# Local-Interaction Learning Dynamics: A Markov Random Field View of Deep Learning

**Wen Dong**
Air Force Research Laboratory
wen.dong@afrl.af.mil

## Abstract

Modern neural-network training is governed by local interactions: residual streams carry signal, normalization changes scale, attention and state-space blocks route information, augmentation changes which directions generalize, and optimizer state determines the actual checkpoint update. Parameter-space curvature and full-weight posterior views reveal global geometry, but they do not show which layer, block, or interface created it. We introduce a computation-graph view in which parameters $\theta$ and intermediate states $u$ define a local Markov–Gibbs system. A potential energy encodes module relations, data loss, and regularization; its Gibbs law gives a probability reference, and its checkpoint-local Fisher/Gauss–Newton precision matrix gives the curvature geometry. This precision matrix is sparse because variable groups interact only when they share a local module or loss factor. In the hard-constraint limit, eliminating intermediate states recovers the reduced parameter curvature. From this common object we prove two results: separator transfers give two-sided certificates for checkpoint-linearized stability and active plasticity, and a resolved KL-to-Gibbs bound gives a bounded-loss PAC-Bayes consequence on a declared comparison space under the admissibility conditions on the prior. Checkpoint diagnostics on CIFAR-10 and GPT-20M/FineWeb-Edu use sampled Gram operators and boundary-gradient probes to examine learning-rate stability, plasticity, and generalization geometry without dense Hessian, Fisher, or covariance matrices.

## 1 Introduction

Modern neural-network training is controlled by local interactions. Residual streams carry signal across depth, normalization changes local scale, attention and state-space layers route information, augmentation changes which directions generalize, and optimizer state determines the actual update taken at a checkpoint. Yet much of the theory used to discuss stability and generalization is written in parameter space: Hessian, Fisher, and Gauss–Newton matrices are parameter-space objects, and posterior or Gibbs comparisons are often stated over all weights at once. These views are powerful, but they can obscure where architecture enters.

This paper makes the computation graph part of the mathematical object. The trainable parameters are $\theta$, and the intermediate states of the forward pass—activations, normalization variables, attention variables, residual-stream states, scan states, and outputs—are $u$. Local module relations, data loss, and regularization define a potential energy over $(\theta, u)$. The same potential has two faces: its Gibbs law is a probability reference, and its checkpoint-local Fisher/Gauss–Newton precision matrix is the curvature geometry. Because each factor touches only a small group of variables, this lifted precision records local interactions inside layers, blocks, residual additions, normalization interfaces, attention routes, and data-loss factors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The viewpoint connects several established tools in learning theory. The edge-of-stability literature studies how large curvature scales interact with learning rate near training checkpoints [7]. Natural gradient, Fisher/Gauss–Newton, and K-FAC methods use local curvature as an optimization geometry [1, 3, 24]. Factor graphs and Gaussian graphical models show how local factors induce sparse precision matrices and how elimination mediates interactions through interfaces [15, 27, 32]. SGLD, stochastic-gradient diffusion views, and PAC-Bayes analyses compare training-induced distributions with Gibbs-like references for generalization [5, 19, 23, 34]; related Bayesian and subspace posterior methods show how local stochastic geometry can matter in overparameterized models [13, 22, 35]. Our contribution is to connect these languages through the neural computation graph: one local precision matrix exposes architecture, supplies the curvature for criticality, and defines the Gibbs geometry for KL/PAC-Bayes comparison.

The first theorem identifies reduced Fisher/Gauss–Newton curvature as the hard-constraint limit of activation elimination. In lifted $(\theta, u)$-coordinates, the precision matrix is sparse on the computation graph; eliminating $u$ gives the Schur complement $F_\tau$, which converges to $F_{\mathrm{red}}$. Figure 1 makes this architecture-visible object concrete for MLP, residual CNN, and transformer blocks. The second theorem turns architecture into criticality. Separator transfers describe how perturbation and backward-sensitivity signals pass through eliminated interfaces, getting preserved, magnified, or damped. They yield upper certificates for the local stability edge and lower certificates for active directions. The resulting criticality theorem is two-sided: useful training needs both a safe optimizer step and active directions with loss-changing curvature.

Theorem 4 gives a distributional companion. Neural networks contain symmetries, gauges, and loss/output tangents: directions that can be weakly explored by minibatch training and barely change predictions, while contributing strongly to inverse-covariance drift-mismatch terms. We therefore compare the training-induced law with a localized Gibbs reference on a declared comparison space $E$. Under a log-Sobolev inequality on $E$, KL contracts toward the reference up to an accumulated defect. For bounded losses, the PAC-Bayes consequence additionally requires the prior and predictor embedding to be independent of the certificate sample, possibly conditional on independent build data. This resolved view turns PAC-Bayes from a full-space posterior claim into a local diagnostic of the probability geometry created by a training scheme.

The experiments study checkpoint diagnostics of these quantities. Across the CIFAR-10 and GPT-20M/FineWeb-Edu runs, the plotted optimizer-step coordinates stay below their sampled-curvature edge proxies, while the active-plasticity diagnostic tracks curvature on the declared gradient span. The KL/PAC-Bayes diagnostics show that low-covariance gauge/tangent directions can dominate the inverse-metric defect while barely moving predictions, and retained directions expose train/validation alignment under augmentation. Separate build/certificate splits yield direct prior-KL clipped-NLL bounds across MNIST and CIFAR-10 architectures.

The paper contributes four theorem-facing tools. First, it identifies reduced Fisher/Gauss–Newton curvature as the hard-constraint limit of activation elimination from a sparse computation-graph precision matrix. Second, it turns separator transfers into architecture-local certificates for both the local stability edge and active plasticity. Third, it proves a resolved KL-to-Gibbs bound with a bounded-loss PAC-Bayes consequence and a local block decomposition of the defect. Fourth, it implements checkpoint diagnostics using small Gram problems and boundary-gradient probes, with exact nonzero spectra for the sampled operators and calibrated proxy interpretations for theoretical curvature and transport. These diagnostics provide a practical language for stability, plasticity, and generalization geometry in neural-network training.

## 2 From computation graph to local Gibbs precision

We make the interaction between parameters, intermediate states, and data explicit. The training set is $D_n = \{(x_i, y_i)\}_{i=1}^n$, the trainable parameters are $\theta$, and the intermediate quantities already computed by the forward pass—activations, normalization variables, attention variables, residual-stream states, scan states, and outputs—are denoted by $u$. We write $z = (\theta, u)$. The lift is bookkeeping for variables already present in the computation graph. Its purpose is to expose local interactions before the intermediate states are substituted out.

Appendix A gives a worked two-layer MLP dictionary of variables, factor scopes, local residual Jacobians, and clique–separator interfaces.

Let $\mathcal{F}_n$ be the collection of local factors induced by module relations, data loss, and regularization. Each factor $f \in \mathcal{F}_n$ depends only on a small scope $S_f$ of variables in $z$. The lifted potential is

$$E_n(z; D_n) = \sum_{f \in \mathcal{F}_n} E_f(z_{S_f}; D_n). \tag{1}$$

A module factor may encode an affine map, normalization, activation, attention update, residual addition, or scan transition; a data factor encodes the loss at the output; a prior or regularizer is a factor on parameter variables. In the hard-constraint limit, low-energy configurations recover the deterministic forward computation and the usual empirical objective. When a law such as $\rho_t$ or $\pi_{n,\beta}$ is placed on lifted states, $u$ is random as part of the joint coordinate $z = (\theta, u)$; conditioned on a deterministic forward pass, $u$ lies on the usual network states.

The same potential has two readings. First, at inverse temperature $\beta > 0$, it defines the lifted Gibbs law

$$\pi_{n,\beta}(dz) = Z_{n,\beta}^{-1} \exp\{-\beta E_n(z; D_n)\}\, m(dz), \tag{2}$$

where $m$ is the reference measure on lifted states. Equivalently,

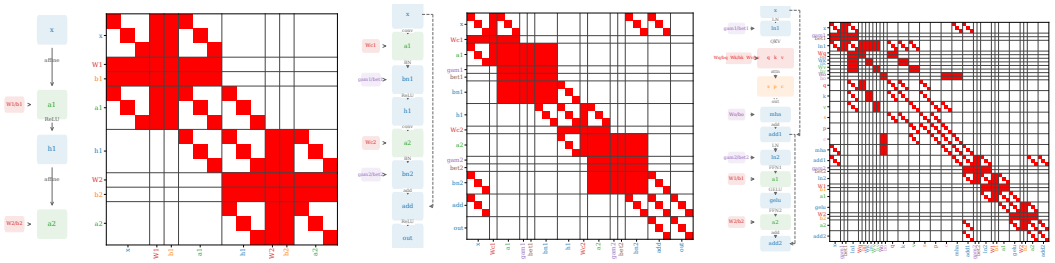
$$\pi_{n,\beta} = \arg\min_{\rho} \left\{ \mathbb{E}_\rho E_n + \beta^{-1} \mathrm{KL}(\rho \| m) \right\}. \tag{3}$$

Thus the potential specifies local interactions, and the Gibbs law turns those interactions into a probability reference. Second, at a checkpoint $z_0$, the Fisher/Gauss–Newton information form of the same factorized potential defines a local precision matrix. A residual-form factor has local first variation $r_f(z_0 + \Delta z_{S_f}) = r_f(z_0) + J_f(z_0)\Delta z_{S_f} + o(\|\Delta z_{S_f}\|)$; with fixed base weight $W_f \succeq 0$, its unscaled information contribution is $J_f(z_0)^\top W_f J_f(z_0)$. Embedding the local contributions with a common module-relation variance $\tau > 0$ gives

$$\begin{aligned} P_{\theta,u}(z_0; \tau) = \Lambda_\theta \oplus 0_u + \tau^{-1} \sum_{f \in \mathcal{F}_n^{\mathrm{mod}}} \mathrm{Embed}_{S_f}\left(J_f(z_0)^\top W_f J_f(z_0)\right) \\ + \sum_{f \in \mathcal{F}_n^{\mathrm{data}}} \mathrm{Embed}_{S_f}\left(J_f(z_0)^\top W_f J_f(z_0)\right), \end{aligned} \tag{4}$$

where $\mathcal{F}_n^{\mathrm{mod}}$ and $\mathcal{F}_n^{\mathrm{data}}$ contain module and data-loss factors, respectively, and $\Lambda_\theta \succeq 0$ contains parameter regularization/prior precision. The Jacobians, base weights, and $\Lambda_\theta$ are fixed at $z_0$ as $\tau \downarrow 0$; increasing the module precision $\tau^{-1}$ enforces the weighted linearized module relations.

Equation (4) is the architecture-visible object. Two variable groups can have a nonzero precision block only when they co-occur in a local factor. A zero block means no direct local interaction in the lifted factorization. Figure 1 shows the resulting support patterns for MLP, residual CNN, and transformer blocks.



(a) MLP block (b) residual CNN block (c) transformer block

Figure 1: **Lifted locality makes curvature architecture-visible.** A red block means that two variable groups co-occur in a local computational factor, so the corresponding block of $P_{\theta,u}$ may be nonzero. The MLP gives chain-like locality, the residual CNN adds skip and normalization interfaces, and the transformer adds residual-stream, attention, and feed-forward couplings. Eliminating $u$ gives $F_\tau$, which converges to $F_{\mathrm{red}}$ in the hard-constraint limit.

We now eliminate intermediate states. Partition $P_{\theta,u}$ into parameter and state blocks. The reduced parameter geometry is the Schur complement.

**Theorem 1** (Activation elimination recovers reduced Fisher/GGN)**.** *At a forward-consistent checkpoint $z_0$, let $P_{\theta,u}(z_0;\tau)$ be the lifted precision in* (4)*. Assume the weighted linearized module equations $W_f^{1/2}(J_{f,\theta}\Delta\theta + J_{f,u}\Delta u) = 0$, $f \in \mathcal{F}_n^{\mathrm{mod}}$, determine a unique $\Delta u$ for every $\Delta\theta$. For $\tau > 0$, eliminating intermediate states gives*

$$F_\tau = P_{\theta\theta}(\tau) - P_{\theta u}(\tau)P_{uu}(\tau)^{-1}P_{u\theta}(\tau). \tag{5}$$

*As $\tau \downarrow 0$, $F_\tau$ increases in Loewner order to the reduced Fisher/GGN curvature $F_{\mathrm{red}}$. On a junction-tree representation, the same Schur complement is computed by local Gaussian elimination over cliques and separators while retaining all parameter variables.* Proof in Appendix C.

Theorem 1 can be read as enforcing the linearized forward pass. For example, the MLP activation factor penalizes $h_1 - \phi(a_1)$ and involves only those two state groups; affine factors also involve their parameter blocks (Appendix A). Elimination minimizes the local quadratic form over state perturbations for a fixed parameter perturbation. At finite $\tau$, the states can partly accommodate that change by departing from the linearized module relations, so the reduced curvature can be softer than $F_{\mathrm{red}}$. Decreasing $\tau$ makes such departures more costly. In the limit, the state response follows the unique linearized forward computation, and substitution into the loss and regularizer quadratics recovers the usual chain-rule Fisher/GGN. The resulting parameter couplings therefore have an architectural explanation through the intermediate states and interfaces that mediate them.

The section gives the shared object used by the rest of the paper:

$$\text{computation graph} \Rightarrow E_n(\theta,u) \Rightarrow \begin{cases} \pi_{n,\beta} \propto e^{-\beta E_n} & \text{Gibbs reference,} \\ P_{\theta,u}(z_0;\tau) & \text{sparse lifted precision,} \\ F_{\mathrm{red}} & \text{hard-constraint reduced curvature.} \end{cases}$$

Section 3 uses $F_{\mathrm{red}}$ for stability and active plasticity. Section 4 uses the Gibbs law and lifted/reduced precision as the local probability geometry for KL/PAC-Bayes comparison.

## 3 Separator transfers and two-sided criticality

After activation elimination in the hard-constraint limit, local training is governed by the reduced curvature $F_{\mathrm{red}}$ from Theorem 1. This reduced object may couple parameter blocks that are distant in the computation graph, but those couplings are created by eliminating variables from the sparse lifted precision in (4). Criticality is therefore an architecture question: which local interfaces route curvature and backward signal after elimination, and do they produce a checkpoint that is both stable and still able to learn?

We read this question in optimizer coordinates. Let $P \succ 0$ denote the checkpoint optimizer metric, with $P = I$ for plain gradient descent, and write $\bar{F} = P^{1/2}F_{\mathrm{red}}P^{1/2}$. The upper side of criticality is the local no-explosion edge

$$\eta_{\max}(\theta) = \frac{2}{\lambda_{\max}(\bar{F}(\theta))}. \tag{6}$$

The lower side is active plasticity. Let $\mathcal{U}_{\mathrm{act}}$ be a declared active subspace: in experiments, a small span of recent or probe gradients; in architecture arguments, a residual carrier, memory path, attention route, or other task-active interface. The lower-side question is whether $\bar{F}$ has nontrivial curvature on this active span.

Separator transfers describe the mechanism. For sample $i$, let $J_{i,v}^{\mathrm{red}} := \partial o_i/\partial\theta_v$. Under the module assumption of Theorem 1, Appendix D constructs local source injections $B_{v,i}^{(C)}$, separator-source transfers $T_e$, and a root output response $\mathcal{R}_i$ from the linearized module constraints. Set $s_{v,i}^{(C)} = \|B_{v,i}^{(C)}\|_{\mathrm{op}}$, $g_e = \|T_e\|_{\mathrm{op}}$, and $\gamma_i = \|\mathcal{R}_i\|_{\mathrm{op}}$ in the declared Euclidean coordinates. With output root $C_{\mathrm{out}}$ and injection cliques $\mathcal{C}(v)$, define

$$S_{v,i} = \gamma_i \sum_{C\in\mathcal{C}(v)} s_{v,i}^{(C)} \prod_{e\in\mathrm{path}(C\rightsquigarrow C_{\mathrm{out}})} g_e. \tag{7}$$

Equation (7) follows a parameter perturbation through the module system. Local injections create information sources, separator maps carry them toward the output root, and the root response converts

their sum into an output displacement. Norms multiply along each route and add across injection sites. The link to curvature is

$$F_{\mathrm{red}} = \Lambda_\theta + \frac{1}{|B|}\sum_{i\in B}(J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}},$$

where $B$ is the sample batch defining the local loss and $H_i \succeq 0$ is its output-space loss Hessian. Thus bounds on transported output sensitivity give an upper curvature certificate.

**Theorem 2** (Separator-transfer certificates for reduced curvature)**.** *Fix the setup of Theorem 1 and use the module-constraint budgets* (7)*.*

***(i) Upper transfer.*** *For every sample $i$ and parameter block $v$, $\|J_{i,v}^{\mathrm{red}}\|_{\mathrm{op}} \le S_{v,i}$. If $0 \preceq H_i \preceq M_H I$ for all $i \in B$, then*

$$\lambda_{\max}(F_{\mathrm{red}}) \le \lambda_{\max}(\Lambda_\theta) + \frac{M_H}{|B|}\sum_{i\in B}\sum_{v} S_{v,i}^2. \tag{8}$$

***(ii) Active coverage.*** *Let $\mathcal{U}_{\mathrm{act}}$ be a declared active subspace. Suppose that*

$$\frac{1}{|B|}\sum_{i\in B}\nu^\top (J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}}\nu \ge \mu_{\mathrm{act}} \quad \textit{for every unit } \nu \in \mathcal{U}_{\mathrm{act}}. \tag{9}$$

*Then*

$$\lambda_{\min}(F_{\mathrm{red}}|_{\mathcal{U}_{\mathrm{act}}}) \ge \lambda_{\min}(\Lambda_\theta|_{\mathcal{U}_{\mathrm{act}}}) + \mu_{\mathrm{act}}. \tag{10}$$

*The same statements hold in optimizer-whitened coordinates with $\bar{F} = P^{1/2}F_{\mathrm{red}}P^{1/2}$, $\bar{\Lambda}_\theta = P^{1/2}\Lambda_\theta P^{1/2}$, and budgets constructed for the block columns of $J_i^{\mathrm{red}}P^{1/2}$.* Proof in Appendix D.

Theorem 2 separates possible amplification from active coverage. Upper norm budgets alone cannot ensure a loss-changing response: transported contributions can cancel or become invisible to the loss. Active coverage separately requires a loss-weighted output response in every declared active direction, averaged over samples. Theorem 3 converts the two certificates into a stable scalar step and a nonzero linearized response on the active span.

**Theorem 3** (Two-sided architecture-local criticality)**.** *Let $\bar{F} = P^{1/2}F_{\mathrm{red}}P^{1/2}$. Suppose that*

$$\lambda_{\max}(\bar{F}) \le S_+^2, \qquad \lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S_-^2 > 0. \tag{11}$$

*Then every scalar step $0 < \eta < 2/S_+^2$ is locally stable for the linearized update in optimizer coordinates. Moreover, for every unit $u \in \mathcal{U}_{\mathrm{act}}$, $\|\eta\bar{F}u\|_2 \ge \eta S_-^2$. Thus the linearized update is nonexpansive on the full parameter space in optimizer coordinates and nontrivial on the active subspace.* Proof in Appendix D.

The checkpoint lower-side measurement is

$$\rho_{\mathrm{act}}(\theta) = \eta_{\mathrm{eff}}(\theta)\,\lambda_{\min}\big(\bar{F}(\theta)|_{\mathcal{U}_{\mathrm{act}}}\big), \tag{12}$$

where $\eta_{\mathrm{eff}}$ is reconstructed from the saved optimizer state. For $\eta_{\mathrm{eff}}(\theta) > 0$, the certificate (11) gives $\rho_{\mathrm{act}}(\theta) \ge \eta_{\mathrm{eff}}(\theta)S_-^2$.

Theorem 3 separates two failure modes. If the actual optimizer step is too large relative to $2/S_+^2$, the local update can become unstable. If $S_-^2$ is small on the directions used by the optimizer, the update can remain stable while becoming inactive. This is why criticality is two-sided: useful training needs both a safe upper edge and an active lower signal. For an eigenmode of $\bar{F}$ with eigenvalue $\lambda$, the linearized update has multiplier $1 - \eta\lambda$; when $\eta\lambda$ is small, the mode changes little even within the stable range.

Section 5.1 estimates checkpoint counterparts of these two coordinates. The upper plot compares a scalar optimizer-step coordinate with a sampled-curvature edge proxy, while the lower plot tracks the active-plasticity diagnostic. Appendix D gives DenseNet, ResNet/WideResNet, Transformer/GPT, and SSM instantiations of the theorem interface. Appendix B.1 shows that residual scale, normalization, width, compression, and scan/state choices move the same two diagnostic coordinates, and that backward-gain diagnostics localize where architecture preserves, magnifies, or damps signal.

## 4 Resolved KL-to-Gibbs comparison

The criticality theorem asks whether a checkpoint is stable and still active. We now ask the complementary distributional question: what local probability law is created by training, and how close is it to a Gibbs reference that favors low bounded loss? Fix a checkpoint and a finite-dimensional comparison space $E$, held fixed during the local comparison. We write $\rho_t^E$ for the law of the assumed local Itô model and $\pi_\beta^E$ for its Gibbs reference on $E$. When both laws are embedded in parameter space with the same fixed $E^\perp$ component, their ambient KL equals the KL between their $E$-coordinate laws. This is a declared local comparison, not an identification of the optimizer's full parameter law. The checkpoint diagnostics in Section 5.2 construct $E$ from a retained positive-spectrum subspace of a blockwise minibatch-gradient covariance estimate.

For a reference law $\pi_0^E$, bounded empirical loss $\widehat{L}_E$, and inverse temperature $\beta > 0$, define

$$\pi_\beta^E(dw) \propto \exp\{-\beta \widehat{L}_E(w)\}\, \pi_0^E(dw). \tag{13}$$

The Gibbs free-energy identity is

$$\widehat{L}_E(\rho) + \beta^{-1}\mathrm{KL}(\rho\|\pi_0^E) - \mathcal{F}_\beta^E = \beta^{-1}\mathrm{KL}(\rho\|\pi_\beta^E), \qquad \mathcal{F}_\beta^E = -\beta^{-1}\log \int e^{-\beta \widehat{L}_E(w)}\, \pi_0^E(dw). \tag{14}$$

Thus KL to the localized Gibbs law is excess bounded-loss/complexity free energy on $E$.

Let $b_E$ and $a_E$ be the drift and covariance of an Itô law $\rho_t^E$ on $E$. Let $b_{\mathrm{rev},E}$ be the reversible drift whose invariant law is $\pi_\beta^E$ with covariance $a_E$; for constant covariance, $b_{\mathrm{rev},E} = \frac{1}{2} a_E \nabla \log \pi_\beta^E$. Define the training-reference defect

$$\mathfrak{D}_E(t) = \mathbb{E}_{\rho_t^E}\left[\|b_E - b_{\mathrm{rev},E}\|^2_{a_E^{-1}}\right]. \tag{15}$$

This defect is zero for reversible Gibbs dynamics on $E$ and otherwise measures drift mismatch in the covariance metric of the assumed Itô model. The checkpoint calculations estimate local first and second stochastic moments. The original diagnostics used a raw-objective-gradient drift surrogate and centered minibatch-gradient probes, with stationary momentum scaling for CIFAR; they did not reconstruct the full Nesterov/AdamW transition. These moment estimates do not establish an Itô limit or the metric log-Sobolev inequality. Appendix B.1 reports an optimizer-aware calibration.

We use the metric log-Sobolev inequality in the following form: $\pi_\beta^E$ has rate $\lambda_E > 0$ if, for every smooth $\rho$,

$$\mathcal{I}_E(\rho\|\pi_\beta^E) \geq 2\lambda_E\, \mathrm{KL}(\rho\|\pi_\beta^E), \qquad \mathcal{I}_E(\rho\|\pi_\beta^E) = \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\pi_\beta^E}\right\|^2_{a_E}\right].$$

This is the condition that turns entropy dissipation into KL contraction.

**Theorem 4** (Resolved KL-to-Gibbs comparison)**.** *Fix a checkpoint and a declared comparison space $E$. Let $\rho_t^E$ be an Itô law on $E$ with drift $b_E$ and covariance $a_E$, and let $\pi_\beta^E$ be the localized Gibbs reference in* (13)*. Suppose the metric log-Sobolev inequality above holds. Then, for any $\varepsilon \in (0,1)$,*

$$\mathrm{KL}(\rho_t^E\|\pi_\beta^E) \leq e^{-(1-\varepsilon)\lambda_E t}\mathrm{KL}(\rho_0^E\|\pi_\beta^E) + \frac{1}{2\varepsilon}\int_0^t e^{-(1-\varepsilon)\lambda_E(t-s)}\mathfrak{D}_E(s)\, ds. \tag{16}$$

*In a checkpoint-local quadratic model with lower curvature $\mu_E$ on $E$ and covariance scale $a_E \succeq a_{\min} I$, one may take $\lambda_E \gtrsim \beta\mu_E a_{\min}$. If $E = \oplus_{B\in\mathcal{B}} E_B$ is an orthogonal block decomposition and $a_E = \oplus_{B\in\mathcal{B}} a_{E_B}$, then*

$$\mathfrak{D}_E(t) = \sum_{B\in\mathcal{B}} \mathfrak{D}_{E_B}(t), \tag{17}$$

*where $\mathfrak{D}_{E_B}(t) := \mathbb{E}_{\rho_t^E}[\|(b_E - b_{\mathrm{rev},E})_B\|^2_{a_{E_B}^{-1}}]$.* Proof in Appendix E.

Equation (16) separates the fading influence of the initial distribution from continued mismatch between the assumed training drift and the reversible Gibbs drift. Older mismatches receive less weight because contraction has acted on them for longer. With zero defect, the KL bound contracts

exponentially. If $\mathfrak{D}_E(s) \le \overline{\mathfrak{D}}$ throughout the comparison interval, taking $\varepsilon = 1/2$ bounds the accumulated contribution by $2\overline{\mathfrak{D}}/\lambda_E$. Faster relaxation and smaller drift mismatch therefore tighten the comparison. By Eq. (14), $\beta^{-1}$ times this KL is excess empirical loss-plus-complexity free energy; under the prior/sample admissibility conditions of Corollary 10, that excess enters the population-risk bound. Under the stated block conditions, Eq. (17) also attributes the defect to the contributing parameter groups. Appendix E gives the conditional-build statement and proof.

The reference geometry is the same computation-graph geometry used for criticality. The potential in (1) defines the lifted Gibbs law (2) through the variational principle (3); its checkpoint-local precision is the sparse lifted precision (4). Eliminating intermediate states gives the finite-$\tau$ Schur-complement geometry in (5); its hard-constraint limit is $F_{\mathrm{red}}$. Thus the criticality theorem uses the reduced precision as curvature, while the KL theorem uses the same lifted/reduced precision as Gibbs reference geometry.

The comparison space $E$ exposes directions in which stochastic training varies locally. In a covariance eigendirection with eigenvalue $\sigma^2$, a drift mismatch $m$ contributes $m^2/\sigma^2$ pointwise to the defect integrand in (15), before averaging under $\rho_t^E$. Prediction sensitivity and inverse-covariance mismatch therefore measure different properties of a direction. Weakly explored gauge or tangent directions can dominate this metric even when finite movement along them barely changes held-out predictions. A retained-space comparison is defined by the prior/posterior support construction and the shared conditional complement; covariance thresholding alone provides a local diagnostic (definitions and audits in Appendix B.2).

At checkpoint $j$, the retained-spectrum implementation has the compact form

$$D_{j,E}^{(\ell)}(\beta) = d_{0,j} + \beta d_{1,j}^{(\ell)} + \beta^2 d_{2,j}^{(\ell)}. \tag{18}$$

Here $d_{0,j}$ is inverse-covariance drift mismatch, which can be dominated by weakly explored gauge or tangent directions; $d_{1,j}^{(\ell)}$ measures alignment between training drift and the bounded-loss Gibbs score; and $d_{2,j}^{(\ell)}$ measures the Gibbs score inside the retained covariance geometry. The coefficient formulas and blockwise Gram computation are given in Appendix E. Figure 3 visualizes this checkpoint geometry in ResNet-32/CIFAR-10. Table 1 reports separate split-based bounds in normalized clipped-NLL units, computed from exact prior KL rather than from Eq. (18) or the dynamical bound in Theorem 4. The split protocol is given in Appendix B.2.

## 5 Experiments

The theorems above define checkpoint-local quantities: an upper stability edge, an active-plasticity coordinate, and a resolved KL/PAC-Bayes geometry. We compute checkpoint diagnostics of these quantities using sampled-gradient Grams and boundary probes, without forming dense Hessian, Fisher, or covariance matrices. The goal is to show what the theory makes visible: when training is stable, when it is still plastic, and which local probability directions matter for generalization.

### 5.1 Two-sided criticality: stable and still active

The criticality theorem concerns a scalar checkpoint-linearized update with both an upper stability condition and an active-curvature condition. Figure 2 reports diagnostics of these two sides on CIFAR-10 [14] residual, dense, and wide convolutional architectures and on a from-scratch GPT-20M run on BPE/FineWeb-Edu tokens [21]. The upper row compares the scalar optimizer-step coordinate with the sampled-curvature edge proxy; the lower row plots the active-plasticity diagnostic. Across these runs, the plotted step coordinate stays below the proxy edge, while the active coordinate remains measurable after early training.

Figure 2 shows the baseline operating regimes. Figures B.1 and B.2 provide the failure comparison: GPT/SSM no-residual runs remain below the reported scalar upper-edge proxy while the active coordinate collapses. Thus the second coordinate changes the diagnosis even when the first remains favorable. Figure B.3 then identifies boundaries with amplified or attenuated directional backward transport. Together, these measurements provide an ordered diagnostic procedure: check the scalar upper edge, inspect active curvature, and examine the interfaces where transport changes. Architecture edits move the same measured coordinates: residual scale, normalization, width, compression, and

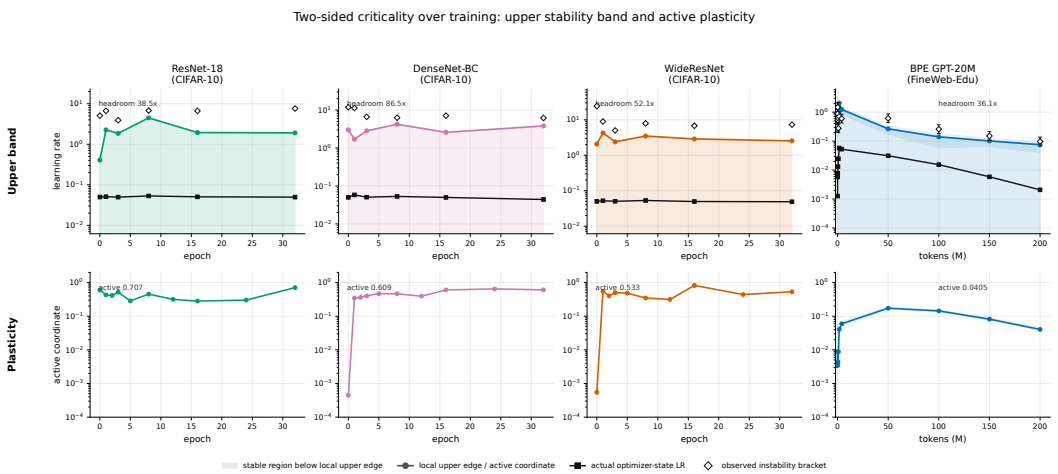



Figure 2: **Two-sided criticality diagnostics over training.** The black curve is the scalar projection of the update reconstructed from the saved optimizer state. The colored curve is the upper-edge proxy computed from the sampled gradient Gram, and shading marks the side indicated as stable by this scalar diagnostic. Diamonds show observed instability brackets from state-matched LR-cliff sweeps when available. The lower row reports the optimizer-step coordinate times the minimum eigenvalue of the sampled operator projected onto the active gradient span. These are calibrated checkpoint diagnostics of the two sides of Theorem 3, not complete stability guarantees for momentum or AdamW.

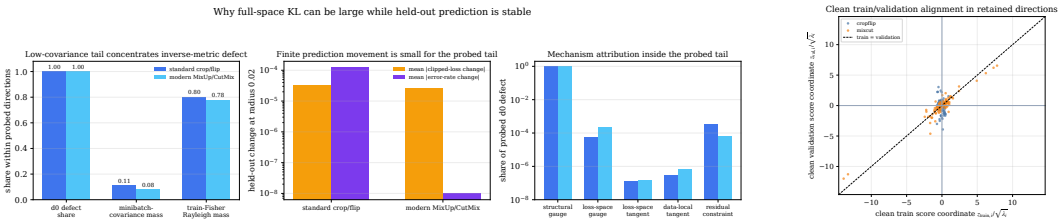
scan/state choices change active plasticity and backward gain (Appendix B.1). These results support the architecture interpretation of Theorem 2: local interfaces change how curvature and backward sensitivity are preserved, magnified, or damped.

All criticality measurements are checkpoint-local. The uncentered sampled-gradient Gram matrix has the same nonzero spectrum as its sampled empirical-Fisher-type operator; this identity does not make that operator equal to the GGN. Active plasticity uses a small recent/probe-gradient span, and boundary VJP norm ratios are directional transport proxies rather than separator operator norms. We calibrate these diagnostics with state-matched LR-cliff sweeps, random-subspace controls, and appendix sensitivity checks.

### 5.2 Resolved KL/PAC-Bayes: the local probability space of training

The inverse-covariance metric in (15) gives larger weight to drift mismatch along directions with small covariance eigenvalues. Figure 3 examines this defect geometry in ResNet-32/CIFAR-10 together with finite prediction movement and directional probes of symmetries, gauges, and loss/output tangents. These measurements do not directly estimate a full-parameter posterior KL. Panel (a) shows that the probed low-covariance tail concentrates the inverse-metric defect while carrying only a small fraction of ordinary covariance mass, and finite held-out prediction movement remains small. Panel (b) shows the generalization-facing retained geometry: MixUp/CutMix improves clean train/validation score alignment in the retained directions. Panel (c) places these directions inside the full blockwise covariance spectrum, with diagnostic overlays for structural gauges, loss-space gauges, loss tangents, residual probes, block eigendirections, and MixUp/CutMix tangent similarity.

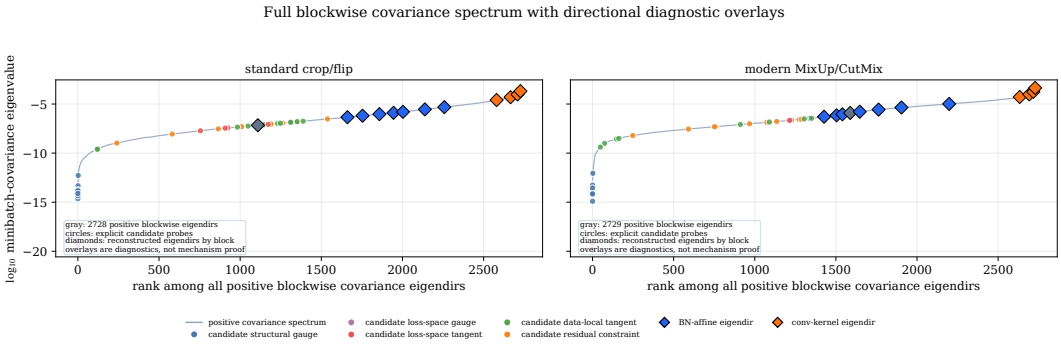
Table 1 reports direct prior-KL bounds from fresh build/certificate runs for five architectures on MNIST [17] and CIFAR-10. Conditional on build data and independent construction randomness, the prior and retained-space construction are fixed independently of the certificate sample; the same frozen posterior is used for KL and risk evaluation. The per-row PAC-Bayes and Monte Carlo failure budgets sum to $5(0.008 + 0.002) = 0.05$. Test data were accessed only after the certificate table was finalized. These reported normalized clipped-NLL bounds concern the randomized retained posterior and do not use the stationary defect proxy or Theorem 4's Itô/LSI assumptions. The split protocol is given in Appendix B.2.




(a) inverse-metric defect versus prediction movement




(b) retained train/validation alignment




(c) blockwise covariance spectrum with gauge/tangent diagnostics

Figure 3: **Resolved stochastic geometry in ResNet-32/CIFAR-10.** Panel (a) shows that the probed low-covariance tail concentrates the inverse-metric defect while carrying little covariance mass and producing small finite held-out prediction changes. Panel (b) shows clean train/validation score alignment in retained directions under augmentation. Panel (c) locates directional diagnostic overlays within the blockwise covariance spectrum; these overlays do not select the certificate posterior. The plots describe checkpoint geometry, not an evaluated full-parameter posterior KL or a population-risk bound. Figure B.4 provides a larger spectrum map.

Table 1: **Split-based prior-KL certificates for randomized retained posteriors.** Each architecture was freshly trained using build data only. The checkpoint, retained space, prior, posterior, confidence allocation, and 4,096 posterior draws per row were frozen before certificate access. The same posterior is used for exact Gaussian KL and posterior-risk evaluation. Per-row budgets $\delta_{\mathrm{PB}} = 0.008$ and $\delta_{\mathrm{MC}} = 0.002$ give simultaneous 95% coverage across the reported population bounds in normalized clipped-NLL units. Posterior test risk is a separate held-out audit. The protocol is given in Appendix B.2.

| Model | $k$ | $\mathrm{KL}(Q\|P)$ | Population clipped-NLL bound | Posterior test risk |
|---|---|---|---|---|
| MLP | 1 | 0.31815 | 0.04258 | 0.02198 |
| LeNet | 1 | 0.31815 | 0.02983 | 0.00898 |
| ResNet-32 | 16 | 5.09035 | 0.07657 | 0.05840 |
| WRN-16-4 | 11 | 3.49962 | 0.06349 | 0.04448 |
| DenseNet-small | 8 | 2.54518 | 0.07669 | 0.06100 |

The certificate compares an explicit prior and posterior supported on the same retained affine space. Their KL is the KL of the retained-coordinate laws; the build/certificate split makes the prior and embedding independent of certificate data. The weak-direction audits in Appendix B.2 describe checkpoint geometry but do not themselves establish these support or independence conditions. The geometry diagnostics use blockwise gradient Grams, retained spectra, and defect coefficients. The split certificate separately evaluates prior KL and posterior clipped-NLL risk on its build-selected space. Basis construction can use the same small-Gram computation, and neither calculation requires a dense full-parameter Hessian, Fisher, or covariance matrix.

## 6 Related work

This work connects four lines of deep-learning theory. Lifted formulations introduce auxiliary variables to expose or optimize through intermediate network states, as in the method of auxiliary coordinates and related lifted optimization methods [4, 9]. Our lift has a different purpose: the intermediate states make the computation-graph interactions visible, so that Fisher/Gauss–Newton curvature can be recovered from a sparse lifted precision by activation elimination in the hard-constraint limit. This places the work close to factor graphs, Gaussian graphical models, and approximate inference, where local factors, sparse precision matrices, and elimination or message passing expose mediated interactions [15, 25–27, 32, 33, 36].

The criticality part of the paper is related to curvature, conditioning, and trainability. Natural-gradient, Fisher/Gauss–Newton, and K-FAC methods use local curvature as optimization geometry [1, 3, 24], while empirical-Fisher studies clarify when common curvature proxies do or do not match the target geometry [16]. Edge-of-stability and signal-propagation studies show that learning rates, large curvature scales, residual paths, normalization, and depth-wise gradient transport shape modern training [2, 7, 10, 28, 29]. Our experiments instantiate the theory on standard residual, dense, and wide architectures [11, 12, 38]. The separator-transfer theorem expresses these architecture effects as local interfaces that preserve, magnify, or damp reduced-Jacobian signal, controlling both the upper stability edge and the lower active-plasticity coordinate.

The KL/PAC-Bayes part relates to stochastic-process views of SGD, stochastic modified equations, stochastic-gradient MCMC, and PAC-Bayes generalization bounds [5, 6, 18–20, 23, 30, 31, 34]. These literatures motivate Gibbs references and diffusion comparisons, but full-parameter comparisons can be too strong for neural networks with symmetries, gauges, and near-singular stochastic-gradient covariance. The resolved KL theorem instead compares an assumed local Itô law with a localized Gibbs reference on a fixed retained space, with any ambient embedding using the declared common conditional complement. Our retained-space experiments also connect to augmentation methods that change train/validation geometry [8, 37, 39].

## 7 Limitations and scope

The claims are checkpoint-local. Activation elimination is a Schur-complement identity for the stated checkpoint-local information model: the global identity does not depend on treewidth. Its identification with the deterministic reduced Fisher/GGN curvature requires the hard-constraint limit in Theorem 1. Controlled separator size makes exact local message passing efficient; for large models, the same identity motivates blockwise, sketched, and retained diagnostics. The criticality theorems certify local stability and active plasticity of the checkpoint-linearized dynamics, and the experiments calibrate the diagnostics with LR-cliff sweeps, random-subspace controls, and architecture ablations rather than claiming a global training guarantee. The KL theorem concerns an assumed local Itô law on a fixed comparison space and reference, under the stated regularity and log-Sobolev conditions. The experiments estimate local moments and report diagnostics; they do not establish these process-level assumptions. The PAC-Bayes consequence additionally requires an admissible prior independent of the certificate sample and applies to bounded risk of the declared randomized posterior. Its direct prior-KL form can be used without the Itô/LSI assumptions.

## 8 Conclusions and discussion

The paper gives a checkpoint-local language for modern training. The computation graph defines a potential; the potential defines a Gibbs reference and a sparse lifted precision; activation elimination recovers the reduced Fisher/Gauss–Newton curvature in the hard-constraint limit. Separator transfers explain how architecture changes the upper stability edge and active-plasticity signal, while resolved KL identifies the stochastic directions that should carry PAC-Bayes comparison. The resulting diagnostics are computable by Gram probes, active-subspace projections, retained spectra, and boundary-gradient summaries.

## Acknowledgments, Disclosure of Funding, and Disclaimer

The author thanks the reviewers for their constructive comments.

The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense or of the United States Air Force. The content or appearance of hyperlinks does not reflect an official DoD, Air Force, Air Force Research Laboratory position or endorsement of the external websites, or the information, products, or services contained therein.

## References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746.

[2] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL `https://arxiv.org/abs/1702.08591`.

[3] Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical gauss–newton optimisation for deep learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017.

[4] Miguel Carreira-Perpinan and Weiran Wang. Distributed optimization of deeply nested systems. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 10–19, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL `https://proceedings.mlr.press/v33/carreira-perpinan14.html`.

[5] Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

[6] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 1683–1691. PMLR, 2014. URL `https://arxiv.org/abs/1402.4102`.

[7] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jh-rTtvkGeM`.

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[9] Fangda Gu, Armin Askari, and Laurent El Ghaoui. Fenchel lifted networks: A lagrange relaxation of neural network training. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3362–3371. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/gu20a.html`.

[10] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1611.04231`.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[13] Pavel Izmailov, Dmitry Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, 2018.

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. In *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[15] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, 47(2):498–519, 2001. doi: 10.1109/18.910572.

[16] Frederik Kunstner, Lucas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent, 2019. URL https://arxiv.org/abs/1905.12558.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016. URL https://arxiv.org/abs/1512.07666.

[19] Lei Li and Yuliang Wang. On uniform-in-time diffusion approximation for stochastic gradient descent. *ArXiv*, abs/2207.04922, 2022. URL https://api.semanticscholar.org/CorpusID:250426296.

[20] Qiang Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20 (40):1–47, 2019. URL https://jmlr.org/papers/v20/17-526.html.

[21] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

[22] Wesley J. Maddox, T. Garipov, Pavel Izmailov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:59608823.

[23] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18:134:1–134:35, 2017. URL https://api.semanticscholar.org/CorpusID:9469223.

[24] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015.

[25] Thomas Minka. *Expectation Propagation for Approximate Bayesian Inference*. PhD thesis, MIT, 2001.

[26] Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 1999.

[27] Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL, 2005. ISBN 9781584884323. doi: 10.1201/9780203492024.

[28] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.nips.cc/paper/7515-how-does-batch-normalization-help-optimization.

[29] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.01232.

[30] Ulas Simsekli, Levent Sagun, Mert Gurbuzbalaban, and Aurelien Lucchi. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019. URL `https://proceedings.mlr.press/v97/simsekli19a.html`.

[31] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

[32] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008. doi: 10.1561/2200000001.

[33] Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 308–315. PMLR, 2003.

[34] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[35] Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świkatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*, pages 10248–10259, 2020.

[36] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. In *IEEE ISIT*, 2005.

[37] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, and Sanghyuk Chun. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[38] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2018.

# A Whole-network lifted-to-reduced operator patterns

**Worked MLP dictionary.** Consider one observed pair $(x, y)$ and the two-layer, one-hidden-layer network

$$a_1 = W_1 x + b_1, \qquad h_1 = \phi(a_1), \qquad o = W_2 h_1 + b_2.$$

Here $o$ is the output denoted $a_2$ in Figure 1(a). For this conditional example, the displayed input $x$ is held fixed: $\Delta x = 0$, and $x$ is excluded from the unobserved coordinates. The parameter groups are $\theta = (W_1, b_1, W_2, b_2)$ and the lifted state groups are $u = (a_1, h_1, o)$; $x, y$ are observed. The module residuals and scopes are

$$\begin{aligned} r_1 &= a_1 - W_1 x - b_1, & S_1 &= \{W_1, b_1, a_1\}, \\ r_\phi &= h_1 - \phi(a_1), & S_\phi &= \{a_1, h_1\}, \\ r_2 &= o - W_2 h_1 - b_2, & S_2 &= \{W_2, b_2, h_1, o\}. \end{aligned}$$

The output-loss factor has scope $\{o\}$, with $y$ observed. For this worked example, prior factors act separately on the parameter blocks. Give each module residual a positive-definite information weight $\Omega_f$, distinct from the network matrices $W_1, W_2$, and energy

$$E_{f,\tau}(z) = \frac{1}{2\tau} \|r_f(z)\|^2_{\Omega_f}.$$

At a checkpoint where the stated linearizations exist, $J_f = Dr_f$ and the factor contributes $\tau^{-1} \operatorname{Embed}_{S_f}(J_f^\top \Omega_f J_f)$ to the lifted precision. The structural primal graph joins groups that share a factor scope. A shared scope permits a nonzero precision block; a particular numerical block can still vanish. The residual Jacobians act as

$$\begin{aligned} J_1 \Delta z &= \Delta a_1 - \Delta W_1 x - \Delta b_1, \\ J_\phi \Delta z &= \Delta h_1 - D\phi(a_1) \Delta a_1, \\ J_2 \Delta z &= \Delta o - \Delta W_2 h_1 - W_2 \Delta h_1 - \Delta b_2. \end{aligned}$$

Enforcing these linearized module equations determines the state displacement uniquely and gives

$$\Delta o = \Delta W_2 h_1 + W_2 D\phi(a_1)(\Delta W_1 x + \Delta b_1) + \Delta b_2.$$

This is the reduced output Jacobian used in the hard-constraint limit.

For this single-example, block-prior construction, a junction-tree factor-scope cover is the chain $C_1 - C_2 - C_3 - C_4$, with

$$\begin{aligned} C_1 &= \{W_1, b_1, a_1\}, & C_2 &= \{a_1, h_1\}, \\ C_3 &= \{W_2, b_2, h_1, o\}, & C_4 &= \{o\}. \end{aligned}$$

Its intersections are $C_1 \cap C_2 = \{a_1\}$, $C_2 \cap C_3 = \{h_1\}$, and $C_3 \cap C_4 = \{o\}$. These describe the factor-scope cover, not an elimination schedule that can discard every clique-internal variable while retaining all parameters. For the state-only elimination in Theorem 1, parameter groups are retained and must be carried by the separators wherever required. A conservative construction augments each cluster and separator by all parameter groups and eliminates only the state variables; more tailored elimination orders may have smaller interfaces. The output cluster $C_4$ need not be maximal. In general, the cover depends on the elimination order and any chordal completion. Dense prior factors or shared parameters across multiple examples can add scopes and enlarge the required clusters.

Figure 1 in the main text gives the block-level rule: local computational factors induce a sparse lifted operator, and eliminating hidden or activation variables produces a denser parameter-only object. The two figures in this appendix show that the same mechanism persists at whole-network scale for a residual CNN and a transformer. These two architectures are the most visually informative whole-network witnesses in our current collection. We omit a whole-network MLP panel because, at this scale, it adds little beyond the chain-like block example already shown in Figure 1(a).

In each figure, the left panel gives the network schematic, the middle panel shows the numeric lifted operator over all variables under a single global ordering, and the right panel shows the corresponding parameter-only operator after eliminating hidden or activation variables at finite hidden/activation precision. These finite-precision operators illustrate elimination-induced coupling; they need not equal the hard-constraint limit $F_{\mathrm{red}}$. The point is structural rather than pixel-exact: repeated local motifs remain sparse and localized in lifted coordinates, while elimination spreads those local interactions into broader parameter couplings.

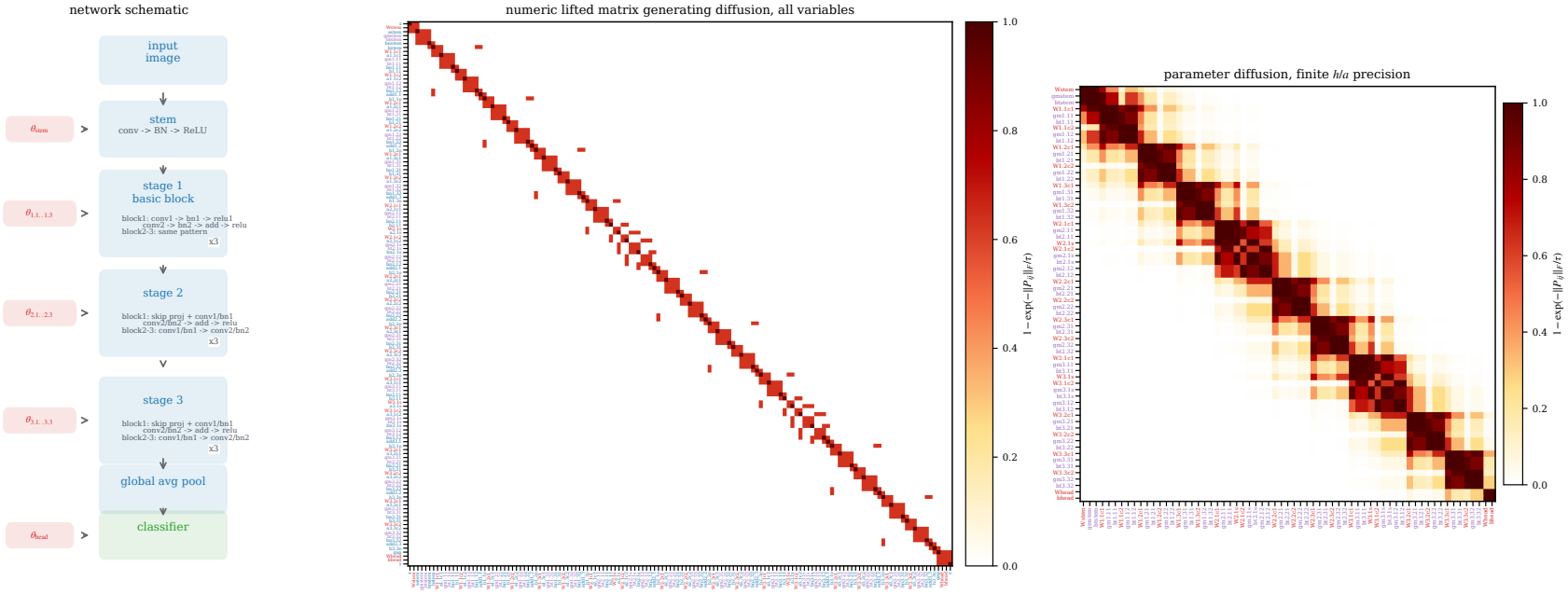



Figure A.1: **Whole-network residual CNN witness of the lifted-to-reduced transition.** Left: schematic of a ResNet20-style network with a stem, three stages of residual blocks, global average pooling, and a classifier. Middle: numeric lifted operator over all variables, ordered according to the schematic. The dominant pattern is still local: each residual block contributes a compact motif, and the stage-transition blocks with projection skips appear as visibly different local interfaces. Right: the corresponding parameter-only operator after eliminating hidden or activation variables at finite hidden/activation precision. The reduced object is substantially denser: local skip, normalization, and within-block interactions induce broader couplings among parameter groups, especially within and across stages. This is the whole-network analogue of Figure 1(b).

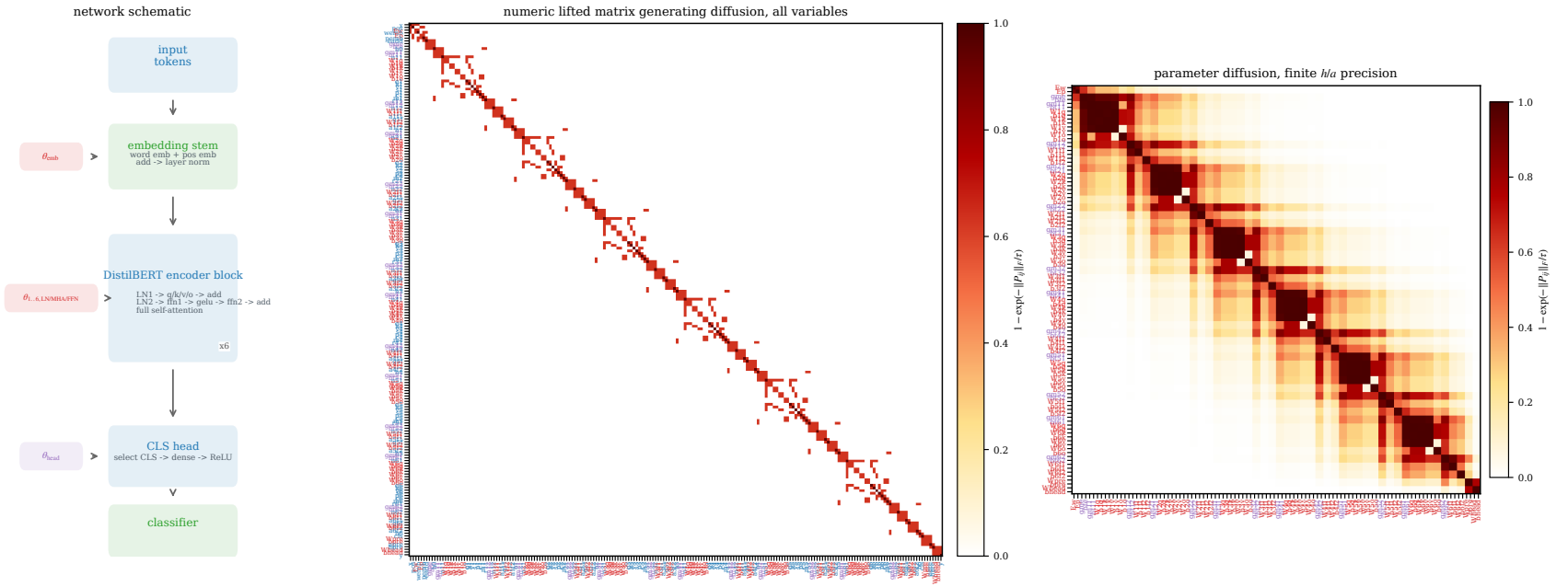



Figure A.2: **Whole-network transformer witness of the lifted-to-reduced transition.** Left: schematic of a DistilBERT-style classifier with token embedding and layer normalization, a stack of six encoder blocks with full attention, and a classifier head. Middle: numeric lifted operator over all variables, again ordered according to the schematic. The six encoder blocks appear as repeated local motifs, reflecting residual-stream, attention, and feed-forward interactions that remain sparse in lifted coordinates. Right: the corresponding parameter-only operator after eliminating hidden or activation variables at finite hidden/activation precision. The resulting parameter couplings are much denser and organized by block structure: embeddings and head couple to the stack, and repeated encoder blocks induce broader off-diagonal interactions after elimination. This is the whole-network analogue of Figure 1(c).

# B Experimental diagnostics and ablations

Section 5 presents the main diagnostic figures and table. This appendix gives measurement definitions, calibration checks, and architecture ablations. The first subsection expands the scalar optimizer-step, sampled upper-edge, active-plasticity, and directional-transport diagnostics. The second describes gauges and tangents, retained stochastic geometry, and the separate split-based prior-KL protocol with its matching fresh-run audit table.

## B.1 Criticality diagnostics and architecture ablations

This subsection supports Section 5.1. The scalar checkpoint-linearized model has an upper stability condition and an active-curvature condition; the experiments evaluate calibrated checkpoint diagnostics of these two quantities. The main text shows the compact operating-regime view in Figure 2. Here we give the measurement definitions, calibration checks, and architecture ablations behind that figure.

**Checkpoint quantities.** All quantities are computed at saved checkpoints. Let $\theta_t$ be the checkpoint parameters, $g_t$ the current training gradient, and $\Delta\theta_t$ the next optimizer update reconstructed from the saved optimizer state. For SGD with momentum, $\Delta\theta_t$ is reconstructed from the momentum buffer; for AdamW, it is reconstructed from the saved first- and second-moment state. The scalar optimizer coordinate plotted in the upper band is the projection of the actual update onto the current gradient,

$$\eta_{\parallel,t} = -\frac{\Delta\theta_t^\top g_t}{\|g_t\|_2^2}.$$

This is the checkpoint's state-dependent step along the current training direction. It is not the nominal scheduler learning rate and not an asymptotic momentum multiplier. The projection $\eta_{\parallel,t}$ summarizes the parameter update along $g_t$; it does not retain every component of the parameter update or the optimizer-state dynamics.

The diagnostic uses an uncentered observed-gradient Gram matrix in the stated optimizer coordinates. Let $P_t$ be the local optimizer metric, $\tilde{g}_t^{(s)} = P_t^{1/2} g_t^{(s)}$, and let $\tilde{G}_t \in \mathbb{R}^{m\times p}$ have rows $(\tilde{g}_t^{(s)})^\top$. The sampled empirical-Fisher-type operator $\widehat{F}_t = m^{-1}\tilde{G}_t^\top \tilde{G}_t$ and the smaller Gram matrix $\widehat{K}_t = m^{-1}\tilde{G}_t\tilde{G}_t^\top$ have exactly the same nonzero eigenvalues. We report the corresponding scalar proxy

$$\widehat{\eta}_{\mathrm{edge},t} = \frac{2}{\lambda_{\max}(\widehat{K}_t)}.$$

For the scalar heavy-ball proxy used in the SGD-momentum vision runs, the same Gram spectral scale is reported in the corresponding momentum-adjusted effective-step units. The black and colored curves use the same scalar coordinate. This is a proxy for the Fisher/GGN stability edge of the stated local linearized model; it is not a direct computation of that GGN or of the complete momentum/AdamW stability region.

The lower-side measurement is the active-plasticity coordinate. Let $U_t \in \mathbb{R}^{p\times r}$ have orthonormal columns spanning a small active subspace, chosen from recent or probe gradient directions. We form the projected curvature

$$\widehat{\widehat{F}}_{\mathrm{act},t} = U_t^\top \widehat{\widehat{F}}_t U_t, \qquad \rho_{\mathrm{act},t} = \eta_{\mathrm{eff},t}\lambda_{\min}(\widehat{\widehat{F}}_{\mathrm{act},t}).$$

Here $\eta_{\mathrm{eff},t}$ is the same actual checkpoint optimizer coordinate used in the upper-band comparison. The quantity $\rho_{\mathrm{act},t}$ is a plasticity coordinate: it asks whether the directions currently used by the optimizer have nontrivial local curvature. It is not a second learning-rate boundary and is not an estimate of the global smallest eigenvalue.

The architecture mechanism diagnostic is backward gain. For retained architecture boundaries $h_{0,t}, \ldots, h_{L,t}$, we backpropagate the training loss $L_t$ and record

$$b_{\ell,t} = \frac{\mathrm{RMS}(\partial L_t/\partial h_{\ell,t})}{\mathrm{RMS}(\partial L_t/\partial h_{\ell+1,t})}.$$

A gain near one means that the boundary transports backward signal close to neutrally in the measured direction; sustained gains above or below one indicate amplification or attenuation. This is a directional VJP proxy for the separator-transfer mechanism in Theorem 2, not a full operator-norm separator proof.

**Estimator status and calibration.** The primitive gradient averages are standard sampling estimates for the chosen checkpoint probe distribution. The spectral quantities derived from them, including largest eigenvalues, retained eigenspaces, and active-subspace minima, are nonlinear functions of the sampled empirical-Fisher-type operator or centered covariance estimate. We treat them as calibrated diagnostics; the Gram spectral identity does not make them unbiased estimators of the corresponding population quantities. Their variance depends on probe count, effective rank, eigengaps, block choice, and the active-subspace construction.

We use three calibration checks. First, when possible, we run state-matched LR-cliff sweeps: starting from the saved weights and optimizer state, we multiply the checkpoint update by trial factors and bracket the onset of local instability. The diamonds in the upper-band plots mark these observed instability brackets. Second, active-subspace random controls check that the measured lower-side signal is not a generic random-subspace artifact. Third, architecture ablations check whether changes in residual paths, normalization, width, compression, and recurrent/state carriers move the diagnostic coordinates in the predicted direction.

**Frozen-checkpoint moment calibration.** Starting from identical submitted ResNet-32 weights and Nesterov optimizer state, we ran $256$ independent stochastic continuations. The one-step empirical mean-update direction had cosine $0.975714$ with the optimizer-aware local Nesterov map, compared with $0.512519$ under the raw-gradient drift convention. The empirical one-step retained covariance trace was $0.386292$ of the optimizer-aware prediction and $0.264958$ of the stationary-momentum convention used in the submitted diagnostic.

Including the saved optimizer state substantially improved the one-step mean-direction prediction at this checkpoint. The covariance-trace ratios show a substantial scale mismatch even at one step. These measurements calibrate local first- and second-moment predictions; they do not establish an Itô limit or a metric log-Sobolev inequality.

**Upper-side ablations.** Figure B.1 extends the main upper-band plot to architecture edits in GPT-20M, a selective diagonal state-space model (SSM), ResNet-18, DenseNet-BC, and WideResNet-16. The upper-edge proxy is a scalar stability diagnostic based on the sampled gradient Gram; it does not describe the full optimizer-state stability region. The actual optimizer step is reconstructed from AdamW or SGD-momentum state and plotted in the same units. Across these ablations, architecture and optimizer state jointly move the local upper edge. The figure also shows why the upper side alone is not a complete training story: some residual-removal runs stay below the upper edge while losing the lower-side active signal in Figure B.2.

**Lower-side ablations.** Figure B.2 shows the active-plasticity coordinate across the same architecture edits. The clearest failure mode is residual removal. In GPT-20M and in the SSM run, no-residual variants remain below the scalar upper-edge proxy while their active coordinate collapses. This is consistent with the theorem's distinction between upper stability and active plasticity; the proxy measurements alone do not establish full optimizer stability. Vision ablations show a broader operating map. Residual scale, normalization, DenseNet compression, growth rate, and width move the same lower-side coordinate, revealing how architectural carriers change the amount of trainable signal in the active subspace.

**Architecture transport.** Figure B.3 localizes the transport mechanism behind the scalar upper and lower coordinates. In GPT, the baseline, scaled-residual, and RMSNorm variants keep most block gains near the neutral value after early blocks, while no-normalization and no-residual variants show stronger deviations. In the SSM run, the residual carrier again controls whether transport remains near neutral. In ResNet and WideResNet, residual scale and normalization change where backward signal spikes or attenuates. In DenseNet, gains stay close to one across many layers, with periodic dips at transition-like boundaries, matching the concatenative-carrier interpretation. The plot connects the scalar checkpoint measurements to the separator-transfer theorem: architecture changes trainability by changing the local interfaces through which curvature and backward sensitivity propagate.

**Computation and scalability.** The dense objects suggested by the theorems are not formed. With $p$ parameters and $m$ probes, the upper-edge spectral problem is $m \times m$, not $p \times p$. The Gram matrices can be formed blockwise or streamed. Active plasticity uses a small $r \times r$ projected problem on a recent or probe-gradient span. Backward gain uses boundary-gradient summaries

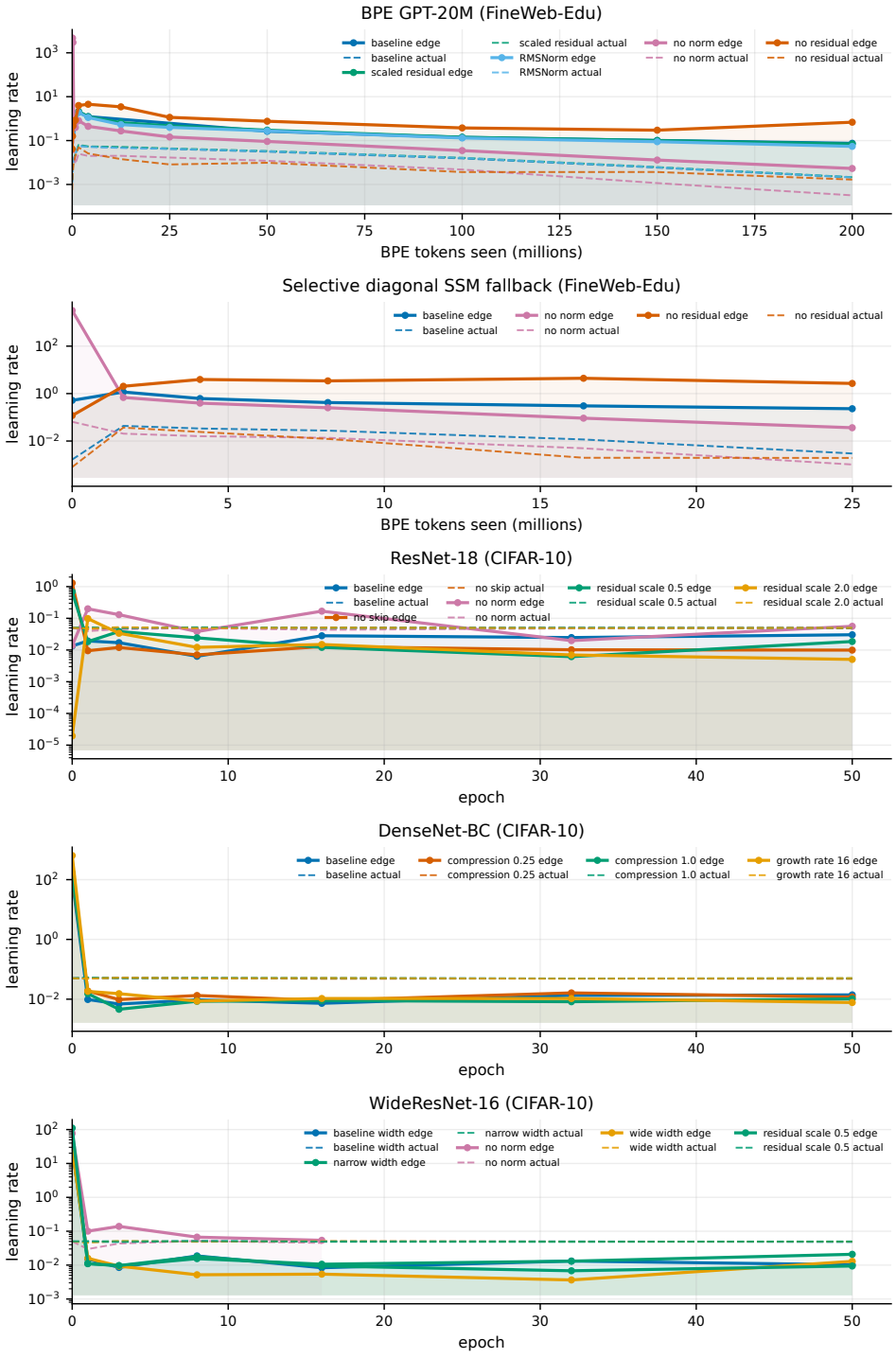



Figure B.1: **Unified upper-stability bands across GPT, SSM, and vision ablations.** Solid curves show sampled-Gram upper-edge proxies; dashed curves show scalar update projections reconstructed from AdamW or SGD-momentum state. The panels compare architecture edits in GPT-20M, a selective diagonal SSM, ResNet-18, DenseNet-BC, and WideResNet-16. The comparison gives a checkpoint diagnostic motivated by the upper side of Theorem 3, rather than a complete stability guarantee for the optimizer-state dynamics.

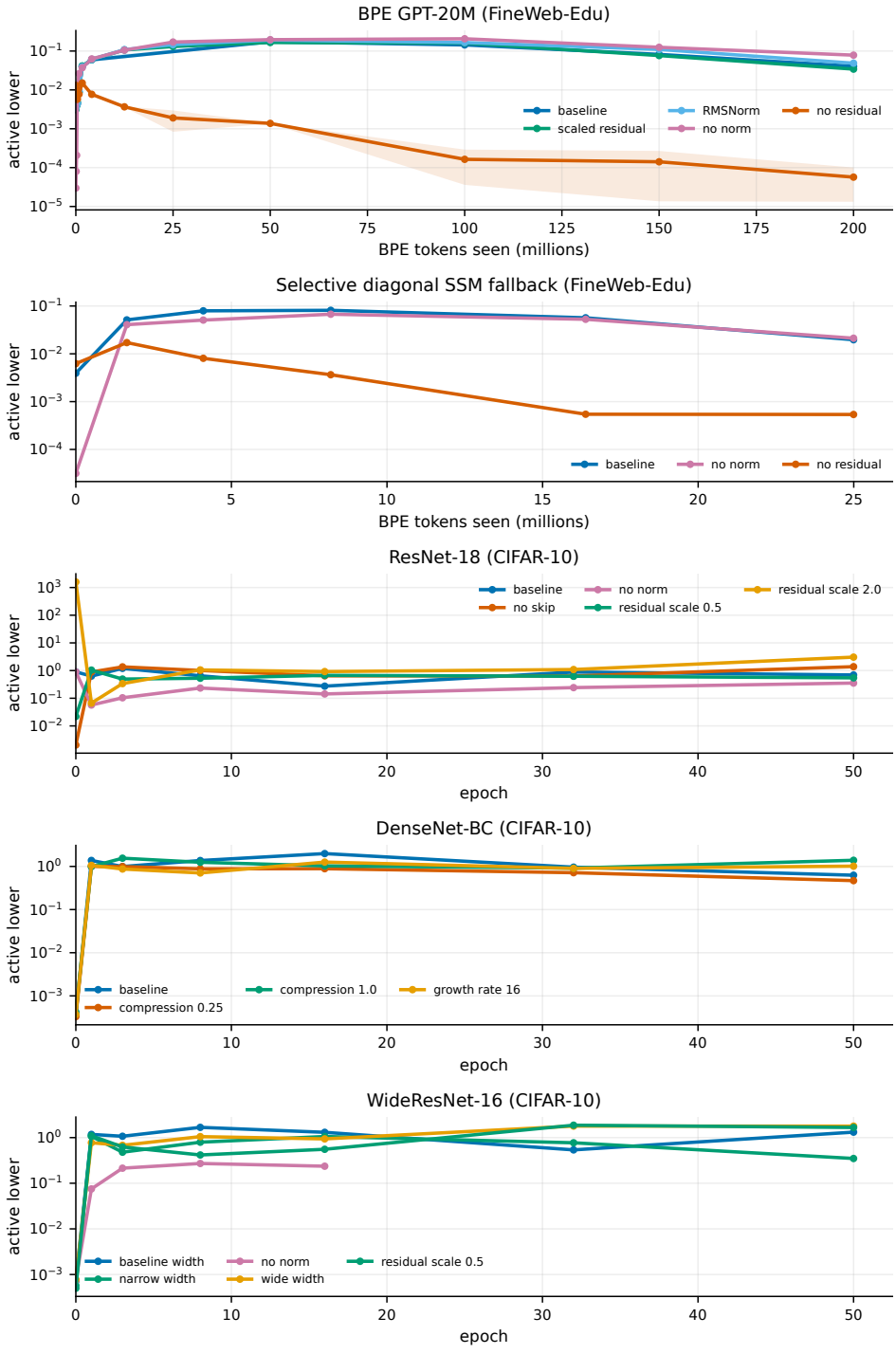



Figure B.2: **Unified active-plasticity signal across GPT, SSM, and vision ablations.** The curves plot the scalar optimizer-step coordinate times the minimum eigenvalue of the sampled operator on the active gradient span. In GPT and the state-space model, residual-path removal collapses this active signal while the scalar step remains below the upper-edge proxy. Vision ablations show how residual scale, normalization, DenseNet compression, growth rate, and width move the same lower-side measurement.

from ordinary backpropagation. For large models, the practical artifacts are checkpoint summaries, small Gram matrices, retained active-subspace coefficients, LR-cliff brackets when available, and boundary-gradient statistics.

The ablations demonstrate the usefulness of the two-sided criticality theorem. The scalar upper-edge diagnostic and active plasticity can move independently: GPT and SSM residual removal keep the scalar step below the upper-edge proxy while collapsing the active signal. Architectural carriers shape the tradeoff: residual scale, normalization, DenseNet compression, growth rate, and width change both the stable-step edge and the active directions that remain trainable. The backward-gain curves then show where these changes enter the computation graph. Well-behaved residual, normalization, and concatenative carriers keep gains near neutral over much of the network; component edits create deviations that later appear as changes in maximum stable step or active plasticity.

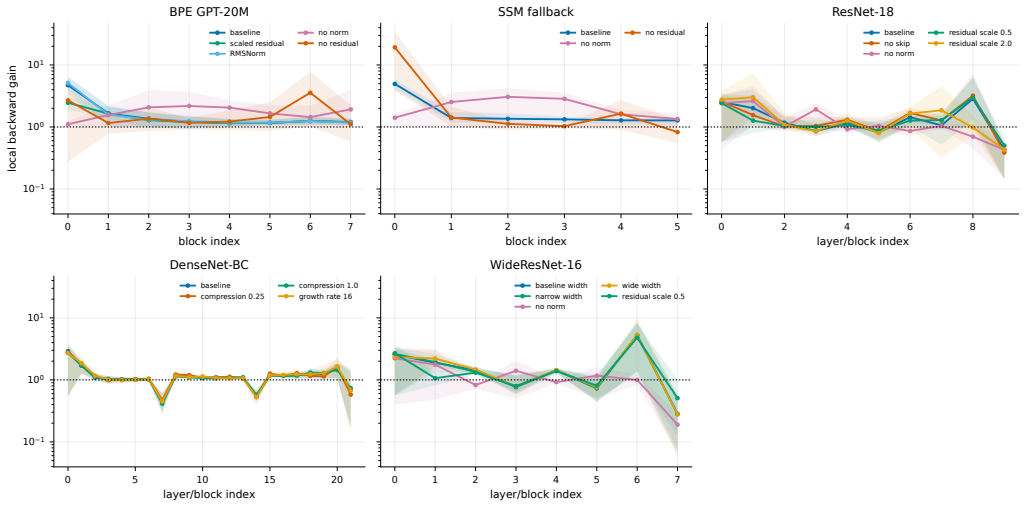



Figure B.3: **Unified backward gain by layer or block.** Each curve reports a local backward gain ratio at retained architecture boundaries, with the neutral value marked at one. Gains near one indicate neutral signal transport through the corresponding boundary; sustained deviations indicate amplification or attenuation. The plot localizes how architecture modifications change the transport mechanism behind the upper-stability and active-plasticity measurements.

### B.2 Localized KL/PAC-Bayes diagnostics: gauges, tangents, and retained probability space

This subsection supports Section 5.2. The resolved KL theorem compares an assumed local Itô law with a localized Gibbs reference on a fixed comparison space $E$. The main text shows the compact ResNet/CIFAR-10 geometry in Figure 3 and the certificate rows in Table 1. Here we make the directional probes, retained-space construction, posterior-risk audit, and scalability assumptions explicit.

**Directional diagnostics.** Let $f_\theta(X)$ be the vector of logits on a measured data batch $X$, and let $J_f(\theta; X)$ be its parameter Jacobian. A structural gauge is a parameter displacement $v$ for which $J_f(\theta; X)v$ is zero or small on the measured data, so the parameter coordinates move while the represented function is nearly unchanged. ReLU scale symmetries and BatchNorm/ReLU scale candidates are the structural probes used here.

A loss-space gauge is an output-space redundancy that is nearly invisible to the loss. For softmax cross-entropy, the common-logit direction is the clean example: shifting all logits by the same scalar does not change probabilities. A loss tangent is a direction whose logit movement has low curvature for the bounded loss. If $q_v = J_f(\theta; X)v$ is the output tangent and $H_\ell$ is the output-space Hessian of the bounded loss, then low-loss tangency means that $\|H_\ell^{1/2} q_v\|$ is small compared with $\|q_v\|$. These directions may move raw logits while changing probabilities, clipped loss, or classification error very little.

An augmentation tangent compares parameter movement with data-space movement. For an augmentation $T$, such as crop, color, MixUp, or CutMix, we compare $J_f(\theta; X)v$ with the centered logit response $f_\theta(TX) - f_\theta(X)$. High cosine similarity indicates that a local parameter displacement resembles a data-transformation response. These directions are not parameter-space gauges; they are probes of how augmentation changes the local train/validation geometry.

These probes are diagnostic overlays. The retained PAC-Bayes comparison space is selected by the blockwise covariance spectrum, not by the semantic labels. The probes explain what kinds of directions appear in the weakly explored tail or in the retained geometry.

**Spectrum map and augmentation geometry.** Figure B.4 gives a larger version of the spectrum map used in Figure 3(c). The gray curve is the positive blockwise minibatch-gradient covariance spectrum for ResNet-32/CIFAR-10 under standard crop/flip and modern MixUp/CutMix. Low-rank positions on the left are weakly explored directions; high-rank positions on the right are directions where stochastic training injects more local variation. The overlays locate candidate structural gauges, loss gauges, loss tangents, augmentation tangents, residual diagnostics, BN-affine eigendirections, convolution-kernel eigendirections, and MixUp/CutMix tangent halos. The map shows how weakly explored directions can dominate an inverse-metric defect calculation, while the retained coordinates describe directions with larger measured stochastic variation.

Figure B.5 gives the ordinary predictor-level companion to the retained-direction alignment plot in the main text. MixUp/CutMix changes the retained train/validation geometry while also improving the clean generalization pattern relative to standard crop/flip. The point is not that augmentation merely lowers a scalar defect; it changes which local directions align between clean train and validation losses.

**Weak directions: output/loss tangents rather than clean input transforms.** Figure B.6 gives a controlled MNIST example from a projected ReLU MLP. Each row shows one weak non-explicit-gauge direction. The finite logit and probability changes are small, and the closest simple input transforms have weak cosine similarity. A sequential logit-space decomposition makes the point quantitative: across the weak directions, the median softmax-loss tangent fraction is $0.758$, the median low-loss input-transform fraction after removing the softmax-loss component is $0.0008$, the median all-transform fraction is $0.0021$, and the median residual-other fraction is $0.241$. Thus these weak directions are primarily output/loss tangents, not clean input-transformation tangents.

Figure B.7 gives the corresponding ResNet/CIFAR-10 audit. Each panel compares the class-probability change induced by a selected parameter direction with the closest input or augmentation response. The examples make the mechanism claims auditable: weak directions can move logits, but their finite probability changes are small, and simple input transformations only partially explain the

movement. The augmentation-sensitive directions are therefore treated as data-space tangents, not as parameterization gauges.

**Split-based prior-KL certificate protocol.** Table B.1 gives the reported audit quantities for the same fresh runs as Table 1. Write

$$\ell_c(\theta; x, y) = \min\{-\log p_\theta(y \mid x)/c, 1\}, \qquad c > 0,$$

for normalized clipped-NLL, and denote the fresh-run cutoff for row $j$ by $c_{\mathrm{cert},j}$. These cutoffs are not identified with the historical diagnostic cutoff specified below.

Each of the five architectures was freshly trained using build data only. Before accessing any certificate record, we froze all five checkpoints, build-only covariance spaces $E$, retained dimensions $k = (1, 1, 16, 11, 8)$, explicit same-support Gaussian priors $P$ and posteriors $Q$ on the retained coordinates, confidence allocations, and $4{,}096$ posterior draws per row. The same frozen $Q$ is used for its exact Gaussian $\mathrm{KL}(Q\|P)$ and posterior-risk evaluation.

Conditional on build data and the independent construction randomness, the prior and construction are fixed relative to the independent certificate sample. Averaging the conditional PAC-Bayes guarantee over build data and construction randomness gives the joint statement. The reported per-row allocation is $\delta_{\mathrm{PB}} = 0.008$ and $\delta_{\mathrm{MC}} = 0.002$; a union bound over the five rows gives simultaneous $95\%$ confidence, including Monte Carlo error. Test data were used only after the certificate-only table had been finalized. These bounds use the prior-KL form of Corollary 10, not $K_{\mathrm{stat}}$, the Itô/LSI assumptions of Theorem 4, or Eq. (E.4).

*Documentation status of the reported split runs.* The available five-model summary does not specify the row-specific cutoffs $c_{\mathrm{cert},j}$, build/certificate sample counts, empirical posterior losses and their Monte Carlo corrections, full Gaussian parameters, or executed concentration and Monte Carlo procedures. These settings and the raw outputs remain to be documented for independent numerical verification; the original $c = 5$ diagnostic convention does not determine the fresh-run cutoffs.

**Relation to the original checkpoint diagnostics.** The original risk/score audit used $\ell_5$, namely $c = 5$. Its retained geometry was formed by choosing architecture-native blocks, sampling stochastic gradients, centering them by a training-drift estimate, and diagonalizing the corresponding block Gram matrices. With $m$ sampled gradients in a block of dimension $p_B$, the empirical covariance is $m^{-1}G_B^\top G_B$; its nonzero spectrum is obtained from the smaller $m \times m$ matrix $m^{-1}G_B G_B^\top$. The earlier stationary score used a frozen-checkpoint defect proxy as its complexity input. This proxy was not established as a prior-KL upper bound, and the original calculation did not execute Eq. (E.3) or Eq. (E.4). That score is therefore not used for the fresh split certificates in Tables 1 and B.1.

Table B.1: **Reported split-certificate audit quantities.** All rows refer to the same fresh build/certificate runs as Table 1. The Gaussian KL is evaluated for the same posterior used for risk estimation. The confidence budgets and $4{,}096$ posterior draws per row were frozen before certificate access; the five per-row budgets sum to $0.05$. Population bounds and posterior test risks are in normalized clipped-NLL units. Test data were accessed only after the certificate-only table had been finalized.

| Model | $k$ | $\mathrm{KL}(Q\|P)$ | $\delta_{\mathrm{PB}}$ | $\delta_{\mathrm{MC}}$ | Draws | Population bound | Posterior test risk |
|---|---|---|---|---|---|---|---|
| MLP | 1 | 0.31815 | 0.008 | 0.002 | 4096 | 0.04258 | 0.02198 |
| LeNet | 1 | 0.31815 | 0.008 | 0.002 | 4096 | 0.02983 | 0.00898 |
| ResNet-32 | 16 | 5.09035 | 0.008 | 0.002 | 4096 | 0.07657 | 0.05840 |
| WRN-16-4 | 11 | 3.49962 | 0.008 | 0.002 | 4096 | 0.06349 | 0.04448 |
| DenseNet-small | 8 | 2.54518 | 0.008 | 0.002 | 4096 | 0.07669 | 0.06100 |

**Estimator status and scalability.** The matrix $m^{-1}G_B^\top G_B$ is the empirical covariance of the centered probes. When the centering mean is estimated from these same $m$ independent probes, the usual unbiased covariance estimator instead uses $G_B^\top G_B/(m-1)$. The Gram spectral identity holds with either normalization when the same normalization is used in both matrices. The reported spectral quantities—eigenvalues, retained eigenspaces, retained-prefix choices, and inverse-metric defect coefficients—are nonlinear functions of the sample covariance. We therefore treat them as

calibrated diagnostics rather than unbiased scalar estimators. Their variance depends on probe count, effective rank, eigengaps, block choice, and the retained-prefix rule.

The computation avoids dense Hessian, Fisher, and covariance matrices. With $p_B$ parameters in a block and $m$ probes, the spectral problem is $m \times m$, not $p_B \times p_B$, and the Gram matrices can be formed blockwise or streamed. Posterior sampling is restricted to the retained directions, and the unresolved complement is shared by posterior and reference. For larger models, the persistent artifacts are checkpoint summaries, small block Gram matrices, retained eigenpairs, defect coefficients, posterior-risk samples, and validation/test audit summaries. The resolved-space construction therefore scales as a checkpoint diagnostic: it asks which local probability directions training has actually explored, rather than attempting a dense full-parameter posterior comparison.

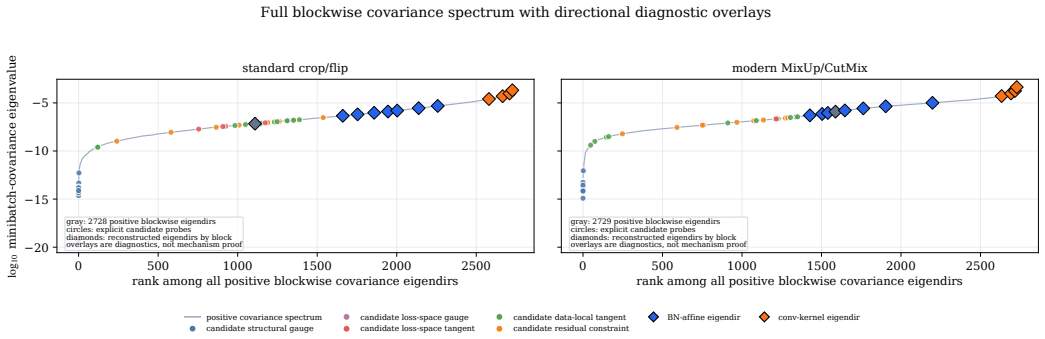



Figure B.4: **Full ResNet/CIFAR-10 blockwise covariance spectrum with diagnostic overlays.** The gray curve shows all positive blockwise covariance eigendirections under standard crop/flip and MixUp/CutMix. Circles mark explicit candidate probes; diamonds mark reconstructed blockwise eigendirections; magenta halos show MixUp/CutMix tangent similarity. The overlays are diagnostics rather than mechanism proofs: they place gauges, loss tangents, residual probes, and augmentation-aligned directions inside the same local covariance geometry used by the retained KL/PAC-Bayes construction.

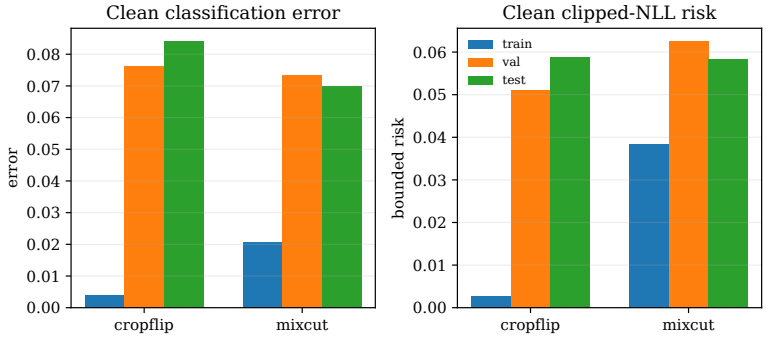



Figure B.5: **Clean risks for the ResNet-32 augmentation comparison.** The panels report clean classification error and clean clipped-NLL risk under standard crop/flip and MixUp/CutMix. Together with the retained-direction train/validation alignment plot in Figure 3, this shows that augmentation changes both predictor-level generalization and the retained local probability geometry.

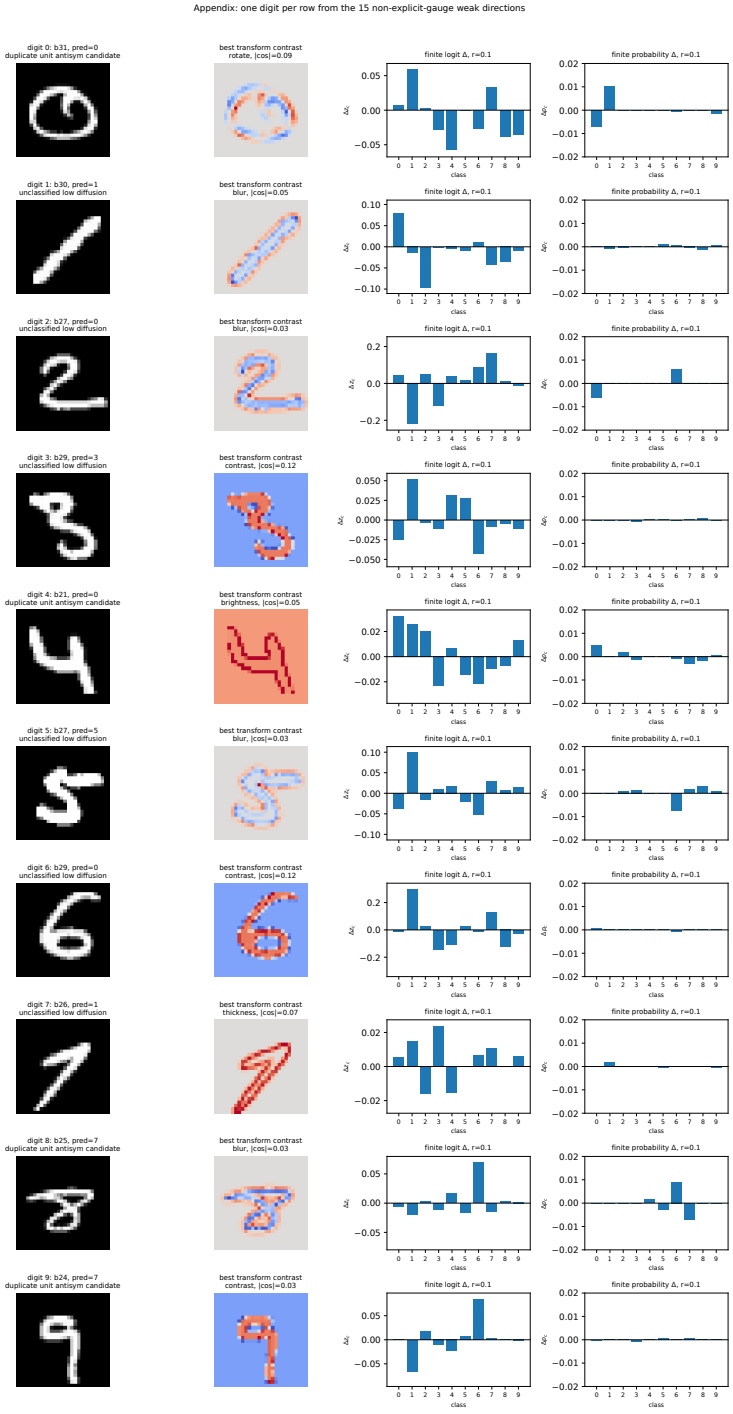



Figure B.6: **MNIST weak-direction examples.** Each row shows one weak non-explicit-gauge direction. The finite logit and probability changes remain small at radius 0.1, and the closest tested input transforms have weak cosine similarity. The examples support the output/loss-tangent interpretation of weak unresolved directions.

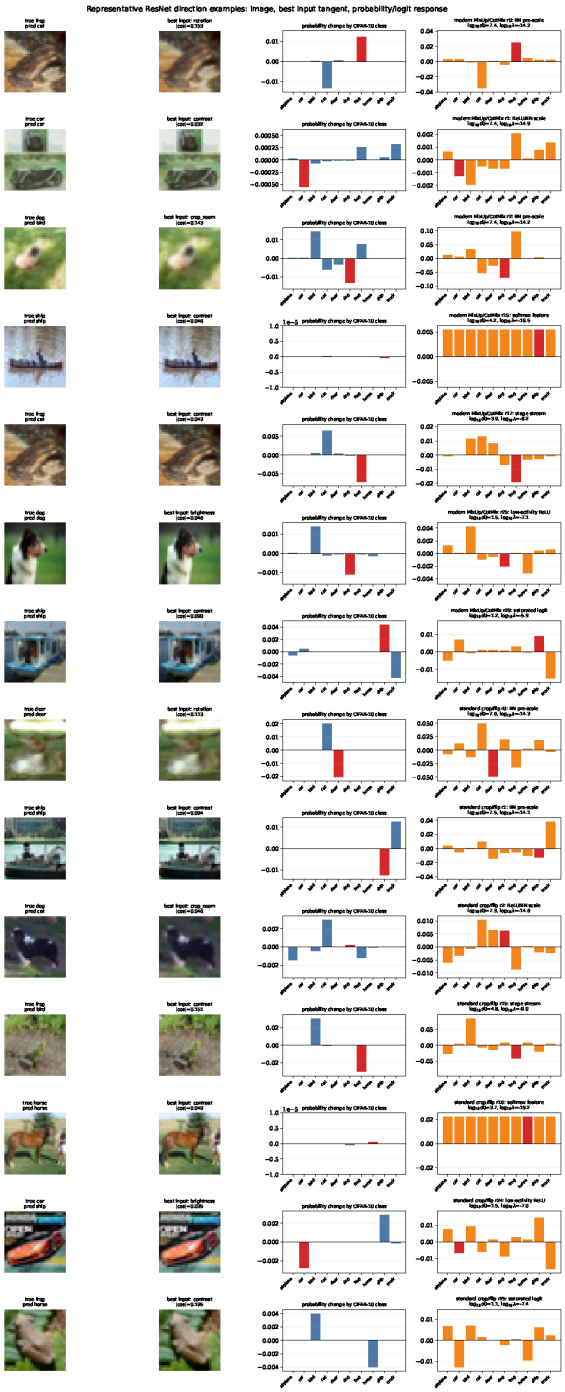



Figure B.7: **ResNet/CIFAR-10 direction examples.** Each panel shows the finite class-probability change induced by a selected weak or diagnostic direction, together with the closest input or augmentation response. The examples support the output-space, loss-tangent, and augmentation-tangent interpretation of directions that contribute strongly to the inverse-metric defect while having small finite predictive effect.

## C Activation elimination and Gaussian message passing

This appendix proves Theorem 1. The theorem connects the checkpoint-local lifted precision $P_{\theta,u}$ from Section 2 to the reduced parameter curvature $F_{\mathrm{red}}$. There are two equivalent views. Directly eliminating the intermediate states $u$ gives the Schur complement (5). If the lifted factor graph is represented as a junction tree, the same Schur complement is obtained by local Gaussian messages over cliques and separators. The first view identifies the reduced Fisher/Gauss–Newton curvature in the hard-constraint limit; the second view explains why architecture interfaces are the places through which reduced interactions are mediated.

**Checkpoint-local information model.** We work with local displacements $\Delta z = (\Delta\theta, \Delta u)$ around a checkpoint $z_0 = (\theta_0, u_0)$. For each factor $f$ in the lifted potential, Section 2 uses the checkpoint-local first variation of that factor. With fixed base weight $W_f \succeq 0$, write $Q_f = J_f^\top W_f J_f$ on the factor scope $S_f$. Module factors contribute $\tau^{-1}Q_f$, while data-loss factors contribute $Q_f$. The linear information term affects the reduced information vector but not the reduced precision; we include it below only to state the Gaussian message formula.

Embedding the scaled factor contributions into the full lifted variable ordering gives the information-form energy

$$\mathcal{E}_\tau(\Delta z) = \frac{1}{2}\Delta z^\top P_{\theta,u}(\tau)\Delta z - h_\tau^\top \Delta z + \mathrm{const},$$

where $P_{\theta,u}(\tau) = P_{\theta,u}(z_0;\tau)$ is the lifted precision in (4). The weighted-module uniqueness assumption in Theorem 1 makes the state-state block positive definite for every $\tau > 0$. The hard-constraint limit is $\tau \downarrow 0$, as the module precision $\tau^{-1}$ increases. The local quadratic model has Gaussian information form with precision $P_{\theta,u}(\tau)$.

**Theorem 1 (Activation elimination recovers reduced Fisher/GGN, restated).**

*At a forward-consistent checkpoint $z_0$, let $P_{\theta,u}(z_0;\tau)$ be the lifted precision in* (4)*. Assume the weighted linearized module equations $W_f^{1/2}(J_{f,\theta}\Delta\theta + J_{f,u}\Delta u) = 0$, $f \in \mathcal{F}_n^{\mathrm{mod}}$, determine a unique $\Delta u$ for every $\Delta\theta$. For $\tau > 0$, eliminating intermediate states gives*

$$F_\tau = P_{\theta\theta}(\tau) - P_{\theta u}(\tau)P_{uu}(\tau)^{-1}P_{u\theta}(\tau).$$

*As $\tau \downarrow 0$, $F_\tau$ increases in Loewner order to the reduced Fisher/GGN curvature $F_{\mathrm{red}}$. On a junction-tree representation, the same Schur complement is computed by local Gaussian elimination over cliques and separators while retaining all parameter variables.*

**Proof sketch.** The proof has two steps. First, eliminating the state variables at finite $\tau$ gives the Schur complement (5). Increasing the module precision enforces the weighted linearized module equations, and the quadratic-penalty limit gives $F_\tau \uparrow F_{\mathrm{red}}$. Second, state-only clique–separator elimination computes the same finite-$\tau$ Schur complement. The running-intersection property ensures that composing these local eliminations gives the global reduction while retaining the parameter variables.

**Complete proof of Theorem 1.**

*Proof.* Partition $P_{\theta,u}(\tau)$ and $h_\tau$ as

$$P_{\theta,u}(\tau) = \begin{pmatrix} A & B \\ B^\top & D \end{pmatrix}, \qquad h_\tau = \begin{pmatrix} h_\theta \\ h_u \end{pmatrix},$$

where $A = P_{\theta\theta}$, $B = P_{\theta u}$, and $D = P_{uu}$. The local information energy is

$$\mathcal{E}_\tau(\Delta\theta, \Delta u) = \frac{1}{2}\Delta\theta^\top A\Delta\theta + \Delta\theta^\top B\Delta u + \frac{1}{2}\Delta u^\top D\Delta u - h_\theta^\top\Delta\theta - h_u^\top\Delta u + \mathrm{const}.$$

For $\tau > 0$, $D$ is invertible. Completing the square in $\Delta u$ gives

$$\begin{aligned}\mathcal{E}_\tau(\Delta\theta, \Delta u) = {} & \frac{1}{2}\left\|\Delta u + D^{-1}(B^\top\Delta\theta - h_u)\right\|_D^2 \\ & + \frac{1}{2}\Delta\theta^\top(A - BD^{-1}B^\top)\Delta\theta - (h_\theta - BD^{-1}h_u)^\top\Delta\theta + \mathrm{const}.\end{aligned}$$

Thus both minimization over $\Delta u$ and Gaussian integration over $\Delta u$ produce the same reduced parameter precision,

$$F_\tau = A - BD^{-1}B^\top = P_{\theta\theta} - P_{\theta u}P_{uu}^{-1}P_{u\theta},$$

which is (5). The reduced information vector is $h^{\mathrm{red}}_{\theta,\tau} = h_\theta - BD^{-1}h_u$; it specifies the reduced linear term and is not needed for the curvature statement.

For the hard-constraint limit, retain only the homogeneous quadratic part, since linear information terms do not affect the reduced precision. Let $R_\theta$ and $R_u$ be the stacked weighted module Jacobians, whose block rows are $W_f^{1/2}J_{f,\theta}$ and $W_f^{1/2}J_{f,u}$. The hypothesis gives, for every $\Delta\theta$, a unique $\Delta u_*$ satisfying $R_\theta\Delta\theta + R_u\Delta u_* = 0$. In particular, $R_u$ has full column rank, so $P_{uu}(\tau) \succ 0$ for every $\tau > 0$. Write the unpenalized quadratic as

$$\mathcal{Q}_0(\Delta\theta, \Delta u) = \frac{1}{2}\Delta\theta^\top\Lambda_\theta\Delta\theta + \frac{1}{2}\sum_{f\in\mathcal{F}_n^{\mathrm{data}}} \|J_{f,\theta}\Delta\theta + J_{f,u}\Delta u\|^2_{W_f} \geq 0.$$

The finite-$\tau$ Schur complement has the variational form

$$\frac{1}{2}\Delta\theta^\top F_\tau\Delta\theta = \min_{\Delta u}\left\{\mathcal{Q}_0(\Delta\theta,\Delta u) + \frac{1}{2\tau}\|R_\theta\Delta\theta + R_u\Delta u\|_2^2\right\}.$$

The right-hand side increases as $\tau$ decreases and is bounded above by $\mathcal{Q}_0(\Delta\theta, \Delta u_*)$. If $\Delta u_\tau$ is its minimizer, nonnegativity of $\mathcal{Q}_0$ gives

$$\|R_u(\Delta u_\tau - \Delta u_*)\|_2^2 \leq 2\tau\mathcal{Q}_0(\Delta\theta,\Delta u_*).$$

Full column rank of $R_u$ implies $\Delta u_\tau \to \Delta u_*$. Continuity of $\mathcal{Q}_0$, together with

$$\mathcal{Q}_0(\Delta\theta,\Delta u_\tau) \leq \frac{1}{2}\Delta\theta^\top F_\tau\Delta\theta \leq \mathcal{Q}_0(\Delta\theta,\Delta u_*),$$

therefore proves convergence to the constrained quadratic. Substituting the unique linearized forward state $\Delta u_*$ into the output-loss factors gives the reduced Jacobians $J_i^{\mathrm{red}}$ and

$$\mathcal{Q}_0(\Delta\theta,\Delta u_*) = \frac{1}{2}\Delta\theta^\top F_{\mathrm{red}}\Delta\theta, \qquad F_{\mathrm{red}} = \Lambda_\theta + \frac{1}{|B|}\sum_{i\in B}(J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}},$$

with the data-loss weights normalized as in the empirical objective. Hence $F_\tau \uparrow F_{\mathrm{red}}$ in Loewner order. In finite dimensions, convergence of these quadratic forms also gives matrix convergence. This proves the direct elimination and hard-constraint claims.

It remains to show that local Gaussian messages implement the same Schur complement. Use a state-elimination order that retains all parameter variables, and a compatible junction-tree representation with cliques $C$ and separators $S = C \cap \mathrm{pa}(C)$. Retained parameter variables are included in separators wherever necessary, so the internal variables eliminated below are state variables only. Each clique carries a local information pair $(Q_C, h_C)$, obtained by summing the factor contributions assigned to that clique. After receiving messages from its children, clique $C$ forms

$$Q_C^{\mathrm{tot}} = Q_C + \sum_{C'\in\mathrm{ch}(C)} Q_{C'\to C}, \qquad h_C^{\mathrm{tot}} = h_C + \sum_{C'\in\mathrm{ch}(C)} h_{C'\to C}.$$

Partition the clique variables into separator variables $S$ and internal variables $I = C \setminus S$. Local elimination of $I$ sends the Gaussian message

$$Q_{C\to S} = Q_{SS}^{\mathrm{tot}} - Q_{SI}^{\mathrm{tot}}(Q_{II}^{\mathrm{tot}})^{-1}Q_{IS}^{\mathrm{tot}}, \qquad h_{C\to S} = h_S^{\mathrm{tot}} - Q_{SI}^{\mathrm{tot}}(Q_{II}^{\mathrm{tot}})^{-1}h_I^{\mathrm{tot}}. \tag{C.1}$$

The scalar normalizing constant is omitted because it does not affect the precision.

The proof that these messages compute the global Schur complement is by induction on the junction tree. For a leaf clique, (C.1) is exactly the Schur complement obtained by eliminating variables that appear only inside that leaf subtree. For an internal clique, the induction hypothesis says that each child message is already the exact Schur complement of the eliminated variables in that child subtree. Adding those messages to the clique information pair therefore gives the exact information form on the variables remaining in the union of the clique and its processed child subtrees. Applying (C.1) then eliminates the internal variables of the current subtree. The running-intersection property

guarantees that once a variable is eliminated across a separator, it never reappears outside the separator path. Induction to the root gives the same reduced precision as eliminating all intermediate states from the global information matrix.

If all parameter variables are retained, and all intermediate state variables are eliminated, the root information form is the reduced parameter precision $F_\tau$. Taking the same $\tau \downarrow 0$ limit gives $F_{\mathrm{red}}$. Thus local Gaussian message passing and direct Schur complement elimination are two computations of the same activation-eliminated curvature. □

**Discussions.** Activation elimination has both a geometric and an information-form interpretation. At finite $\tau$, minimization and Gaussian integration over the state variables give the same reduced precision $F_\tau$. The hard-constraint limit $F_\tau \uparrow F_{\mathrm{red}}$ recovers the curvature used in the subsequent analysis.

The proof also explains why architecture interfaces matter later. The parameter interaction $F_{\mathrm{red}}$ may connect distant modules after elimination, but that interaction is mediated by clique–separator messages. Section 3 turns the corresponding transfer operators into upper and lower criticality certificates. Section 4 uses the lifted/reduced precision as the Gibbs reference geometry for the resolved KL-to-Gibbs comparison. Thus the activation-elimination identity is the bridge from the computation graph to both the criticality theorem and the KL-bound theorem.

# D Separator-transfer proofs and architecture instantiations

This appendix proves the two criticality theorems from Section 3 and then instantiates them in standard architecture families. The proof structure is separated from the architecture catalog. The separator-transfer theorem describes the architecture mechanism: local interfaces pass perturbation and backward-sensitivity signal through the eliminated computation graph, possibly preserving, magnifying, or damping that signal. The two-sided criticality theorem then uses the resulting upper and lower curvature certificates to give a checkpoint-local statement about stable steps and active plasticity.

**Setup.** We use the lifted information model and activation elimination from Section 2 and Appendix C. In the hard-constraint limit of activation elimination, the reduced parameter curvature is

$$F_{\mathrm{red}} = \Lambda_\theta + \frac{1}{|B|} \sum_{i \in B} (J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}},$$

where $J_i^{\mathrm{red}}$ is the reduced Jacobian from parameters to the output/logit block of sample $i$, and $H_i \succeq 0$ is the output-space loss Hessian. The module-constraint state system below gives a local route to the reduced Jacobians. On a junction tree with bounded clique and separator dimensions, exact state elimination has bounded local solve cost per clique. The experiments use boundary-wise backward gains and sketched Gram probes as scalable diagnostics.

**Constructing the separator transfers.** Fix sample $i$ and the checkpoint, suppressing $i$ on local matrices when unambiguous. Stack the weighted module Jacobians as $R_u$ and $R_\theta$, as in Appendix C. The module hypothesis gives a unique solution of $R_u \Delta u + R_\theta \Delta\theta = 0$ for every $\Delta\theta$. Therefore

$$K := R_u^\top R_u \succ 0, \qquad H_v := -R_u^\top R_{\theta,v}, \qquad \Delta u = K^{-1} H_v \Delta\theta_v.$$

This state system contains only module constraints: data-loss and parameter-prior factors are excluded, and the common module precision $\tau^{-1}$ cancels from the state response. It is not the finite-$\tau$ state block of the complete loss-coupled precision.

Choose a junction tree for $K$, rooted at a clique $C_{\mathrm{out}}$ containing the required output states. Assign each module factor to a clique containing its state scope. In each clique's state-coordinate ordering, with local terms embedded as needed, set

$$Q_C = \sum_{f \mapsto C} J_{f,u}^\top W_f J_{f,u}, \qquad B_{v,i}^{(C)} = -\sum_{f \mapsto C} J_{f,u}^\top W_f J_{f,\theta_v}.$$

Thus $B_{v,i}^{(C)} \Delta\theta_v$ is the local information source; let $\mathcal{C}(v)$ contain its nonzero injection cliques. After accumulating child precision messages, write $Q_C^{\mathrm{tot}}$ for the clique precision. For a nonroot clique, order

coordinates as $(I, S)$, where $S$ is the parent separator and $I = C \setminus S$. The complete source-message map is

$$M_C = \begin{bmatrix} -Q_{SI}^{\mathrm{tot}}(Q_{II}^{\mathrm{tot}})^{-1} & I_S \end{bmatrix}, \qquad T_{C\to \mathrm{pa}(C)} = E_{S\to \mathrm{pa}(C)} M_C,$$

where $E_{S\to \mathrm{pa}(C)}$ embeds separator coordinates into the parent's clique ordering. Sources already on $S$ pass through the identity block. Empty internal sets use only that block. Positive definiteness of $K$ makes all eliminated state pivots and the final root precision positive definite. Let $O_i$ select the output displacement from root coordinates and define

$$\mathcal{R}_i = O_i (Q_{C_{\mathrm{out}}}^{\mathrm{tot}})^{-1}.$$

If $Z_{C,v}$ is the total source-response matrix after processing the subtree at $C$, Gaussian information elimination gives

$$Z_{C,v} = B_{v,i}^{(C)} + \sum_{D\in \mathrm{ch}(C)} T_{D\to C} Z_{D,v}.$$

Expanding this recursion and applying the root state response proves

$$J_{i,v}^{\mathrm{red}} = \mathcal{R}_i \sum_{C\in \mathcal{C}(v)} \left( \overrightarrow{\prod}_{e\in \mathrm{path}(C\rightsquigarrow C_{\mathrm{out}})} T_e \right) B_{v,i}^{(C)}.$$

Products compose maps from the injection clique toward the root; an empty product is the identity. All norms here are Euclidean in the declared source coordinates, with Frobenius norms for vectorized matrix-parameter blocks. Setting $\gamma_i = \|\mathcal{R}_i\|_{\mathrm{op}}$, $s_{v,i}^{(C)} = \|B_{v,i}^{(C)}\|_{\mathrm{op}}$, and $g_e = \|T_e\|_{\mathrm{op}}$ gives (7). The source-message maps alone are not primal forward Jacobians: the source injections and root response are essential.

**Direct forward construction.** For an explicit feedforward graph, linearizing each module gives $\Delta u_k = \sum_{j\in \mathrm{pa}(k)} A_{kj}\Delta u_j + \sum_v B_{kv}\Delta\theta_v$. Successive substitution in topological order expresses each output Jacobian as a sum of products along directed module paths. Operator-norm products, including the final output readout, give another verified budget. For the two-layer MLP of Appendix A, solving its three linearized module equations successively gives

$$\Delta o = W_2 D\phi(a_1)(\Delta W_1 x + \Delta b_1) + \Delta W_2 h_1 + \Delta b_2.$$

The module-only state solve gives the same output response, but its individual Gaussian source-message maps need not equal $D\phi(a_1)$ or $W_2$. The architecture examples below use direct forward-carrier budgets. The curvature argument requires only the verified inequalities $\|J_{i,v}^{\mathrm{red}}\|_{\mathrm{op}} \le S_{v,i}$, so either construction is admissible; no equality between their intermediate maps is asserted.

The upper part of the separator theorem controls how large the reduced Jacobian blocks can become. Since curvature is built from $(J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}}$, this gives an upper stability certificate. The lower part records an active-coverage condition: on a declared active subspace, the architecture and loss must still carry nonzero loss-changing signal.

**Theorem 2 (Separator-transfer certificates for reduced curvature, restated).**

*Fix the setup of Theorem 1 and use the module-constraint budgets* (7)*.*

***(i) Upper transfer.*** *For every sample $i$ and parameter block $v$,*

$$\|J_{i,v}^{\mathrm{red}}\|_{\mathrm{op}} \le S_{v,i}.$$

*If $0 \preceq H_i \preceq M_H I$ for all $i \in B$, then*

$$\lambda_{\max}(F_{\mathrm{red}}) \le \lambda_{\max}(\Lambda_\theta) + \frac{M_H}{|B|} \sum_{i\in B} \sum_v S_{v,i}^2.$$

***(ii) Active coverage.*** *Let $\mathcal{U}_{\mathrm{act}}$ be a declared active subspace. Suppose that*

$$\frac{1}{|B|} \sum_{i\in B} \nu^\top (J_i^{\mathrm{red}})^\top H_i J_i^{\mathrm{red}} \nu \ge \mu_{\mathrm{act}} \qquad \textit{for every unit } \nu \in \mathcal{U}_{\mathrm{act}}.$$

*Then*

$$\lambda_{\min}(F_{\mathrm{red}}|_{\mathcal{U}_{\mathrm{act}}}) \ge \lambda_{\min}(\Lambda_\theta|_{\mathcal{U}_{\mathrm{act}}}) + \mu_{\mathrm{act}}.$$

*The same statements hold in optimizer-whitened coordinates with $\bar{F} = P^{1/2} F_{\mathrm{red}} P^{1/2}$, $\bar{\Lambda}_\theta = P^{1/2}\Lambda_\theta P^{1/2}$, and budgets constructed for the block columns of $J_i^{\mathrm{red}} P^{1/2}$.*

**Proof sketch.** The proof has three steps. First, the constructed state-source elimination, its root response, the triangle inequality, and submultiplicativity give $\|J^{\mathrm{red}}_{i,v}\|_{\mathrm{op}} \le S_{v,i}$. Second, the reduced curvature is the positive-semidefinite prior term plus the averaged output-loss GGN terms, so these Jacobian bounds control its largest eigenvalue. Third, the active lower statement is a Rayleigh-quotient argument on the declared active subspace.

*Proof.* Fix a sample $i$, parameter block $v$, and unit perturbation $\delta\theta_v$. The state-source recursion proved above gives the exact reduced-Jacobian identity, so

$$\begin{aligned}\|J^{\mathrm{red}}_{i,v}\delta\theta_v\| &\le \|\mathcal{R}_i\|_{\mathrm{op}} \sum_{C\in\mathcal{C}(v)} \left(\prod_{e\in \mathrm{path}(C\rightsquigarrow C_{\mathrm{out}})} \|T_e\|_{\mathrm{op}}\right) \|B^{(C)}_{v,i}\|_{\mathrm{op}} \\ &= S_{v,i}.\end{aligned}$$

Taking the supremum over unit $\delta\theta_v$ proves $\|J^{\mathrm{red}}_{i,v}\|_{\mathrm{op}} \le S_{v,i}$. The following curvature argument uses only this inequality and therefore also applies to independently verified forward-Jacobian budgets.

We next bound the top curvature. Write $J^{\mathrm{red}}_i$ in block columns as $J^{\mathrm{red}}_i = [J^{\mathrm{red}}_{i,v}]_v$. For any parameter perturbation $\delta\theta = (\delta\theta_v)_v$ with $\|\delta\theta\| = 1$,

$$\|J^{\mathrm{red}}_i\delta\theta\| = \left\|\sum_v J^{\mathrm{red}}_{i,v}\delta\theta_v\right\| \le \sum_v S_{v,i}\|\delta\theta_v\| \le \left(\sum_v S^2_{v,i}\right)^{1/2}.$$

Thus $\|J^{\mathrm{red}}_i\|^2_{\mathrm{op}} \le \sum_v S^2_{v,i}$. Since $0 \preceq H_i \preceq M_H I$,

$$\lambda_{\max}\big((J^{\mathrm{red}}_i)^\top H_i J^{\mathrm{red}}_i\big) \le M_H\|J^{\mathrm{red}}_i\|^2_{\mathrm{op}} \le M_H \sum_v S^2_{v,i}.$$

Using subadditivity of $\lambda_{\max}$ for symmetric matrices,

$$\lambda_{\max}(F_{\mathrm{red}}) \le \lambda_{\max}(\Lambda_\theta) + \frac{M_H}{|B|}\sum_{i\in B}\sum_v S^2_{v,i}.$$

This proves the upper-transfer part.

For the active lower part, let $\nu \in \mathcal{U}_{\mathrm{act}}$ be unit norm. By the reduced curvature identity,

$$\nu^\top F_{\mathrm{red}}\nu = \nu^\top\Lambda_\theta\nu + \frac{1}{|B|}\sum_{i\in B}\nu^\top(J^{\mathrm{red}}_i)^\top H_i J^{\mathrm{red}}_i\nu.$$

The active-coverage assumption gives

$$\nu^\top F_{\mathrm{red}}\nu \ge \nu^\top\Lambda_\theta\nu + \mu_{\mathrm{act}} \ge \lambda_{\min}(\Lambda_\theta|_{\mathcal{U}_{\mathrm{act}}}) + \mu_{\mathrm{act}}.$$

Taking the infimum over unit $\nu \in \mathcal{U}_{\mathrm{act}}$ proves

$$\lambda_{\min}(F_{\mathrm{red}}|_{\mathcal{U}_{\mathrm{act}}}) \ge \lambda_{\min}(\Lambda_\theta|_{\mathcal{U}_{\mathrm{act}}}) + \mu_{\mathrm{act}}.$$

For optimizer-whitened coordinates, set $\delta\theta = P^{1/2}\widetilde{\theta}$ and replace the module parameter Jacobian by $R_\theta P^{1/2}$. The state precision $K$ is unchanged, while each local source uses the corresponding block columns of $-J^\top_{f,u}W_f J_{f,\theta}P^{1/2}$. A dense $P^{1/2}$ may create injections at multiple cliques, all of which are included in $\mathcal{C}(v)$. The output Jacobian is $J^{\mathrm{red}}_i P^{1/2}$ and the prior precision is $\bar{\Lambda}_\theta = P^{1/2}\Lambda_\theta P^{1/2}$. The same norm and Rayleigh-quotient arguments then apply. □

The next theorem uses the certificates just proved. It does not depend on the details of the separator construction; its inputs are an upper curvature certificate $S^2_+$, a lower active certificate $S^2_-$, an optimizer metric $P$, and an active subspace $\mathcal{U}_{\mathrm{act}}$. The architecture corollaries below follow this interface: define the carrier and transfer gains, obtain an upper curvature budget, verify or measure active coverage, and then apply the two-sided criticality theorem.

**Theorem 3 (Two-sided architecture-local criticality, restated).**

*Let $\bar{F} = P^{1/2} F_{\mathrm{red}} P^{1/2}$ be the optimizer-whitened reduced curvature at a checkpoint. Suppose architecture-local certificates give constants $S_+^2$ and $S_-^2 > 0$ such that*

$$\lambda_{\max}(\bar{F}) \le S_+^2, \qquad \lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S_-^2.$$

*Then every scalar step $0 < \eta < 2/S_+^2$ is locally stable for the linearized update in optimizer coordinates. Moreover, for every unit $u \in \mathcal{U}_{\mathrm{act}}$,*

$$\|\eta \bar{F} u\|_2 \ge \eta S_-^2.$$

*Thus the linearized update is nonexpansive on the full parameter space in optimizer coordinates and nontrivial on the active subspace. For a positive scalar coordinate $\eta_{\mathrm{eff}}$, the corresponding theoretical active-plasticity coordinate is $\rho_{\mathrm{act}} = \eta_{\mathrm{eff}} \lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge \eta_{\mathrm{eff}} S_-^2$.*

**Proof sketch.** The proof is the stable-plastic wrapper around the certificates. In optimizer coordinates, the linearized update is multiplication by $I - \eta\bar{F}$. The upper certificate bounds all eigenvalues of $\bar{F}$, so $0 < \eta < 2/S_+^2$ makes the update nonexpansive on the positive-curvature subspace. The lower certificate says that every unit vector in the active subspace has Rayleigh curvature at least $S_-^2$. Since $\|\bar{F}u\| \ge u^\top \bar{F} u$ for unit $u$, the update has nontrivial movement on the active subspace.

*Proof.* The optimizer-whitened linearized error update is

$$\widetilde{e}^{+} = (I - \eta\bar{F})\widetilde{e}.$$

Since $\bar{F}$ is symmetric positive semidefinite and $\lambda_{\max}(\bar{F}) \le S_+^2$, every eigenvalue $\lambda$ of $\bar{F}$ satisfies $0 \le \lambda \le S_+^2$. If $0 < \eta < 2/S_+^2$, then $|1 - \eta\lambda| \le 1$ for all $0 \le \lambda \le S_+^2$. Thus $I - \eta\bar{F}$ is nonexpansive in optimizer-whitened coordinates. It is strictly contractive on eigenvectors with positive curvature and neutral on zero-curvature directions. This proves local stability for the linearized update.

Now let $u \in \mathcal{U}_{\mathrm{act}}$ be unit norm. The lower certificate means $u^\top \bar{F} u \ge S_-^2$. Since $\bar{F}$ is symmetric positive semidefinite, $u^\top \bar{F} u \le \|u\| \, \|\bar{F}u\| = \|\bar{F}u\|$. Therefore

$$\|\eta \bar{F} u\| \ge \eta u^\top \bar{F} u \ge \eta S_-^2.$$

This proves the active movement statement.

Equation (12) defines the corresponding theoretical coordinate using the active curvature; for $\eta_{\mathrm{eff}} > 0$, it is bounded below by $\eta_{\mathrm{eff}} S_-^2$. The experimental diagnostic replaces that curvature by the projected probe estimate described in Appendix B.1. It is a plasticity coordinate, not a second learning-rate boundary. □

**Discussions.** The separator-transfer theorem converts verified reduced-Jacobian budgets into an upper curvature bound. The lower bound additionally requires the stated active-coverage condition. The two-sided criticality theorem converts these bounds into a stable-step condition and an active-plasticity lower bound for the fixed checkpoint-linearized model.

The corollaries below use direct forward-Jacobian carrier bounds, not an identification of those carriers with Gaussian source-message maps. Each budget includes its final readout and any conversion from the declared hidden-state norms to the Euclidean output norm. For the displayed carrier formulas, initial inputs/states and residual scales are held fixed. Any trainable initial state, scale, or readout parameter must be included through its additional parameter-injection term. Each architecture example specifies the carrier, such as a residual stream, concatenative stream, attention route, scan state, or another module interface; bounds the transfer products that give $S_+^2$; identifies an active subspace $\mathcal{U}_{\mathrm{act}}$ and lower certificate $S_-^2$, either structurally or through a recent/probe-gradient span; and evaluates the theorem with the checkpoint optimizer coordinate $\eta_{\mathrm{eff}}$. For compactness, the path budgets $S_{v,i}$ below are written in the coordinate system being certified: for Euclidean gradient descent this is the original parameter coordinate, while for optimizer-whitened certificates each parameter-injection map is composed with the appropriate $P^{1/2}$ factor and $\bar{\Lambda}_\theta = P^{1/2} \Lambda_\theta P^{1/2}$. The output is the sufficient stable-step condition $0 < \eta < 2/S_+^2$ for the fixed linearized model and, for $\eta_{\mathrm{eff}} > 0$, the lower bound $\rho_{\mathrm{act}} \ge \eta_{\mathrm{eff}} S_-^2$.

This is the distinction measured in Section 5.1. The upper-band plots compare the actual optimizer-state scalar step with the measured local edge. The active-plasticity plots track the lower-side signal on a gradient-derived active subspace. The backward-gain plots in Appendix B.1 are directional VJP measurements across architecture boundaries; they are scalable proxies for where separator transfers magnify, damp, or preserve signal.

**Block Gershgorin bridge.** The separator-transfer bound also recovers a familiar block-row control of the top curvature. This connects the path-budget view to classical diagonal-dominance reasoning.

**Proposition 5** (Block Gershgorin control from separator budgets)**.** *Partition the parameter vector into blocks $v$. Let $F_{vw}$ denote the $(v,w)$-block of $F_{\mathrm{red}}$, and suppose $0 \preceq H_i \preceq M_H I$. If $\|J_{i,v}^{\mathrm{red}}\|_{\mathrm{op}} \le S_{v,i}$, then*

$$\|F_{vw}\|_{\mathrm{op}} \le \|\Lambda_{\theta,vw}\|_{\mathrm{op}} + \frac{M_H}{|B|}\sum_{i\in B} S_{v,i}S_{w,i}.$$

*Consequently,*

$$\lambda_{\max}(F_{\mathrm{red}}) \le \max_v \left\{ \|F_{vv}\|_{\mathrm{op}} + \sum_{w\neq v} \|F_{vw}\|_{\mathrm{op}} \right\}.$$

*Proof.* The block formula for $F_{\mathrm{red}}$ gives

$$F_{vw} = \Lambda_{\theta,vw} + \frac{1}{|B|}\sum_{i\in B} (J_{i,v}^{\mathrm{red}})^\top H_i J_{i,w}^{\mathrm{red}}.$$

Since $0 \preceq H_i \preceq M_H I$,

$$\|(J_{i,v}^{\mathrm{red}})^\top H_i J_{i,w}^{\mathrm{red}}\|_{\mathrm{op}} \le M_H \|J_{i,v}^{\mathrm{red}}\|_{\mathrm{op}} \|J_{i,w}^{\mathrm{red}}\|_{\mathrm{op}} \le M_H S_{v,i}S_{w,i}.$$

This proves the block norm bound. For the block Gershgorin inequality, let $x$ be a unit eigenvector of $F_{\mathrm{red}}$ with eigenvalue $\lambda$, and choose a block $k$ such that $\|x_k\| \ge \|x_v\|$ for all $v$. The $k$-th block equation is $\lambda x_k = F_{kk}x_k + \sum_{w\neq k} F_{kw}x_w$. Taking norms and using $\|x_w\| \le \|x_k\|$ gives

$$|\lambda|\|x_k\| \le \left( \|F_{kk}\|_{\mathrm{op}} + \sum_{w\neq k} \|F_{kw}\|_{\mathrm{op}} \right) \|x_k\|.$$

Since $F_{\mathrm{red}} \succeq 0$, $\lambda_{\max}$ is bounded by the displayed maximum over block rows. □

**DenseNet and concatenative carriers.** A DenseNet block carries old features forward by concatenation and appends a new innovation. Write the local linearized layer as

$$\delta h_{\ell+1,i} = \begin{bmatrix} \delta h_{\ell,i} \\ C_{\ell,i}\delta h_{\ell,i} + B_{\ell,i}\delta\theta_\ell \end{bmatrix},$$

where the innovation has dimension $k_\ell$ and the output carrier has dimension $d_{\ell+1} = d_\ell + k_\ell$. Use the RMS norm $\|x\|_{\mathrm{rms},d} = d^{-1/2}\|x\|_2$. Suppose $\|C_{\ell,i}\|_{\mathrm{rms}\leftarrow\mathrm{rms}} \le \gamma_{\ell,i}$ and $\|B_{\ell,i}\|_{\mathrm{rms}\leftarrow 2} \le \rho_{\ell,i}$. Then the carrier gain satisfies

$$a_{\ell,i}^2 \le 1 + \frac{k_\ell}{d_{\ell+1}}(\gamma_{\ell,i}^2 - 1)_+.$$

Compression or transition layers are included by multiplying the downstream budget by their operator-norm gains.

**Corollary 6** (DenseNet concatenative criticality)**.** *Consider a DenseNet-style carrier satisfying the bounds above, with transition gains $\kappa_{b,i}$. Let $S_{v,i}$ be the resulting path budget, using the concatenative gains $a_{\ell,i}$, injection gains $\rho_{\ell,i}$, and transition gains $\kappa_{b,i}$. Suppose*

$$\lambda_{\max}(\bar{\Lambda}_\theta) + \frac{M_H}{|B|}\sum_{i\in B}\sum_v S_{v,i}^2 \le S_+^2,$$

*and suppose active coverage gives $\lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S_-^2 > 0$ in optimizer coordinates. Then Theorem 3 gives the sufficient stable-step condition $0 < \eta < 2/S_+^2$ for the fixed linearized model and, for $\eta_{\mathrm{eff}} > 0$, the bound $\rho_{\mathrm{act}} \ge \eta_{\mathrm{eff}} S_-^2$.*

*Proof.* For the carrier response, set $\delta\theta_\ell = 0$. Then

$$\|\delta h_{\ell+1,i}\|^2_{\mathrm{rms},d_{\ell+1}} = \frac{d_\ell}{d_{\ell+1}}\|\delta h_{\ell,i}\|^2_{\mathrm{rms},d_\ell} + \frac{k_\ell}{d_{\ell+1}}\|C_{\ell,i}\delta h_{\ell,i}\|^2_{\mathrm{rms},k_\ell}.$$

Using $\|C_{\ell,i}\| \le \gamma_{\ell,i}$ gives

$$\|\delta h_{\ell+1,i}\|^2_{\mathrm{rms},d_{\ell+1}} \le \left(1 + \frac{k_\ell}{d_{\ell+1}}(\gamma^2_{\ell,i} - 1)_+\right)\|\delta h_{\ell,i}\|^2_{\mathrm{rms},d_\ell}.$$

Thus the layer has carrier gain $a_{\ell,i}$, while parameter injections are bounded by $\rho_{\ell,i}$. Transition layers multiply downstream norms by $\kappa_{b,i}$. Therefore the resulting $S_{v,i}$ satisfies $\|J^{\mathrm{red}}_{i,v}\|_{\mathrm{op}} \le S_{v,i}$, and Theorem 2 gives the displayed upper certificate. The assumed active coverage gives the lower certificate. Applying Theorem 3 gives the stated sufficient stable-step condition and active-plasticity lower bound. □

The DenseNet insight is that concatenation preserves old coordinates until a transition or compression layer acts. Depth amplification is charged through width-weighted innovation energy and transition gains, rather than through a single fixed-width residual product.

**Residual networks and WideResNets.** A residual block has local form

$$h_{\ell+1,i} = C_{\ell,i}h_{\ell,i} + \alpha_\ell\phi_{\ell,i}(h_{\ell,i}, \theta_\ell),$$

where $C_{\ell,i}$ is the skip or projection path and $\alpha_\ell$ is a fixed residual scale. Suppose $\|C_{\ell,i}\|_{\mathrm{op}} \le c_{\ell,i}$, $\|D_h\phi_{\ell,i}\|_{\mathrm{op}} \le \gamma_{\ell,i}$, and $\|D_{\theta_v}\phi_{\ell,i}\|_{\mathrm{op}} \le \rho_{\ell,v,i}$. Let $\rho^{\mathrm{skip}}_{\ell,v,i}$ bound $\|D_{\theta_v}(C_{\ell,i}h_{\ell,i})\|_{\mathrm{op}}$ with $h_{\ell,i}$ held fixed; it is zero when the skip is parameter-independent. The carrier and injection gains may then be taken as $a_{\ell,i} = c_{\ell,i} + |\alpha_\ell|\gamma_{\ell,i}$ and $\rho^{\mathrm{res}}_{\ell,v,i} = \rho^{\mathrm{skip}}_{\ell,v,i} + |\alpha_\ell|\rho_{\ell,v,i}$.

**Corollary 7** (Residual-stream criticality)**.** *Consider a residual or WideResNet carrier satisfying the bounds above. Let $S_{v,i}$ be the resulting path budget. Suppose*

$$\lambda_{\max}(\bar{\Lambda}_\theta) + \frac{M_H}{|B|}\sum_{i\in B}\sum_v S^2_{v,i} \le S^2_+,$$

*and suppose active coverage gives $\lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S^2_- > 0$ in optimizer coordinates. Then Theorem 3 gives the sufficient stable-step condition $0 < \eta < 2/S^2_+$ for the fixed linearized model and, for $\eta_{\mathrm{eff}} > 0$, the bound $\rho_{\mathrm{act}} \ge \eta_{\mathrm{eff}}S^2_-$.*

*Proof.* With each parameter derivative taken at fixed $h_{\ell,i}$, linearization gives

$$\begin{aligned}\delta h_{\ell+1,i} = {} & (C_{\ell,i} + \alpha_\ell D_h\phi_{\ell,i})\delta h_{\ell,i} \\ & + \sum_v \left[D_{\theta_v}(C_{\ell,i}h_{\ell,i}) + \alpha_\ell D_{\theta_v}\phi_{\ell,i}\right]\delta\theta_v.\end{aligned}$$

The state-transport norm is at most $c_{\ell,i} + |\alpha_\ell|\gamma_{\ell,i}$, and the block-$v$ injection norm is at most $\rho^{\mathrm{skip}}_{\ell,v,i} + |\alpha_\ell|\rho_{\ell,v,i}$. These bounds give valid forward path budgets $S_{v,i}$. The upper certificate follows from Theorem 2; the lower certificate is the assumed active coverage. The conclusion follows from Theorem 3. □

The residual insight is that the skip path is both an upper-side and lower-side carrier. Residual scale changes the upper budget through $c_{\ell,i} + |\alpha_\ell|\gamma_{\ell,i}$, while the skip path can preserve active directions that would otherwise be damped by the nonlinear branch. Width changes in WideResNet alter the same quantities through branch gains, injection norms, and active coverage.

**Transformer and GPT residual streams.** A pre-norm transformer block has two residual sublayers. In local form,

$$h_{\ell+1/2,i} = h_{\ell,i} + \alpha_\ell^{\mathrm{attn}} A_{\ell,i}(\mathrm{LN}(h_{\ell,i})), \qquad h_{\ell+1,i} = h_{\ell+1/2,i} + \alpha_\ell^{\mathrm{mlp}} M_{\ell,i}(\mathrm{LN}(h_{\ell+1/2,i})).$$

Let $\gamma_{\ell,i}^{\mathrm{attn}}$ and $\gamma_{\ell,i}^{\mathrm{mlp}}$ bound the state derivatives of the attention and MLP branches, including layer normalization on the measured local region. Let the corresponding parameter injection bounds be $\rho_{\ell,v,i}^{\mathrm{attn}}$ and $\rho_{\ell,v,i}^{\mathrm{mlp}}$. Then the block carrier gain may be taken as

$$a_{\ell,i} = (1 + |\alpha_\ell^{\mathrm{attn}}|\gamma_{\ell,i}^{\mathrm{attn}})(1 + |\alpha_\ell^{\mathrm{mlp}}|\gamma_{\ell,i}^{\mathrm{mlp}}),$$

with parameter-injection norm bounds scaled by the absolute values of the corresponding fixed residual scales.

**Corollary 8** (Transformer residual-stream criticality)**.** *Consider a transformer or GPT residual stream satisfying the local bounds above. Let $S_{v,i}$ be the resulting attention/MLP path budget. Suppose*

$$\lambda_{\max}(\bar{\Lambda}_\theta) + \frac{M_H}{|B|} \sum_{i \in B} \sum_v S_{v,i}^2 \le S_+^2,$$

*and suppose active coverage gives $\lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S_-^2 > 0$ in optimizer coordinates. Then Theorem 3 gives the sufficient stable-step condition $0 < \eta < 2/S_+^2$ for the fixed linearized model and, for $\eta_{\mathrm{eff}} > 0$, the bound $\rho_{\mathrm{act}} \ge \eta_{\mathrm{eff}} S_-^2$.*

*Proof.* Each transformer sublayer is a residual carrier. The attention sublayer has gain at most $1 + |\alpha_\ell^{\mathrm{attn}}|\gamma_{\ell,i}^{\mathrm{attn}}$, and the MLP sublayer has gain at most $1 + |\alpha_\ell^{\mathrm{mlp}}|\gamma_{\ell,i}^{\mathrm{mlp}}$. Sequential composition gives the product gain $a_{\ell,i}$. Parameter injections enter through the attention and MLP branches and are multiplied by the appropriate downstream products. Hence the transformer block gives valid path budgets $S_{v,i}$. The upper certificate follows from Theorem 2; the lower certificate is the assumed active coverage. The conclusion follows from Theorem 3. □

The transformer insight is that the residual stream is the primary carrier, while attention and MLP branches are routed innovations. Removing the residual carrier can leave the upper edge finite while destroying active coverage, which is the stable-but-inactive failure mode measured in the GPT-20M ablations.

**State-space and recurrent carriers.** A recurrent or state-space carrier propagates signal over a sequence as well as through depth. A local linearized recurrence has the form

$$\delta s_{t+1,i} = A_{t,i}\delta s_{t,i} + \sum_{v \in \mathcal{V}_t} B_{t,v,i}\delta\theta_v, \qquad \delta o_{t,i} = R_{t,i}\delta s_{t,i}.$$

Suppose $\|A_{t,i}\|_{\mathrm{op}} \le a_{t,i}$, $\|B_{t,v,i}\|_{\mathrm{op}} \le \rho_{t,v,i}$, and $\|R_{t,i}\|_{\mathrm{op}} \le r_{t,i}$. A parameter injection at time $\tau$ enters state $\tau + 1$ and has temporal downstream budget

$$S_{\tau,v,i} = \sum_{t \ge \tau+1} r_{t,i} \left( \prod_{m=\tau+1}^{t-1} a_{m,i} \right) \rho_{\tau,v,i},$$

where an empty product is one. If $a_{t,i} \le \bar{a}_i < 1$ and $r_{t,i} \le \bar{r}_i$, this is bounded by $\bar{r}_i \rho_{\tau,v,i}/(1 - \bar{a}_i)$.

**Corollary 9** (State-space/recurrent criticality)**.** *Consider a recurrent, gated, or state-space carrier satisfying the bounds above, possibly combined with a residual depth stack. Form $S_{v,i}$ by summing over all parameter injections and complete output paths, multiplying temporal and any intervening residual-depth gains along each path, including the final readout. Suppose*

$$\lambda_{\max}(\bar{\Lambda}_\theta) + \frac{M_H}{|B|} \sum_{i \in B} \sum_v S_{v,i}^2 \le S_+^2,$$

*and suppose active coverage gives $\lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}}) \ge S_-^2 > 0$ in optimizer coordinates. Then Theorem 3 gives the sufficient stable-step condition $0 < \eta < 2/S_+^2$ for the fixed linearized model and, for $\eta_{\mathrm{eff}} > 0$, the bound $\rho_{\mathrm{act}} \ge \eta_{\mathrm{eff}} S_-^2$.*

*Proof.* Unrolling the recurrence, the contribution of an injection $B_{\tau,v,i}\delta\theta_v$ to output time $t \geq \tau+1$ is $R_{t,i}A_{t-1,i}\cdots A_{\tau+1,i}B_{\tau,v,i}\delta\theta_v$, with an empty transition product equal to the identity. Its norm is bounded by

$$r_{t,i}\left(\prod_{m=\tau+1}^{t-1} a_{m,i}\right)\rho_{\tau,v,i}\|\delta\theta_v\|.$$

Summing over output times gives the displayed temporal budget. If $a_{t,i} \leq \bar{a}_i < 1$ and $r_{t,i} \leq \bar{r}_i$, the geometric bound follows from $\sum_{k\geq 0}\bar{a}_i^k = (1-\bar{a}_i)^{-1}$. Multiplying gains along each complete temporal/depth path and summing over injections and output paths gives a valid $S_{v,i}$. The upper certificate follows from Theorem 2; the lower certificate is the assumed active coverage. The conclusion follows from Theorem 3. □

The recurrent insight is that temporal memory and residual depth contribute products of forward state-Jacobian gains. These give reduced-Jacobian budgets without identifying the carrier maps with Gaussian information-source maps. Stable memory kernels control the upper budget; the lower bound still requires active coverage.

**Further discussions.** The same interface can be used for other training systems when the relevant carrier, transfer gains, active subspace, and optimizer coordinate can be specified. For example, preference-optimization pipelines or deterministic Bayesian-message architectures can be described by additional factors and interfaces in the lifted graph. We do not state separate formal corollaries for these cases here because the main empirical evidence in this paper is on vision architectures, GPT-style residual streams, and state-space/recurrent models. The theorem interface above is the reusable part: prove an upper budget, identify active coverage, and then apply Theorem 3.

The architecture examples should be read together with the diagnostic appendix. The unified upper-band, active-plasticity, and backward-gain plots show how residual scale, normalization, width, compression, concatenation, transformer residual paths, and state-space memory move the same two theorem quantities. The upper edge tests stability, and the active coordinate tests whether the stable update still moves through loss-changing directions.

# E Resolved KL-to-Gibbs proof and PAC-Bayes computation

This appendix proves Theorem 4 and the localized PAC-Bayes consequence, and describes the retained-spectrum defect calculations. We distinguish the formal entropy/LSI argument from checkpoint moment estimates and from direct prior-KL certification. In particular, the original numerical scores did not evaluate the finite-time dynamical bound.

**Setup.** Fix a checkpoint, a finite-dimensional Euclidean comparison space $E$, and a Gibbs reference, and hold this space and reference fixed over the interval considered below. The local coordinate describes displacements from the checkpoint. For an ambient comparison, embed both laws using the same fixed complement; more generally, they may share the same conditional law on $E^\perp$ given the $E$-coordinate. The KL chain rule then reduces the ambient KL to the KL on $E$. Merely projecting a training trajectory onto $E$, or matching the marginal law on $E^\perp$, does not establish this reduction. The experimental retained covariance spaces are checkpoint-local constructions.

We write $w \in E$ for the local coordinate. Let $\pi_0^E$ be a reference law on $E$, usually a Gaussian reference whose precision is inherited from the lifted or reduced checkpoint geometry. Let $\widehat{L}_E(w)$ be a bounded empirical loss evaluated under displacement $w$. For inverse temperature $\beta > 0$, define

$$\pi_\beta^E(dw) = Z_{\beta,E}^{-1}e^{-\beta\widehat{L}_E(w)}\,\pi_0^E(dw), \qquad \mathcal{F}_\beta^E = -\beta^{-1}\log\int_E e^{-\beta\widehat{L}_E(w)}\,\pi_0^E(dw).$$

The Gibbs identity is

$$\widehat{L}_E(\rho) + \beta^{-1}\mathrm{KL}(\rho\|\pi_0^E) - \mathcal{F}_\beta^E = \beta^{-1}\mathrm{KL}(\rho\|\pi_\beta^E).$$

Thus KL to $\pi_\beta^E$ is excess empirical free energy on the comparison space.

Let $\rho_t^E$ be the law of an Itô diffusion on $E$, $dW_t = b_E(t,W_t)\,dt + \sigma_E(t,W_t)\,dB_t$, with covariance $a_E(t,w) = \sigma_E(t,w)\sigma_E(t,w)^\top$. We assume $a_E(t,w)$ is symmetric positive definite on $E$, smooth

enough for the Fokker–Planck equation, and that either the domain is compact with no-flux boundary conditions or the densities decay fast enough for the integrations by parts below. A finite numerical bound also requires finite initial $\mathrm{KL}(\rho_0^E \| \pi_\beta^E)$. These are assumptions on the local stochastic model; finite checkpoint moment measurements do not verify the diffusion or log-Sobolev hypotheses. The Fokker–Planck equation is

$$\partial_t \rho_t = -\nabla \cdot (b_E \rho_t) + \frac{1}{2} \nabla \cdot \nabla \cdot (a_E \rho_t).$$

For the covariance field $a_E$ and reference $\pi_\beta^E$, define the reversible Gibbs drift

$$b_{\mathrm{rev},E} = \frac{1}{2} a_E \nabla \log \pi_\beta^E + \frac{1}{2} \nabla \cdot a_E,$$

where $(\nabla \cdot a_E)_i = \sum_j \partial_j a_{E,ij}$. If $a_E$ is constant, this reduces to $b_{\mathrm{rev},E} = \frac{1}{2} a_E \nabla \log \pi_\beta^E$. The defect is

$$\mathfrak{D}_E(t) = \mathbb{E}_{\rho_t^E} \left[ \| b_E - b_{\mathrm{rev},E} \|_{a_E^{-1}}^2 \right].$$

It is the squared mismatch between the local law created by training and the reversible Gibbs dynamics on $E$, measured in the inverse covariance metric.

We use the following log-Sobolev inequality. The reference $\pi_\beta^E$ satisfies a metric log-Sobolev inequality with rate $\lambda_E > 0$ if every smooth probability density $\rho$ on $E$ satisfies

$$\mathrm{KL}(\rho \| \pi_\beta^E) \le \frac{1}{2\lambda_E} \mathcal{I}_E(\rho \| \pi_\beta^E), \qquad \mathcal{I}_E(\rho \| \pi_\beta^E) = \mathbb{E}_\rho \left[ \left\| \nabla \log \frac{\rho}{\pi_\beta^E} \right\|_{a_E}^2 \right].$$

Equivalently, $\mathcal{I}_E(\rho \| \pi_\beta^E) \ge 2\lambda_E \mathrm{KL}(\rho \| \pi_\beta^E)$. For a local quadratic reference with lower curvature $\mu_E$ and covariance $a_E \succeq a_{\min} I$, the standard strongly log-concave estimate gives $\lambda_E \ge \beta \mu_E a_{\min}$.

The block decomposition used later is declared at the level of local coordinates. Let $\mathcal{B}$ be a finite collection of architecture-native blocks, such as convolution kernels, normalization affine parameters, attention projections, residual-stream modules, or other layer/block groups. A blockwise comparison space has the orthogonal decomposition $E = \oplus_{B \in \mathcal{B}} E_B$. If $a_E = \oplus_{B \in \mathcal{B}} a_{E_B}$, the inverse-metric quadratic form splits into block contributions. Each block contribution is the expectation under $\rho_t^E$ of the squared inverse-metric norm of $(b_E - b_{\mathrm{rev},E})_B$. This is the proof-level statement behind the blockwise Gram computations used in Section 5.2.

The comparison space $E$ is chosen to avoid charging KL on directions that are poorly matched to the local probability law created by training. A *symmetry direction* is a parameter displacement that leaves the represented function unchanged on the measured data. If $f_\theta(X)$ denotes model outputs on data $X$, a displacement $v$ is an infinitesimal symmetry when $J_f(\theta; X)v = 0$, or approximately so. ReLU scale symmetry is a canonical example: scaling incoming weights of a hidden unit and inversely scaling outgoing weights can leave the represented function unchanged.

A *gauge direction* is a coordinate or output-space redundancy that is invisible, or nearly invisible, to the loss. At the logit level, the softmax cross-entropy Hessian satisfies $H_{\mathrm{CE}}(p)\mathbf{1} = 0$, so a common shift of all logits changes coordinates but not probabilities. BatchNorm/ReLU scale candidates are structural gauges in parameter space; common-logit shifts are loss-space gauges in output space.

A *loss tangent* is a displacement that changes logits while remaining low-curvature for the loss. If $q_v = J_f v$ is the output tangent and $H_\ell$ is the output-space Hessian of the bounded loss, then $v$ is a low-loss tangent when $\| H_\ell^{1/2} q_v \| \ll \| q_v \|$. These directions may move raw logits but have little effect on probabilities or clipped loss.

An *augmentation tangent* compares parameter movement with data-space movement. Let $T$ be an augmentation such as crop, color, MixUp, or CutMix, and define the augmentation response $q_T(x) = f_\theta(Tx) - f_\theta(x)$, or the corresponding centered logit response. A parameter direction $v$ is augmentation-aligned when $J_f v$ has high cosine similarity with $q_T$. Such a direction is not a parameter-space gauge; it indicates that a local parameter displacement resembles a data-transformation response.

These notions help interpret the inverse-metric defect. A direction with a tiny minibatch-gradient covariance eigenvalue receives a large inverse-metric weight; if it is also a symmetry, gauge, or low-loss tangent, finite movement along it may barely affect held-out predictions. A retained comparison uses a declared common complement and keeps directions with larger stochastic variation. The ResNet/CIFAR-10 diagnostics in Section 5.2 illustrate the defect geometry: the low-covariance tail concentrates the inverse-metric defect, finite prediction movement remains small, and retained directions reveal train/validation alignment under augmentation. These measurements do not evaluate an ambient posterior KL.

**Theorem 4 (Resolved KL-to-Gibbs comparison, restated).**

*Fix a checkpoint and a declared finite-dimensional comparison space $E$. Let $\rho_t^E$ be an Itô law on $E$ with drift $b_E$ and covariance $a_E$, and let $\pi_\beta^E$ be the localized Gibbs reference. Suppose that $\pi_\beta^E$ satisfies the metric log-Sobolev inequality*

$$\mathcal{I}_E(\rho\|\pi_\beta^E) \geq 2\lambda_E \,\mathrm{KL}(\rho\|\pi_\beta^E)$$

*for all smooth $\rho$. Then, for every $\varepsilon \in (0,1)$,*

$$\mathrm{KL}(\rho_t^E\|\pi_\beta^E) \leq e^{-(1-\varepsilon)\lambda_E t}\mathrm{KL}(\rho_0^E\|\pi_\beta^E) + \frac{1}{2\varepsilon}\int_0^t e^{-(1-\varepsilon)\lambda_E(t-s)}\mathfrak{D}_E(s)\,ds.$$

*In a checkpoint-local quadratic model with lower curvature $\mu_E$ on $E$ and covariance scale $a_E \succeq a_{\min} I$, one may take $\lambda_E \geq \beta\mu_E a_{\min}$.*

*If $E = \oplus_{B\in\mathcal{B}} E_B$ is an orthogonal block decomposition and $a_E = \oplus_{B\in\mathcal{B}} a_{E_B}$, then*

$$\mathfrak{D}_E(t) = \sum_{B\in\mathcal{B}} \mathfrak{D}_{E_B}(t), \qquad \mathfrak{D}_{E_B}(t) := \mathbb{E}_{\rho_t^E}\left[\|(b_E - b_{\mathrm{rev},E})_B\|^2_{a_{E_B}^{-1}}\right].$$

**Proof sketch.** The proof has four steps. First, write the Fokker–Planck current as the sum of a reversible Gibbs current and a drift mismatch. Differentiating $\mathrm{KL}(\rho_t^E\|\pi_\beta^E)$ gives an exact balance: a negative metric Fisher-information term plus a cross term involving $b_E - b_{\mathrm{rev},E}$. Second, bound the cross term by Young's inequality in the covariance metric. This trades part of the Fisher information for the defect $\mathfrak{D}_E(t)$. Third, use the metric log-Sobolev inequality to lower-bound Fisher information by KL. Fourth, solve the resulting scalar differential inequality by Grönwall. If the comparison space and covariance decompose over blocks, the inverse-metric defect is the sum of the blockwise inverse-metric defects.

**Complete proof of Theorem 4.**

*Proof.* We suppress the superscript $E$ during the proof and write $\rho_t, \pi_\beta, b, a, b_{\mathrm{rev}}$. Let $h_t = \rho_t/\pi_\beta$ and $s_t = \nabla \log h_t$.

*Step 1: entropy balance.* The Fokker–Planck equation can be written as $\partial_t\rho_t = -\nabla\cdot J_t$, where $J_t = b\rho_t - \frac{1}{2}\nabla\cdot(a\rho_t)$. The reversible drift is $b_{\mathrm{rev}} = \frac{1}{2}a\nabla\log\pi_\beta + \frac{1}{2}\nabla\cdot a$, so the reversible current associated with the same covariance field and density $\rho_t$ is $J_{\mathrm{rev},t} = b_{\mathrm{rev}}\rho_t - \frac{1}{2}\nabla\cdot(a\rho_t)$. Since $\nabla\cdot(a\rho_t) = (\nabla\cdot a)\rho_t + a\nabla\rho_t$, we have

$$J_{\mathrm{rev},t} = \frac{1}{2}a\rho_t\nabla\log\pi_\beta - \frac{1}{2}a\nabla\rho_t = -\frac{1}{2}a\rho_t\nabla\log\frac{\rho_t}{\pi_\beta} = -\frac{1}{2}a\rho_t s_t.$$

Therefore $J_t = (b - b_{\mathrm{rev}})\rho_t - \frac{1}{2}a\rho_t s_t$. Differentiating KL, using $\int \partial_t\rho_t\,dw = 0$, the current form of the Fokker–Planck equation, and integration by parts gives

$$\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi_\beta) = -\frac{1}{2}\mathcal{I}(\rho_t\|\pi_\beta) + \mathbb{E}_{\rho_t}\left[(b - b_{\mathrm{rev}})^\top s_t\right], \tag{E.1}$$

where $\mathcal{I}(\rho_t\|\pi_\beta) = \mathbb{E}_{\rho_t}[s_t^\top a s_t]$.

*Step 2: defect bound.* For every $\varepsilon \in (0,1)$, Young's inequality in the $a$-metric gives

$$(b - b_{\mathrm{rev}})^\top s_t \leq \frac{1}{2\varepsilon}\|b - b_{\mathrm{rev}}\|^2_{a^{-1}} + \frac{\varepsilon}{2}\|s_t\|^2_a.$$

Taking expectation and using (E.1),

$$\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi_\beta) \le -\frac{1-\varepsilon}{2}\mathcal{I}(\rho_t\|\pi_\beta) + \frac{1}{2\varepsilon}\mathfrak{D}(t).$$

*Step 3: log-Sobolev contraction.* By the metric log-Sobolev inequality, $\mathcal{I}(\rho_t\|\pi_\beta) \ge 2\lambda_E \mathrm{KL}(\rho_t\|\pi_\beta)$. Hence

$$\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi_\beta) \le -(1-\varepsilon)\lambda_E \mathrm{KL}(\rho_t\|\pi_\beta) + \frac{1}{2\varepsilon}\mathfrak{D}(t).$$

*Step 4: Grönwall and block decomposition.* Let $K(t) = \mathrm{KL}(\rho_t\|\pi_\beta)$ and $\alpha = (1-\varepsilon)\lambda_E$. Then $K'(t) \le -\alpha K(t) + \frac{1}{2\varepsilon}\mathfrak{D}(t)$. Multiplying by $e^{\alpha t}$ and integrating from 0 to $t$ yields

$$K(t) \le e^{-\alpha t}K(0) + \frac{1}{2\varepsilon}\int_0^t e^{-\alpha(t-s)}\mathfrak{D}(s)\,ds.$$

This proves the KL bound.

For the local quadratic rate, suppose the localized Gibbs reference has potential $U_E$ satisfying $\nabla^2 U_E \succeq \mu_E I$, so that $\pi_\beta^E \propto e^{-\beta U_E}$, and suppose $a_E \succeq a_{\min} I$. Strong log-concavity gives the Euclidean log-Sobolev rate $\beta\mu_E$. Since the metric Fisher information is at least $a_{\min}$ times the Euclidean Fisher information, the metric log-Sobolev rate satisfies $\lambda_E \ge \beta\mu_E a_{\min}$.

Finally, assume $E = \oplus_{B\in\mathcal{B}} E_B$ is an orthogonal block decomposition and $a_E = \oplus_B a_{E_B}$. Writing $r_B = (b_E - b_{\mathrm{rev},E})_B$, the inverse-metric norm splits pointwise:

$$\|b_E - b_{\mathrm{rev},E}\|^2_{a_E^{-1}} = \sum_{B\in\mathcal{B}} \|r_B\|^2_{a_{E_B}^{-1}}.$$

Taking expectation under $\rho_t^E$ gives $\mathfrak{D}_E(t) = \sum_{B\in\mathcal{B}} \mathfrak{D}_{E_B}(t)$ with the block contributions defined above. This proves Theorem 4.

□

**Discussions.** The theorem has three interpretable terms. The first is initial mismatch: if the training-induced law begins far from the localized Gibbs reference, that mismatch decays at the log-Sobolev rate. The second is accumulated defect: it measures how much the training drift differs from the reversible Gibbs drift on the comparison space. The third is the comparison space $E$: KL is charged only on the directions where the posterior and reference are allowed to differ.

The role of $E$ is central in neural networks. Full parameter space contains symmetries, gauges, and loss tangents. Some of these directions have tiny minibatch-gradient covariance and small finite prediction effect, while the inverse covariance metric can make them dominate the drift-mismatch defect. The resolved comparison space focuses the KL/PAC-Bayes certificate on directions where training creates substantial local variation. This is the theorem-level explanation for the ResNet/CIFAR-10 plots in Section 5.2: the full blockwise covariance spectrum places gauge, tangent, residual, and augmentation-aligned directions inside the local geometry, and the low-covariance tail concentrates inverse-metric defect while held-out prediction movement remains small.

The block decomposition is the reason the computation is architecture-local. The potential and lifted precision from Section 2 decompose over computation-graph factors. The retained KL computation uses declared architecture-native groups. Under the block-diagonal covariance assumption above, the defect is a sum of block contributions. Finite selection among declared groups and prefixes is distinct from constructing their covariance bases and priors. The latter objects must satisfy the independence conditions of Corollary 10; a split construction can use independent build data for this purpose.

**Geometry and computation.** At checkpoint $j$, the experiments instantiate the comparison space as $E_j = \bigoplus_{B\in\mathcal{B}} E_{j,B}$, where $B$ ranges over declared architecture-native parameter groups. Examples include all parameters in a small model, convolution kernels in a CIFAR model, BatchNorm affine parameters, attention projections, or residual-stream modules. The choice of blocks determines the factorization used by the localized posterior/reference pair.

For a block $B$, collect $m$ stochastic gradients at checkpoint $j$, centered by subtracting the training drift estimate: $\delta g_{j,B}^{(s)} = g_{j,B}^{(s)} - \bar{g}_{j,B}$ for $s = 1,\dots,m$. Let $G_{j,B} \in \mathbb{R}^{m\times p_B}$ have rows $(\delta g_{j,B}^{(s)})^\top$. The

block covariance estimate is $A_{j,B} = \frac{1}{m} G_{j,B}^\top G_{j,B}$. Rather than diagonalizing this $p_B \times p_B$ matrix, we diagonalize the smaller Gram matrix $K_{j,B} = \frac{1}{m} G_{j,B} G_{j,B}^\top \in \mathbb{R}^{m \times m}$. If $K_{j,B}\alpha_{j,B,r} = \lambda_{j,B,r}\alpha_{j,B,r}$, then the corresponding parameter-space eigenvector is

$$v_{j,B,r} = \frac{1}{\sqrt{m\lambda_{j,B,r}}} G_{j,B}^\top \alpha_{j,B,r}.$$

The retained subspace $E_{j,B}$ is spanned by selected eigenvectors $v_{j,B,r}$. A finite union factor accounts for prefix selection only when the family of candidate spaces, embeddings, and priors is fixed independently of the certificate sample, conditional on any independent build data. It does not by itself account for learning that basis or a checkpoint-shaped prior from the certificate sample.

Let $r_{j,B}$ be the training drift component after subtracting the reference-prior reversible component in block $B$. Let $q_{j,B}^{(\ell)}$ be the negative bounded-loss gradient component, so that $q^{(\ell)} = -\nabla \widehat{L}^{(\ell)}$. With this convention, the reversible Gibbs drift contributes $(\beta/2)A_{j,B}q_{j,B}^{(\ell)}$ on the retained block. The checkpoint defect is

$$D_{j,E}^{(\ell)}(\beta) = \sum_{B \in \mathcal{B}} \sum_{r \in E_{j,B}} \frac{\left(\langle r_{j,B}, v_{j,B,r}\rangle - \frac{\beta}{2}\lambda_{j,B,r}\langle q_{j,B}^{(\ell)}, v_{j,B,r}\rangle\right)^2}{\lambda_{j,B,r}}. \tag{E.2}$$

Equivalently, as in (18),

$$D_{j,E}^{(\ell)}(\beta) = d_{0,j} + \beta d_{1,j}^{(\ell)} + \beta^2 d_{2,j}^{(\ell)}.$$

Expanding the square gives

$$d_{0,j} = \sum_{B,r} \frac{\langle r_{j,B}, v_{j,B,r}\rangle^2}{\lambda_{j,B,r}}, \qquad d_{1,j}^{(\ell)} = -\sum_{B,r} \langle r_{j,B}, v_{j,B,r}\rangle \langle q_{j,B}^{(\ell)}, v_{j,B,r}\rangle,$$

and

$$d_{2,j}^{(\ell)} = \frac{1}{4} \sum_{B,r} \lambda_{j,B,r} \langle q_{j,B}^{(\ell)}, v_{j,B,r}\rangle^2.$$

The square formula in (E.2) defines the checkpointwise retained defect. The term $d_{0,j}$ is inverse-covariance drift mismatch; it grows when the comparison includes directions with tiny covariance eigenvalues. The term $d_{1,j}^{(\ell)}$ measures alignment between training drift and the bounded-loss Gibbs score. The term $d_{2,j}^{(\ell)}$ measures the size of the bounded-loss score inside the retained covariance geometry.

This expansion also explains the discarded full-positive gap. If the full positive spectrum is used, directions with very small $\lambda_{j,B,r}$ enter $d_{0,j}$ with a $1/\lambda_{j,B,r}$ weight. When these directions are gauges or loss tangents, they can have small finite predictive effect but large contributions to the inverse-metric defect. A comparison on a retained factor is defined by the prior/posterior support construction. This restriction alone does not establish PAC-Bayes admissibility or turn the defect estimate into a prior KL.

For a fixed comparison space $E$ and fixed reference $\pi_\beta^E$, let $K_j = \mathrm{KL}(\rho_{t_j}^E \| \pi_\beta^E)$. Suppose that on $[t_j, t_{j+1}]$ the contraction coefficient is bounded below by $\alpha_j > 0$, and the defect is bounded above by $\overline{D}_j$. Theorem 4 then gives

$$K_{j+1} \le e^{-\alpha_j \Delta t_j} K_j + \frac{1 - e^{-\alpha_j \Delta t_j}}{2\varepsilon\alpha_j} \overline{D}_j. \tag{E.3}$$

A checkpoint estimate from (E.2) is not, by itself, a verified upper bound on the defect throughout the interval. Changing the comparison space or reference between intervals also requires additional accounting. Equation (E.3) is a formal recursion under these conditions; it was not executed for the original numerical calculation, which used a stationary defect proxy.

**Formal localized PAC-Bayes certificate.** Let $S_{\mathrm{build}}$ denote the build data together with all independent randomness used to construct the checkpoint, spaces, and priors, and suppose it is independent of the certificate sample $D_n = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$. Conditional on $S_{\mathrm{build}}$, fix a finite family $\mathcal{A}(S_{\mathrm{build}})$ of comparison spaces $E$, parameter embeddings $T_E$, priors $\pi_0^E$, and positive parameters $\beta$, without accessing $D_n$. The embedding includes any checkpoint center, basis, and fixed complement used to map $w \in E$ to a predictor. For each member, let $\ell_E(w; x, y) = \ell(T_E(w); x, y) \in [0, 1]$, and define

$$L_E(\rho) = \mathbb{E}_{w\sim\rho}\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell_E(w; x, y), \qquad \widehat{L}_{n,E}(\rho) = \mathbb{E}_{w\sim\rho}\frac{1}{n}\sum_{i=1}^n \ell_E(w; x_i, y_i).$$

We suppress the subscript $E$ below. Define $\pi_\beta^E(dw) \propto e^{-\beta\widehat{L}_{n,E}(w)}\pi_0^E(dw)$. The prior and admissible family are independent of $D_n$ conditional on build data; the Gibbs law and the posterior $\rho^E$ may depend on $D_n$.

**Corollary 10** (Localized bounded-loss PAC-Bayes certificate)**.** *Fix $\delta \in (0, 1)$. Conditional on $S_{\mathrm{build}}$, with probability at least $1 - \delta$ over $D_n$, the following inequalities hold simultaneously for every admissible choice in $\mathcal{A}(S_{\mathrm{build}})$ and every posterior $\rho^E$, including posteriors selected using $D_n$:*

$$L(\rho^E) \le \widehat{L}_n(\rho^E) + \beta^{-1}\mathrm{KL}(\rho^E\|\pi_0^E) + \beta^{-1}\log\frac{|\mathcal{A}|}{\delta} + \frac{\beta}{8n}.$$

*Equivalently,*

$$L(\rho^E) \le \mathcal{F}_\beta^E + \beta^{-1}\mathrm{KL}(\rho^E\|\pi_\beta^E) + \beta^{-1}\log\frac{|\mathcal{A}|}{\delta} + \frac{\beta}{8n}.$$

*Taking $\rho^E = \rho_t^E$ and applying Theorem 4 gives*

$$\begin{aligned} L(\rho_t^E) \le \mathcal{F}_\beta^E &+ \beta^{-1}e^{-(1-\varepsilon)\lambda_E t}\mathrm{KL}(\rho_0^E\|\pi_\beta^E) \\ &+ \frac{1}{2\varepsilon\beta}\int_0^t e^{-(1-\varepsilon)\lambda_E(t-s)}\mathfrak{D}_E(s)\,ds + \beta^{-1}\log\frac{|\mathcal{A}|}{\delta} + \frac{\beta}{8n}. \end{aligned} \tag{E.4}$$

**Proof of the PAC-Bayes certificate.**

*Proof.* Condition on $S_{\mathrm{build}}$, so that the admissible family, each embedding, and each prior are fixed independently of $D_n$. All expectations and probabilities over $D_n$ in this proof are conditional on $S_{\mathrm{build}}$. For each admissible choice and each fixed $w$, Hoeffding's lemma gives

$$\mathbb{E}_{D_n}\exp\left(\beta[L(w) - \widehat{L}_n(w)]\right) \le \exp\left(\frac{\beta^2}{8n}\right).$$

Thus

$$\mathbb{E}_{D_n}\mathbb{E}_{\pi_0^E}\exp\left(\beta[L(w) - \widehat{L}_n(w)] - \frac{\beta^2}{8n}\right) \le 1.$$

Applying Markov's inequality and a union bound over the conditionally fixed finite family $\mathcal{A}(S_{\mathrm{build}})$, with conditional probability at least $1 - \delta$, for every admissible choice,

$$\mathbb{E}_{\pi_0^E}\exp\left(\beta[L(w) - \widehat{L}_n(w)] - \frac{\beta^2}{8n}\right) \le \frac{|\mathcal{A}|}{\delta}.$$

On this event, the Donsker–Varadhan variational inequality gives, for every posterior $\rho^E$,

$$\mathbb{E}_{\rho^E}\left[\beta(L(w) - \widehat{L}_n(w)) - \frac{\beta^2}{8n}\right] \le \mathrm{KL}(\rho^E\|\pi_0^E) + \log\frac{|\mathcal{A}|}{\delta}.$$

Rearranging gives

$$L(\rho^E) \le \widehat{L}_n(\rho^E) + \beta^{-1}\mathrm{KL}(\rho^E\|\pi_0^E) + \beta^{-1}\log\frac{|\mathcal{A}|}{\delta} + \frac{\beta}{8n}.$$

The Gibbs free-energy identity

$$\widehat{L}_n(\rho^E) + \beta^{-1}\mathrm{KL}(\rho^E\|\pi_0^E) = \mathcal{F}_\beta^E + \beta^{-1}\mathrm{KL}(\rho^E\|\pi_\beta^E)$$

then gives the Gibbs-form PAC-Bayes bound. For a posterior $\rho_t^E$ satisfying Theorem 4's assumptions, substituting its KL bound gives (E.4). The event holds for every posterior, including a certificate-data-dependent posterior. Finally, averaging the conditional probability over $S_{\mathrm{build}}$ gives the same $1 - \delta$ guarantee jointly over build and certificate data.

□

The reported split-based bounds in Table 1 directly use $\mathrm{KL}(Q\|P)$ for an explicit prior and posterior on the same support, with the prior and construction fixed from independent build data. They use the prior-KL form of Corollary 10; they do not use $K_{\mathrm{stat}}$, the Itô/LSI assumptions of Theorem 4, or Eq. (E.4). The split protocol is given in Appendix B.2.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 describe the lifted information model, checkpoint-local criticality, and retained-space KL analysis. Section 7 distinguishes the exact local results, diagnostic proxies, and the separate split-based prior-KL certificates.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses local linearization, the stated information model, separator complexity, and the retained-space construction. Section 5 and Appendix B distinguish sampled diagnostics from the exact operators in the theory.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sections 2–4 state the local-model, state-elimination transport, and retained-space assumptions. Appendices C, D, and E give the corresponding proofs and the separate prior/posterior construction for PAC-Bayes comparison.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Section 5 and Appendix B define the diagnostics and the split-based prior-KL protocol. Complete per-run configurations and all certificate-sample and Monte Carlo settings are not reported for every empirical result.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper cites public data sources and describes the diagnostic computations, but does not identify a code-release location with complete commands and environment instructions mapping to every reported experiment.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Section 5 and Appendix B specify model families, checkpoint measurements, and the certificate protocol, but do not enumerate all training schedules, checkpoint selections, seeds, and probe settings used in the reported runs.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The LR-cliff brackets locate instability thresholds, and the PAC-Bayes and Monte Carlo failure budgets quantify certificate confidence. The paper does not report run-level error bars or significance tests for the checkpoint-diagnostic comparisons.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix B explains the Gram-matrix, active-subspace, and boundary-gradient computations and their dimensions, but does not report a complete per-experiment inventory of hardware, memory use, and execution time.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses existing vision and text datasets for theoretical analysis and checkpoint diagnostics, with no new participant recruitment or deployed decision system.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work develops foundational theory and checkpoint diagnostics for neural-network training, without an application-specific deployment. No direct application-specific societal impact is identified.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work reports checkpoint diagnostics and does not introduce a deployed model, a new high-risk pretrained model, or a newly collected dataset. The GPT-20M/FineWeb-Edu run is a diagnostic testbed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites datasets, model architectures, and augmentation methods, but does not explicitly enumerate the versions, licenses, and terms of use of all experimental assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The diagnostic and certificate-audit procedures are described in Appendix B, but the paper does not identify a code package with accompanying documentation and complete reproduction instructions.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No new crowdsourcing or human-participant study was conducted for this work; the experiments use existing vision and text datasets.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No participants were recruited and no new human-subject study was conducted for this work; the experiments analyze existing benchmark data.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An LLM assisted this revision with prose editing and mathematical consistency checks, including clarification of assumptions and derivations. The GPT-20M training experiment is a testbed for the checkpoint diagnostics.