Delta}$	Count
no	0.000597	0.000000	10
br	0.000516	0.000000	10

Table 8: **Negative regime as control (TV).** Means and medians across setups under no-break and break conditions.

Table 9: **TV setups with non-significant or borderline effects after BH correction.**

Dataset	Model	Regime	Stage	Break	mean Δ_{BF}	95% CI	q
cifar100	resnet18	S	br	0.004149	[-0.000556, 0.008643]	8.45e-02	
cifar100	smallcnn	O	br	0.011148	[-0.000787, 0.022086]	6.76e-02	
cifar100	vgg11	RM	no	-0.001379	[-0.003891, 0.000891]	2.64e-01	
imagenette	smallcnn	O	no	0.002243	[-0.006734, 0.011739]	6.22e-01	
imagenette	smallcnn	S	no	0.001042	[-0.001016, 0.003169]	3.55e-01	
imagenette	vgg11	RM	br	0.001541	[-0.001617, 0.004896]	3.55e-01	