AFRL

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

WEN DONG

AIR FORCE RESEARCH LABORATORY · NEURIPS 2026

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.

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

Local interactions create global curvature

Why do some network designs learn and generalize well, while others fail?

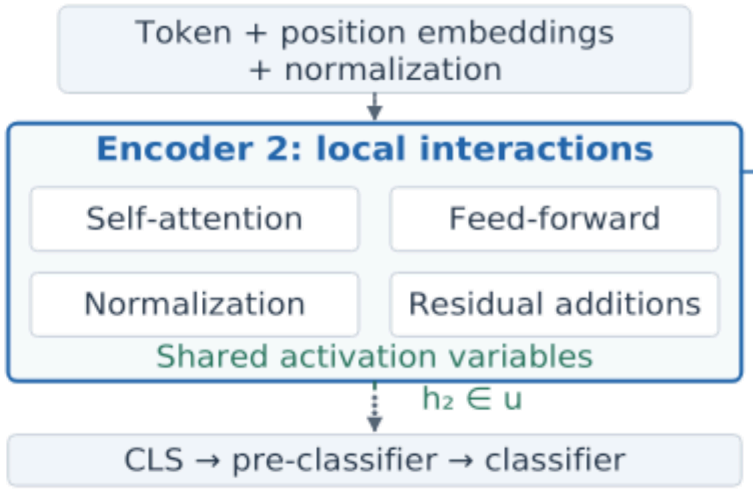
$$E_n(z; D_n) = \sum_{f \in \mathcal{F}_n} E_f(z_{S_f}; D_n), \qquad z = (\theta, u)$$

θ : parameters, u : activations, S_f : factor scope, D_n : data

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

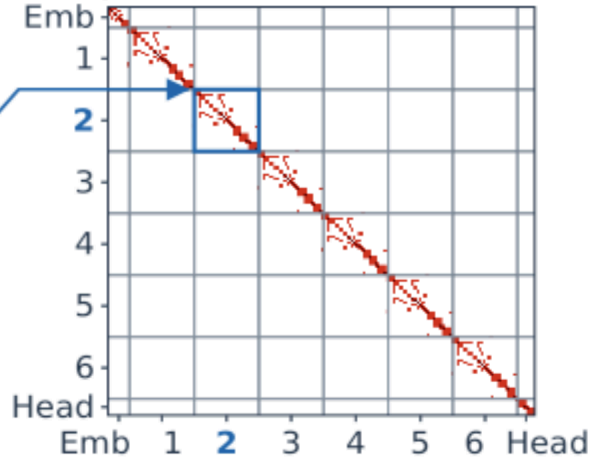
$Z_{n,\beta}$: partition function, $\beta > 0$: inverse temperature, m : reference measure

Transformer architecture



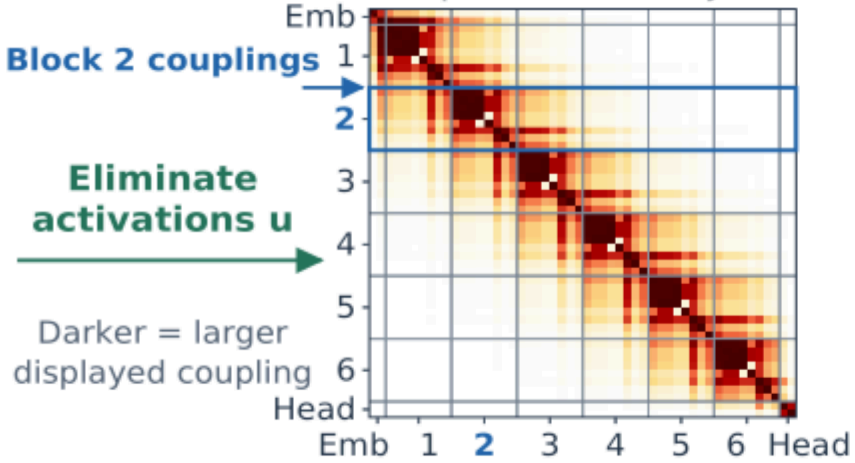
Joint precision $P_{\theta,u}$

parameters + activations



Reduced precision F_τ

parameters only



Eliminate activations:

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

Eliminating activations induces longer-range weight couplings.

Camera-ready paper · Section 2 and Figure A.2 · finite-precision structural illustration

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

Theorem 1: activation elimination recovers curvature

At a forward-consistent checkpoint, keep the local Jacobians and information weights fixed.

1 Assemble the local quadratic costs

Each local Jacobian maps joint perturbations into changes of a module or loss residual.

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

$$J_f = Dr_f:\ \text{residual response}, \quad W_f \succeq 0:\ \text{information weight}, \quad \Lambda_\theta:\ \text{prior/regularizer precision}$$

Embed places each contribution on its participating variables. τ is the module-relation variance.

2 Hold the parameter change fixed; let activations respond

$$\min_{\Delta u} \frac{1}{2}\Delta z^\top P_{\theta,u}(\tau)\Delta z = \frac{1}{2}\Delta\theta^\top F_\tau \Delta\theta, \qquad \Delta z = (\Delta\theta, \Delta u)$$

Complete the square: activation changes choose their least-cost response in this joint quadratic.

3 Tighten the module constraints

Assume the weighted linearized module equations uniquely determine the activation response.
As τ decreases to zero, that response is forced to follow the linearized forward computation.

$$F_\tau \uparrow F_{\mathrm{red}} \qquad (\tau \downarrow 0)$$

The limit is the chain-rule Fisher / generalized Gauss–Newton curvature.

Camera-ready paper · Theorem 1 and Appendix C · checkpoint information model; fixed local Jacobians and weights

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

Theorems 2–3: stable updates can lose plasticity

Theorem 2 · Trace a parameter perturbation to the output

$$\|J^{\mathrm{red}}_{i,v}\|_{\mathrm{op}} \le S_{v,i} = \gamma_i \sum_{C\in\mathcal{C}(v)} s^{(C)}_{v,i} \prod_{e\in\mathrm{path}(C\rightsquigarrow C_{\mathrm{out}})} g_e$$

J: output sensitivity for sample i and parameter block v. C: injection clique; s: injection norm.
g: interface-transfer norm; γ: readout norm. Products follow interfaces; the sum adds source routes.
Loss-weighted sensitivities bound upper curvature. Active coverage rules out loss-invisible directions.

Theorem 3 · Stability and response in optimizer coordinates

$$\bar{F} = P^{1/2} F_{\mathrm{red}} P^{1/2}, \qquad P \succ 0 \text{ is the optimizer metric}$$

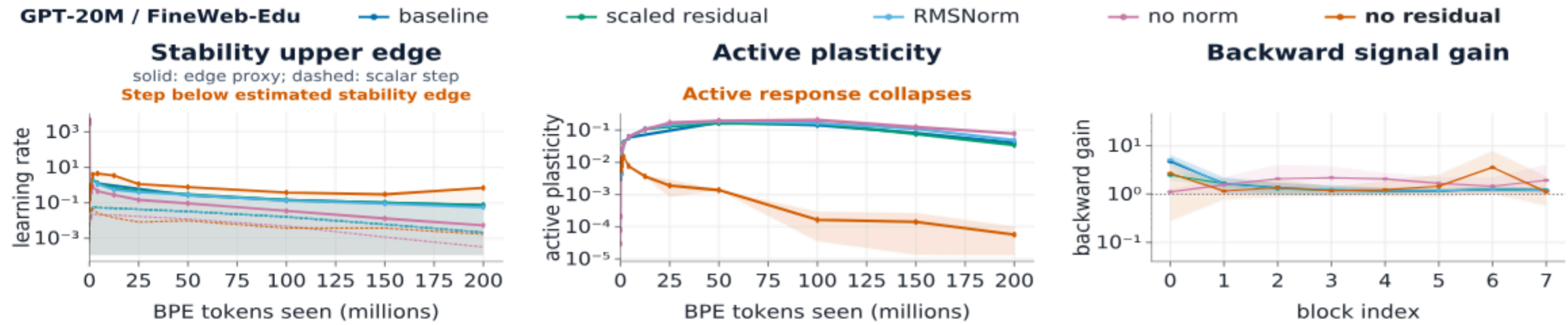
$$\eta_{\max} = \frac{2}{\lambda_{\max}(\bar{F})}, \qquad \rho_{\mathrm{act}} = \eta_{\mathrm{eff}}\, \lambda_{\min}(\bar{F}|_{\mathcal{U}_{\mathrm{act}}})$$

Step limit — Effective step × minimum curvature on the probed active span

Each linearized mode is multiplied by

$$1 - \eta\lambda$$

magnitude > 1 → growth
near +1 → little response



Camera-ready paper · Theorems 2–3 and Figures B.1–B.3 · empirical-Fisher / directional-gradient proxies

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

Theorem 4: entropy decay competes with drift mismatch

Use graph-derived precision to choose a Gibbs reference on a fixed parameter subspace E.
Minibatch dynamics separately provide mean motion (drift) and fluctuations (covariance).

$$\pi_\beta^E(dw) \propto e^{-\beta \hat{L}_E(w)} \pi_0^E(dw) \qquad \mathfrak{D}_E(t) = \mathbb{E}_{\rho_t^E}\left[\|b_E - b_{\mathrm{rev},E}\|_{a_E^{-1}}^2\right]$$

ρ_t^E : local training law, $\hat{L}_E$: bounded empirical loss, π_0^E : reference prior, λ_E : relaxation rate
b_E : training drift, $b_{\mathrm{rev},E}$: Gibbs-preserving drift, a_E : covariance rate

$$\frac{d}{dt}\mathrm{KL}(\rho_t^E \| \pi_\beta^E) \le \underbrace{-(1-\varepsilon)\lambda_E \,\mathrm{KL}(\rho_t^E \| \pi_\beta^E)}_{\text{relaxation toward Gibbs}} + \underbrace{\frac{\mathfrak{D}_E(t)}{2\varepsilon}}_{\text{drift-mismatch bound}}$$

Integrate relaxation and the drift-mismatch bound to control KL.

Fixed reference · local diffusion · positive covariance · metric log-Sobolev inequality · 0 < ε < 1

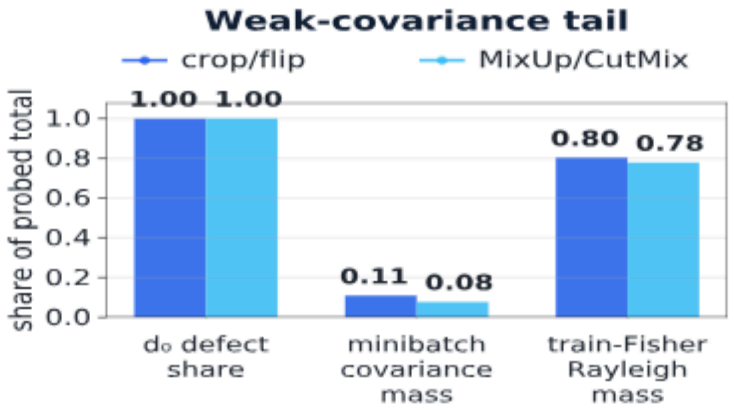
Where does the cost live?

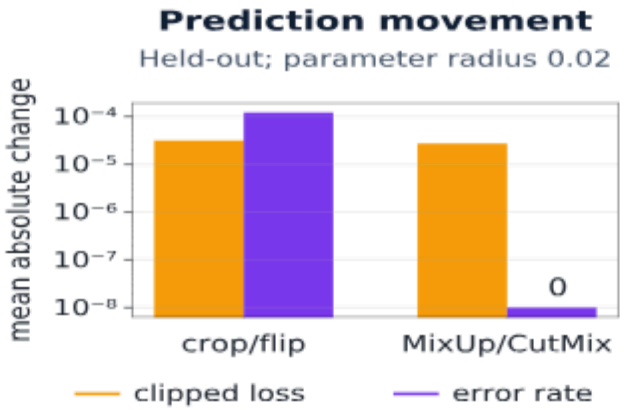
A mismatch costs more
where covariance is small.

$$\frac{m^2}{\sigma^2}$$

m: drift mismatch
σ²: directional variance

The tail dominates the drift term d_0, yet finite moves barely change held-out predictions.





Camera-ready paper · Theorem 4 proof, Appendix E · Figure 3(a), ResNet-32 / CIFAR-10

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

PAC-Bayes certificates for retained posteriors

Free energy balances data fit and a temperature-weighted prior cost.
A posterior is a distribution over weights. Risk averages over draws and new examples.

$$\underbrace{\bar{L}_E(\rho) + \frac{\mathrm{KL}(\rho\|\pi_0^E)}{\beta}}_{\text{free-energy objective}} = \underbrace{\mathcal{F}_\beta^E}_{\text{Gibbs minimum}} + \underbrace{\frac{\mathrm{KL}(\rho\|\pi_\beta^E)}{\beta}}_{\text{excess free energy}}$$

For the certificate, use the same bounded certificate-sample loss in the Gibbs reference and free energy.

$$L(\rho^E) \le \underbrace{\bar{L}_n(\rho^E)}_{\text{certificate loss}} + \underbrace{\frac{\mathrm{KL}(\rho^E\|\pi_0^E)}{\beta}}_{\text{posterior complexity}} + \underbrace{\frac{\log(|\mathcal{A}|/\delta)}{\beta} + \frac{\beta}{8n}}_{\text{confidence and sample-size cost}}$$

β: inverse temperature n: certificate sample size |A|: fixed family size δ: failure probability

Direct prior-KL certificate: freeze the construction on build data, then certify on independent data.

Model	k	Prior KL	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

ResNet-32: bound 0.07657; held-out posterior loss 0.05840. All five bounds are nonvacuous.

k: retained dimension · normalized clipped NLL · simultaneous 95% · MNIST: first two rows; CIFAR-10: remaining rows

Camera-ready paper · Free-energy identity, Corollary 10 and Table 1 · table uses direct prior KL

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited

Small Gram matrices make both analyses practical

Millions of parameters, but only m backpropagated gradient probes.
Stack them as rows of G; diagonalize the m × m Gram matrix.

$$\operatorname{spec}_+\left(\underbrace{G^\top G/m}_{p\times p}\right) = \operatorname{spec}_+\left(\underbrace{GG^\top/m}_{m\times m}\right), \qquad G \in \mathbb{R}^{m\times p}$$

Paper

Proof idea: both matrices contain the same nonzero squared singular values, divided by m.

Uncentered gradients: curvature proxies

$$\tilde{g}^{(s)} = P^{1/2} g^{(s)}, \qquad \tilde{G} \text{ has rows } (\tilde{g}^{(s)})^\top$$

$$\tilde{F} = \frac{\tilde{G}^\top \tilde{G}}{m}$$

Estimate the upper edge and active response from Slide 4 in the optimizer metric P.

The spectrum is exact for these probes;
the sampled curvature is an empirical-Fisher proxy.

Centered gradients: covariance estimates

$$\delta g_B^{(s)} = g_B^{(s)} - \bar{g}_B, \qquad G_B \text{ has rows } (\delta g_B^{(s)})^\top$$

$$A_B = \frac{G_B^\top G_B}{m}, \qquad K_B = \frac{G_B G_B^\top}{m}$$

$$K_B \alpha_r = \lambda_r \alpha_r \quad \Longrightarrow \quad v_{B,r} = \frac{G_B^\top \alpha_r}{\sqrt{m\lambda_r}}$$

$$\lambda_r > 0, \quad \|\alpha_r\|_2 = 1, \qquad B: \text{ parameter block}$$

Recover directions for E and inverse-variance weights in the drift defect on Slide 5.

Local interactions explain the geometry. Gradient probes make it measurable.

Camera-ready paper · Appendices B and E · exact sampled spectra; covariance needs optimizer/time calibration

DISTRIBUTION STATEMENT A / Approved for public release; distribution is unlimited