RELIABLE CHARACTERIZATION OF LOCAL MODEL COMPLEXITY WITH VALID LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding training dynamics and learned models requires more than evaluating loss or predictive performance. Modern deep neural networks (DNNs) are often overparameterized, and solutions with similar performance can have different local structure, which motivates a principled measure of model complexity around a solution. Singular learning theory provides such a framework, and the local learning coefficient (LLC) quantifies local model complexity, including cases where parameter symmetries and redundancies make distinct parameter values represent the same predictive function. A scalable sampling-based approach estimates LLC around a target solution using Gaussian localization. Because the localization penalty is part of the sampled target, its strength can change the measured local model complexity. Weak localization can admit remote low-loss regions, whereas excessive localization can impose an artificial contraction scale and distort the target complexity. By resolving the loss and localizer in a common geometric representation, we derive model-agnostic localization conditions under which localization-based measurements asymptotically converge to the LLC of the target solution across a broad class of analytic models, without estimating model-specific singular geometry. We further establish classwise counterexamples showing that neither the lower nor upper localization requirement admits a uniform relaxation over this analytic model class. The guarantee is asymptotic. For finite computations, we separate the discrepancy between the population localized measurement and the target LLC from empirical and numerical errors, and derive exact diagnostics that quantify sensitivity to localization. Experiments on truth-known singular models and DNN training trajectories show that theory-guided localization yields reliable LLC measurements and reveals reproducible complexity changes during grokking and learning-rate transitions. These results establish when localization-based LLC measurements can be interpreted as properties of a target solution rather than consequences of the localization choice.

1 INTRODUCTION

Deep neural networks (DNNs) have driven major advances in computer vision, language modeling, and scientific discovery (Krizhevsky et al., 2012; He et al., 2016; Vaswani et al., 2017; Brown et al., 2020; Jumper et al., 2021). Yet, their learning mechanisms remain difficult to characterize. Training can depart greatly from the classical picture of smooth convergence to an isolated minimum. Examples include delayed generalization and non-monotone dynamics at large learning rates, while solutions obtained from independent training runs can remain connected through low-loss regions (Power et al., 2022; Cohen et al., 2021; Lewkowycz et al., 2020; Garipov et al., 2018; Draxler et al., 2018). These observations motivate geometric descriptions of parameter states encountered during training and solutions obtained at its end, beyond loss or predictive performance alone.

A natural question is whether small parameter perturbations substantially change the model or whether many nearby parameter values realize similar behavior. A local complexity measure should capture this distinction. Parameter count alone does not distinguish such local structure, while Hessian-based curvature and parameter-space flatness measures can change under reparameterizations that leave the model's input-output function unchanged (Dinh et al., 2017; Jiang et al., 2020; Simsek et al., 2021; Lim et al., 2024). Classical asymptotic statistics provides a rigorous local complexity theory when nearby parameters are locally distinguishable and the log likelihood is well approximated by a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

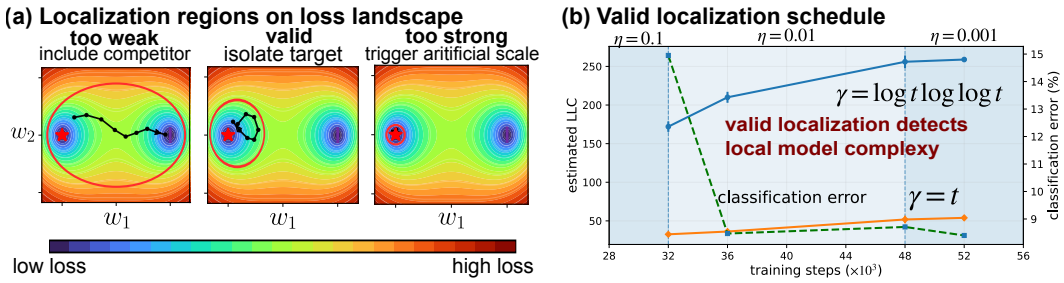


Figure 1: Valid localization is required before local model complexity is interpreted as a property of a target solution. (a) Three localization regimes on a loss landscape. Weak localization admits a remote low-loss competitor, valid localization isolates the target neighborhood, and excessive localization imposes an artificial local scale. (b) Example measurement during grokking. Two admissible schedules show the same qualitative decrease as test generalization emerges.

full-rank quadratic form (van der Vaart, 1998; Schwarz, 1978). DNNs often violate these conditions because symmetries and redundant parameterizations allow distinct weights to represent the same function, while higher-order flat directions make the loss depart from quadratic behavior (Wei et al., 2023; Simsek et al., 2021; Lim et al., 2024). Singular learning theory (SLT) replaces ordinary dimension in this setting with geometric invariants. Its learning coefficient (LC), governs leading corrections to generalization error and model complexity (Watanabe, 2009; 2013).

The local learning coefficient (LLC) is the corresponding solution-specific quantity and admits scalable estimation in DNNs (Lau et al., 2025). LLC and related local observables have been used to study developmental stages, attention specialization, transient computation, grokking, and compressibility (Chen et al., 2023; Hoogland et al., 2025; Wang et al., 2025; Carroll et al., 2025; Urdshals & Urdshals, 2025; Cullen et al., 2026; Urdshals et al., 2025). A scalable sampling-based estimator localizes sampling around a target solution using a Gaussian penalty (Lau et al., 2025). Because this penalty changes the sampled target, its strength can affect the estimated local complexity. Fig. 1(a) shows the two failure modes. Weak localization can leave a remote low-loss region dominant, whereas excessive localization can impose an artificial contraction scale. The central question is when Gaussian localization isolates the target solution without changing its intrinsic local complexity.

This study develops a validity theory for the Gaussian localization used in scalable LLC estimation (Lau et al., 2025). We characterize when localization isolates a target solution while preserving its pointwise LLC. Our analysis compares the loss and localization scales in a common geometric representation and yields model-agnostic conditions without estimating model-specific singular geometry. We also separate finite-scale target error from empirical and numerical errors and derive exact response diagnostics for sensitivity to localization. Finally, we apply the framework to controlled singular models and DNN training trajectories to determine whether observed LLC changes reflect model geometry rather than localization artifacts. Fig. 1(b) illustrates this use during learning. Our contributions are summarized as follows.

- **Problem formulation.** We formulate localization validity as isolating a target solution while preserving its pointwise LLC, and separate localization discrepancy from empirical and numerical errors.
- **Theoretical localization guarantee.** We prove model-agnostic localization conditions and class-wise boundary counterexamples without estimating Hessian rank, singular order, or model-specific thresholds.
- **Empirical characterization.** We verify the predicted localization regimes under known truth and use validated LLC measurements to characterize geometric changes during grokking and learning-rate transitions.

2 Related Work

SLT and LLC. SLT extends statistical learning theory to models where regular quadratic approximations fail, replacing ordinary dimension with geometric quantities that govern generalization and

model complexity (Watanabe, 2001; 2009; 2013; 2018). Exact or analytic learning coefficients (LCs) are known for Gaussian mixtures, reduced-rank regression, hidden Markov models, restricted Boltzmann machines, deep linear networks (DLNs), and semiregular models (Yamazaki & Watanabe, 2003; Aoyagi & Watanabe, 2005; Yamazaki & Watanabe, 2005; Aoyagi, 2013; 2024; Kurumadani, 2025a;b; 2026). These model-specific results provide analytic reference points. We build on this foundation and study localization validity toward LLC estimation for large-scale DNNs without requiring model-specific singular geometry.

LLC estimation. Scalable sampling-based estimation makes LLC practical for DNN analysis (Lau et al., 2025). A scalable local-posterior benchmark compares stochastic-gradient MCMC methods on DNNs (Hitchcock & Hoogland, 2025). Finite-temperature analysis shows that practical LLC estimates characterize the geometry of a temperature-dependent effective data distribution and can remain meaningful away from an exact population minimizer (Chen & Murfet, 2025). Shift-invariant variance estimation removes dependence on an unknown additive loss baseline and separates geometric loss fluctuations from mini-batch noise in off-equilibrium measurements (Cai, 2026). Complementary diagnostics characterize singular geometry from Fisher degeneracies and architecture-specific structure without posterior sampling (Shirodkar, 2026; Shirodkar & Narayanan, 2026a;b). We complement these advances by deriving validity conditions for the Gaussian localization used in scalable LLC estimation.

LLC for learning dynamics and learned solutions. LLC are increasingly used to characterize learning dynamics and the structure of learned solutions, such as stagewise development, attention specialization, transient computation, grokking, and compressibility (Chen et al., 2023; Hoogland et al., 2025; Wang et al., 2025; Carroll et al., 2025; Urdshals & Urdshals, 2025; Cullen et al., 2026; Urdshals et al., 2025). LLC has also been connected theoretically to long-time the stochastic gradient descent (SGD) dynamics (Hennick & Baerdemacker, 2026). Related local-posterior response methods use susceptibilities for structural inference and clustering, while linear-response theory establishes consistent susceptibility estimators and local Bayesian influence functions extend this covariance-based view to Hessian-free data attribution (Baker et al., 2026; Gordon et al., 2026; Elliott & Murfet, 2026; Kreer et al., 2026). These applications largely build on localization-based sampling (Lau et al., 2025). We provide a model-agnostic theory that supplies validity conditions for this established localization mechanism.

In summary, prior work develops SLT theory, scalable LLC estimation, and applications to learning dynamics and learned solutions. Our contribution provides theoretical support for localization-based LLC estimation by identifying when Gaussian localization preserves the intended local model complexity.

3 Problem Setup

We define the local model complexity target and localization needed to isolate a target learned solution while preserving its local model complexity.

Data and excess loss. Let $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote a random variable for an input or feature and $Y \in \mathcal{Y}$ a random variable for output or target, where $\mathcal{Y} \subset \mathbb{R}^{d_y}$ for regression and $\mathcal{Y} = \{1, \ldots, c_Y\}$ for classification with $c_Y \in \mathbb{N}$. Uppercase letters denote random variables and lowercase letters their realizations. The true joint distribution is $q_{X,Y}(X, Y) = q_{Y|X}(Y \mid X)q_X(X)$, from which we observe n i.i.d. samples $D_n = \{(x_i, y_i)\}_{i=1}^n$. A parameter $w \in \mathcal{W} \subset \mathbb{R}^{d_w}$ specifies a conditional model $p(Y \mid X; w)$, or more generally an input-output predictor, with per-example loss $\ell(x, y; w)$. For likelihood-based models, $\ell(x, y; w) = -\log p(y \mid x; w)$. For a population minimizer w^*, we define $L(w) \triangleq \mathbb{E}_{(X,Y)\sim q_{X,Y}}[\ell(X, Y; w)]$, $L_n(w) \triangleq n^{-1}\sum_{i=1}^n \ell(x_i, y_i; w)$, $K(w) \triangleq L(w) - L(w^*)$, and $K_n(w) \triangleq L_n(w) - L_n(w^*)$. For a well-specified likelihood model, $K(w) = \mathbb{E}_{X\sim q_X}[D_{\mathrm{KL}}(q_{Y|X}(t \mid X)\|p(Y \mid X; w))]$, where, for each x such that $q_{Y|X}(Y \mid x)$ is absolutely continuous with respect to $p(Y \mid x; w)$, $D_{\mathrm{KL}}(q_{Y|X}(Y \mid x)\|p(Y \mid x; w)) \triangleq \mathbb{E}_{Y\sim q_{Y|X}(Y|x)}[\log\left(dq_{Y|X}(Y \mid x)/dp(Y \mid x; w)\right)]$ (Watanabe, 2009).

Neighborhood and pointwise LLC. We define $\mathcal{W}_0 \triangleq \{w \in \mathcal{W} \mid K(w) = 0\}$. In singular models, $\mathcal{W}_0$ can contain positive-dimensional sets of optima with location-dependent degeneracies. Let $\omega : \mathcal{W} \to \mathbb{R}_{\geq 0}$ be a fixed smooth background density, or prior weight, positive near w^*. For measurable $\mathcal{A} \subset \mathcal{W}$, define $Z_{\mathcal{A}}(t) \triangleq \int_{\mathcal{A}} e^{-tK(w)}\omega(w)\,dw$. Then, $Z_{\mathcal{W}}(t)$ is the full-space partition function, while relatively compact $\mathcal{V} \subset \mathcal{W}$ that contains w^* gives $Z_{\mathcal{V}}(t)$, whose leading power-law exponent is the neighborhood LLC $\lambda(\mathcal{V})$. The pointwise LLC $\lambda(w^*)$ is obtained by shrinking $\mathcal{V}$ to w^*, equivalently through sufficiently small balls $B_r(w^*)$ followed by $r \downarrow 0$ under standard SLT conditions (Watanabe, 2009; Lau et al., 2025). Appendix B gives the formal asymptotic definition and role of $\mathcal{W}_0$. A smaller LLC means slower decay of localized partition mass as tolerance around zero excess loss tightens, leaving comparatively more weighted parameter volume near the solution; a larger LLC means faster contraction. Therefore, LLC distinguishes locally broader from more constrained singular geometry.

Example 1 (Location-dependent local model complexity)**.** For $K(x, y) = x^2y^4$, every point on either coordinate axis has zero excess loss. Near $(0, 1)$, K behaves like x^2 and LLC is $1/2$, whereas near $(1, 0)$ it behaves like y^4 and LLC is $1/4$. An unrestricted integral is eventually dominated by the slower-decaying $1/4$ region even when $(0, 1)$ is the target solution. Therefore, localization is necessary to isolate the intended part of $\mathcal{W}_0$.

Background density and algorithmic localizer. The fixed background density ω and Gaussian localizing prior used in scalable LLC estimation have different roles. In SLT integrals, $\omega(w)$ is fixed and independent of localization. In the sampling-based estimator (Lau et al., 2025), locality around w^* is imposed by $\phi_\gamma(w - w^*) \propto \exp\left(-\gamma\|w - w^*\|_2^2/2\right)$. We separate this tunable localization by writing $\omega(w)e^{-\gamma q(w)}$, where $q : \mathcal{W} \to \mathbb{R}_{\geq 0}$ is nonnegative, satisfies $q(w^*) = 0$, and is independent of γ. For the Gaussian localizer, $q(w) \triangleq \|w - w^*\|_2^2/2$, so $e^{-\gamma q(w)} = \exp\left(-\gamma\|w - w^*\|_2^2/2\right)$ while $\omega(w)$ remains fixed. When ω is locally constant, $\omega(w)e^{-\gamma q(w)}$ reproduces the localization dependence of ϕ_γ up to normalization. More generally, it applies the same Gaussian localization over a fixed smooth positive background density. Therefore, ω is the fixed SLT background weight, whereas ϕ_γ contains the γ-dependent sampling localization. For measurable $\mathcal{A} \subset \mathcal{W}$, we define the partition function and the Gibbs distribution by

$$Z_{\mathcal{A}}(t, \gamma) \triangleq \int_{\mathcal{A}} e^{-tK(w)-\gamma q(w)}\omega(w)\,dw, \quad \pi^{\mathcal{A}}_{t,\gamma}(dw) \triangleq \frac{e^{-tK(w)-\gamma q(w)}\omega(w)\,dw}{Z_{\mathcal{A}}(t, \gamma)}, \quad w \in \mathcal{A}. \tag{1}$$

The unlocalized case is $Z_{\mathcal{A}}(t) = Z_{\mathcal{A}}(t, 0)$. Thus, $Z_{\mathcal{V}}(t, \gamma)$ is localized to the target neighborhood, whereas $Z_{\mathcal{V}}(t) = Z_{\mathcal{V}}(t, 0)$ defines $\lambda(\mathcal{V})$. For the full space, we write $Z(t, \gamma) \triangleq Z_{\mathcal{W}}(t, \gamma)$ and $\pi_{t,\gamma} \triangleq \pi^{\mathcal{W}}_{t,\gamma}$. We define the finite-temperature local model complexity observable by $\Lambda(t, \gamma) \triangleq t\mathbb{E}_{\pi_{t,\gamma}}[K]$. For the widely applicable Bayesian information criterion (WBIC), $t = n\beta$ with conventional $\beta = 1/\log n$ (Watanabe, 2013; Lau et al., 2025).

Valid localization problem. The strength γ must suppress competing low-loss regions while preserving the target solution's local model complexity. We Define $B_{\mathrm{loc}}(t, \gamma) \triangleq |\Lambda(t, \gamma) - \lambda(w^*)|$. Our goal is to characterize localization strengths that isolate the target while preserving the intended LLC and connect this validity question to finite-scale measurements. Section 4 derives model-agnostic growth conditions on γ_t ensuring $B_{\mathrm{loc}}(t, \gamma_t) \to 0$ as $t \to \infty$ without estimating model-specific singular geometry, together with analytic counterexamples at both boundaries. Section 5 develops the corresponding finite-scale workflow and diagnostics.

4 Valid Localization Theory

In this section, we derive localization conditions under which the localized measurement converges to the target pointwise LLC without computing model-specific singular exponents. In standard SLT, resolution converts singular geometry around the target solution w^* into comparable power laws. Fig. 2(A) shows this standard construction. Fig. 2(B) adds the localizer q to the same resolved coordinates, and Fig. 2(C)–(D) present our localization-specific analysis and model-agnostic schedule. We refer to the corresponding parts of Fig. 2 as the argument proceeds.

Resolved local geometry. LLC summarizes contraction of weighted low-loss volume around w^* under the excess loss K, while the localizer q restricts measurement to that target. Original

coordinates obscure comparison because K can be nonquadratic and symmetries or redundancies can make different parameters represent the same model. Standard SLT resolves $\mathcal{V} \subset \mathcal{W}$ through a nonsingular real-analytic manifold $\mathcal{M}$ covered by finite charts (Hironaka, 1964; Watanabe, 2009; 2013; 2018). Real analytic means local formulas equal their convergent Taylor series. Resolution preserves dimension d_w and replaces singular behavior by regular coordinates. In each chart, K and the transformed weighted measure have normal-crossing form, a monomial times a positive nonvanishing analytic factor, exposing their vanishing orders. Standard principalization (Hironaka, 1964; Atiyah, 1970; Bierstone & Milman, 1997) gives a common refinement where q has the same form. Fig. 2(B) makes the loss, localization, and weighted-volume orders directly comparable.

Let $\mathcal{I}$ index the charts. For each $\alpha \in \mathcal{I}$, $U_\alpha \subset \mathbb{R}^{d_w}$ denotes a local coordinate domain on the resolved manifold $\mathcal{M}$, and $g_\alpha : U_\alpha \to \mathcal{V}$ is the resolution map expressed in these coordinates. Their images cover $\mathcal{V}$ but need not all contain w^*. For $u = (u_1, \ldots, u_{d_w}) \in U_\alpha$ and $\xi = (\xi_1, \ldots, \xi_{d_w})$, we define $M_\xi(u) \triangleq \prod_{j=1}^{d_w} u_j^{\xi_j}$. A zero exponent removes that coordinate from the leading order.

Assumption 1 (Resolved localization setting)**.** *We assume relatively compact $\mathcal{V} \subset \mathcal{W}$ that contains w^* and finite analytic maps $g_\alpha : U_\alpha \to \mathcal{V}$, $\alpha \in \mathcal{I}$, whose images cover $\mathcal{V}$ but need not all contain w^*. After finite orthant decomposition, write $u = (u_1, \ldots, u_{d_w}) \in U_\alpha$ and $M_\xi(u) \triangleq \prod_{j=1}^{d_w} u_j^{\xi_j}$, where zero ξ_j means coordinate j is absent from the leading order. For every α, assume (i) $K \circ g_\alpha = A_\alpha M_{k_\alpha}$, (ii) $q \circ g_\alpha = B_\alpha M_{r_\alpha}$, and (iii) $\omega(g_\alpha)|\det Dg_\alpha| = C_\alpha M_{d_\alpha - \mathbf{1}}$, where $k_\alpha, r_\alpha \in \mathbb{N}_0^{d_w}$, $d_\alpha \in \mathbb{R}_{>0}^{d_w}$, and positive analytic units $A_\alpha, B_\alpha, C_\alpha : U_\alpha \to \mathbb{R}_{>0}$ are bounded above and below on compact chart supports. We further require $\inf_{w \notin \mathcal{V}} q(w) \geq q_0 > 0$ and that there exists a fixed $\bar{\gamma} \geq 0$ such that $\int_{\mathcal{W}} \exp(-\bar{\gamma} q(w))\omega(w)\, dw < \infty$.*

Remark. Standard SLT gives the normal-crossing forms for K and the weighted measure (Watanabe, 2009; 2013). Our localization-specific refinement adds q, making its vanishing order directly comparable with that of K. The condition $q \geq q_0$ outside $\mathcal{V}$ separates the target neighborhood from remote regions, holds when q vanishes only at the target solution and the relevant complement is compact, and is stated explicitly for noncompact parameter spaces. The weighted integrability condition is weaker than requiring $\int_{\mathcal{W}} \omega < \infty$ and, for a coercive Gaussian localizer, also allows backgrounds that are not integrable without localization. We call q target-isolating on $\overline{\mathcal{V}}$ when $q^{-1}(0) \cap \overline{\mathcal{V}} = \{w^*\}$. Let $\mathcal{I}_* \subseteq \mathcal{I}$ index charts whose compactly supported images approach w^*, equivalently whose closures contain w^*. These are the charts relevant to the pointwise LLC. For target-isolating q, including $q(w) = \|w - w^*\|_2^2/2$, every chart in $\mathcal{I} \setminus \mathcal{I}_*$ has a strictly positive localizer gap on its compact support. Assumption 1 permits nonisolated minima, rank-deficient Hessians, and nonquadratic loss behavior. The proof-only vectors k_α, r_α, and d_α encode loss vanishing, localization vanishing, and transformed volume cost, respectively, and are not estimated numerically.

Fixed power phase diagnostic. To test when localization changes the leading power law, we temporarily set $\gamma = t^a$ for fixed $a > 0$, placing loss and localization growth on the common scale

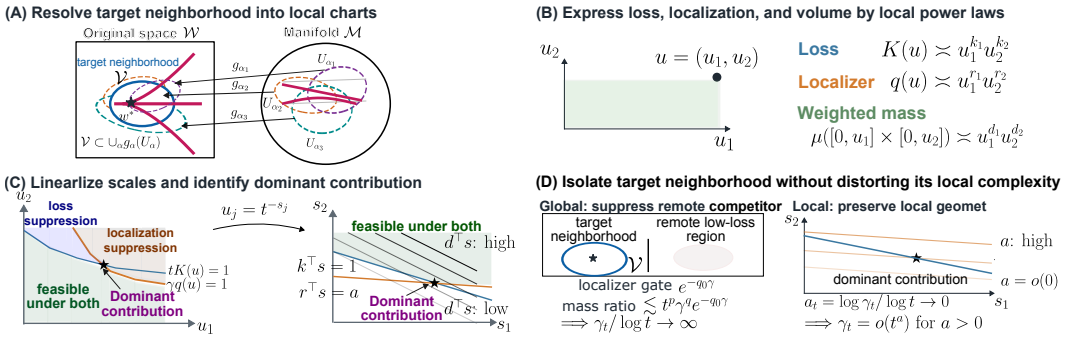


Figure 2: Overview of the valid localization theory. (A) Standard SLT resolves $\mathcal{V}$ into local charts. (B) Adding q puts K, q, and the weighted measure in comparable monomial form. (C) Minimum transformed volume cost on the feasible contraction-rate polyhedron gives the chartwise localized decay exponent. (D) A model-agnostic schedule suppresses remote competitors while preserving the minimizing face.

$\log t$. Write $u_j = t^{-s_j}$ with $s = (s_1, \ldots, s_{d_w}) \in \mathbb{R}^{d_w}_{\geq 0}$. As in Figure 2(C), weighted volume has order $t^{-d_\alpha^\top s}$, while tK and γq have orders $t^{1-k_\alpha^\top s}$ and $t^{a-r_\alpha^\top s}$. Therefore, we define

$$\mathcal{S}_\alpha(a) \triangleq \left\{ s \in \mathbb{R}^{d_w}_{\geq 0} \mid k_\alpha^\top s \geq 1,\ r_\alpha^\top s \geq a \right\}, \quad \psi_\alpha(a) \triangleq \min_{s \in \mathcal{S}_\alpha(a)} d_\alpha^\top s. \tag{2}$$

We call $\mathcal{S}_\alpha(a)$ the *feasible contraction-rate polyhedron* and $\psi_\alpha(a)$ the *chartwise localized decay exponent*. Equivalently, $\psi_\alpha(a)$ is the minimum transformed volume cost compatible with both the loss and localization constraints. At $a = 0$, the localization constraint is automatic and $\psi_\alpha(0) = \lambda_\alpha$ is the standard chartwise LLC from normal-crossing asymptotics (Watanabe, 2009; Atiyah, 1970); for $a > 0$, the second constraint is localization-specific. For the target-relevant charts, a chart contributes at power-law order $t^{-\psi_\alpha(a)}$ up to logarithmic factors, while $\chi(a) \triangleq \min_{\alpha \in \mathcal{I}_*} \psi_\alpha(a)$ selects the chart and minimizing face of the feasible contraction-rate polyhedron $\mathcal{S}_\alpha(a)$, where $d_\alpha^\top s$ is minimal. Charts outside $\mathcal{I}_*$ are separated from the target and are suppressed by their positive localizer gap. Our contribution tracks how this moving constraint changes the minimizing face and chooses a schedule avoiding such changes without estimating chart exponents. Because $\chi(a)$ combines loss scale t and localization scale $\gamma = t^a$, on an affine regime $\chi(a) = c_\chi + \beta a$ the LLC-relevant quantity is the loss-scaling intercept $c_\chi = \chi(a) - a\chi'(a)$.

Proposition 1 (Piecewise affine localization regimes)**.** *Under Assumption 1, χ is piecewise affine on every compact interval of $a \geq 0$. There exists $a_0 > 0$ such that $\chi(a) = \lambda(w^*) + \beta_0 a$ for $0 < a < a_0$. On any affine regime $\chi(a) = c_\chi + \beta a$, the loss-scaling intercept satisfies $c_\chi = \chi(a) - a\chi'(a)$ at differentiability points.*

Remark. The first regime has the target pointwise LLC as its intercept. A later regime can appear stable under small finite changes in t or γ while having a different intercept. Therefore, finite-scale numerical stability alone does not identify the LLC-preserving regime. The model-specific endpoint a_0 is unnecessary for the model-agnostic schedule below.

Example 2 (Strong localization distortion)**.** For $K(u) = u^4$ and $q(u) = u^2$, LLC is $\lambda(u = 0) = 1/4$ and Eq. (2) gives $\psi(a) = \max\{1/4, a/2\}$. The loss controls the dominant contraction scale for $0 < a < 1/2$, whereas the localizer controls it for $a > 1/2$. The latter regime is stable but has loss-scaling intercept 0 rather than $1/4$.

Model-agnostic localization class. The fixed-power diagnostic shows that a model-independent numerical value of the first breakpoint is unnecessary. Outside $\mathcal{V}$, $q \geq q_0$ suppresses remote mass by $e^{-q_0\gamma_t}$, so $\gamma_t / \log t \to \infty$ dominates any unknown polynomial factor in t. Inside $\mathcal{V}$, $\log \gamma_t / \log t \to 0$ stays below every fixed positive breakpoint, as summarized in Figure 2(D).

Theorem 2 (Model-agnostic localization consistency)**.** *We assume Assumption 1, with q target-isolating on $\overline{\mathcal{V}}$, and require*

$$\lim_{t\to\infty} \gamma_t / \log t = \infty, \quad \lim_{t\to\infty} \log \gamma_t / \log t = 0. \tag{3}$$

For $c \neq 1$, we define $\lambda_t^{\mathrm{ratio}}(c) \triangleq -(\log c)^{-1} \log\left(Z(ct, \gamma_t)/Z(t, \gamma_t)\right)$. Here γ_t is held fixed in the numerator and denominator: the ratio varies the loss scale t locally at the chosen localized target rather than moving to γ_{ct} along the schedule. For every compact $C \subset (0, \infty) \setminus \{1\}$, we have $\lim_{t\to\infty} \sup_{c\in C} \left|\lambda_t^{\mathrm{ratio}}(c) - \lambda(w^*)\right| = 0$. Therefore, $\lambda_t^{\mathrm{ratio}} \to \lambda(w^*)$ locally uniformly on $(0, \infty) \setminus \{1\}$. The exclusion of $c = 1$ is necessary because the ratio representation divides by $\log c$. Consequently, $\lim_{t\to\infty} \Lambda(t, \gamma_t) = \lambda(w^*)$.

Remark. Theorem 2 gives a model-agnostic sufficient schedule class converging to the target pointwise LLC. The lower condition suppresses remote low-loss regions despite unknown polynomial advantages, while the upper condition keeps the resolved localization constraint below every fixed positive breakpoint. The result is asymptotic; finite-scale bias, empirical replacement of K, and numerical sampling error remain separate.

Corollary 3 (Equivalent model-agnostic growth class)**.** *Eq. (3) is equivalent to $\log t = o(\gamma_t)$ and $\gamma_t = o(t^a)$ for every fixed $a > 0$. In particular, $\gamma_t = (\log t)^p$ for fixed $p > 1$ is admissible.*

We denote this broad model-agnostic (MA) sufficient schedule class by $\mathcal{G}_{\mathrm{MA}}$; it does not select a unique finite-t formula.

Proposition 4 (Classwise boundary counterexamples). *For $q(w) = \|w - w^*\|_2^2/2$, classwise counterexamples exist among one-dimensional real analytic models with positive smooth ω. If $\gamma_t/\log t$ does not diverge, a remote minimum-loss region can dominate along some unbounded sequence $t_j \to \infty$. If $\limsup_{t\to\infty} \log \gamma_t/\log t > 0$, a finite-order analytic singularity can be localizer dominated along some unbounded sequence $t_j \to \infty$. In either case, $\Lambda(t, \gamma_t)$ does not necessarily converge to the target pointwise LLC.*

Remark. Proposition 4 does not state that every schedule outside $\mathcal{G}_{\mathrm{MA}}$ fails for every model. Rather, neither boundary can be removed from a guarantee uniform over the stated analytic class without model-specific constants or geometry.

Finite scale geometry and error separation. Theorem 2 gives asymptotic consistency but no universal finite-t rate because local orders are model-specific. For finite-scale intuition, we consider a leading local term K_0 that satisfies $K_0(cz) = c^{2P}K_0(z)$ under radial rescaling. Here, $P = 1$ includes regular quadratic behavior and $P > 1$ pure higher-order singular behavior. Gaussian localization at scale $\gamma^{-1/2}$ makes a degree-$2P$ term scale as γ^{-P}, giving competition ratio t/γ^P. This is only a diagnostic, not an assumption of Theorem 2 or the numerical method. General analytic or locally polynomial models, including neural-network regions with such expansions, can mix degrees or minimizing faces of Eq. (2). Appendix D treats these cases and gives exact response identities without assuming a first-order correction formula.

Let $\widehat{\lambda}$ be a numerical estimate, $\Lambda_n(t, \gamma)$ the exact empirical-target quantity, and $\Lambda(t, \gamma)$ the population quantity. Then, we have the following erro decomposition:

$$|\widehat{\lambda} - \lambda(w^*)| \le |\widehat{\lambda} - \Lambda_n(t, \gamma)| + |\Lambda_n(t, \gamma) - \Lambda(t, \gamma)| + |\Lambda(t, \gamma) - \lambda(w^*)|.$$

We denote these terms by B_{num}, B_{emp}, and B_{loc}. B_{num} covers finite-chain Monte Carlo error and approximate-Langevin discretization and stochastic-gradient bias (Vehtari et al., 2021; Welling & Teh, 2011; Teh et al., 2016; Vollmer et al., 2016). B_{emp} is the population–empirical error of the data-dependent Gibbs target, related to generalization and PAC Bayesian analyses (McAllester, 1999; Catoni, 2007; Alquier et al., 2016). B_{loc} is the population-level localization discrepancy addressed here. For finite-temperature diagnosis of B_{loc}, we use the compact-window ratio defect $\mathcal{D}_\epsilon$ and secant correction $\kappa(\epsilon)$.

Proposition 5 (Localization error bound). *For $\lambda \ge 0$ and $0 < \epsilon < 1$, we define $\mathcal{D}_\epsilon(t, \gamma; \lambda) \triangleq \sup_{c\in[1-\epsilon,1+\epsilon]} |\log [Z(ct, \gamma)/Z(t, \gamma)] + \lambda \log c|$ and $\kappa(\epsilon) \triangleq \max\{1-\log(1+\epsilon)/\epsilon, -\log(1-\epsilon)/\epsilon - 1\}$. Then, we have $|\Lambda(t, \gamma) - \lambda| \le \frac{\mathcal{D}_\epsilon(t,\gamma;\lambda)}{\epsilon} + \lambda\kappa(\epsilon)$.*

Note that the first term measures finite-window departure from the candidate power law, while $\kappa(\epsilon) = O(\epsilon)$ is the secant correction.

Limitation and future perspective. The separation condition $q(w) \ge q_0 > 0$ outside $\mathcal{V}$ holds for the standard Gaussian localizer $q(w) = \|w - w^*\|_2^2/2$ with any fixed target neighborhood (Lau et al., 2025). The theory also covers analytic localizers whose zero set isolates the target solution and that admit a common resolution with K (Hironaka, 1964; Bierstone & Milman, 1997; Watanabe, 2009). Extending the guarantee to multiple-zero, nonanalytic, or data-dependent localizers and deriving finite-scale certificates without resolved exponents remain for future work.

5 Numerical Method

In this section, we briefly describe a numerical method. At each training step, we freeze the trained parameters w^*, define the empirical excess loss $K_n(w) = L_n(w) - L_n(w^*)$, and use the Gaussian localizer $q(w) = \|w - w^*\|_2^2/2$. In all experiments below, we set $t = n/\log n$ following the WBIC scaling used in scalable LLC estimation (Watanabe, 2013; Lau et al., 2025). For a chosen localization strength γ, SGLD targets $\pi_{t,\gamma}(dw) \propto \exp(-tK_n(w) - \gamma q(w))\omega(w)\,dw$ and estimates $\widehat{\Lambda}(t, \gamma) = tM^{-1}\sum_{m=1}^{M} K_n(w_m)$ from retained draws $w_1, \ldots, w_M$. The primary DNN measurements use $\gamma_t = \log t \log\log t \in \mathcal{G}_{\mathrm{MA}}$. At a fixed checkpoint, localization comparisons share the same anchor, data bank, and t, and use matched SGLD settings when the comparison is intended to isolate localization strength. Section 6 specifies the compared schedules and controls, and Appendix E gives the reproduction settings.

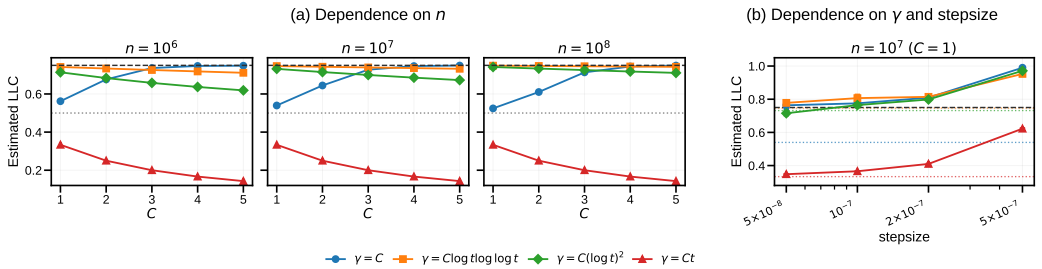


Figure 3: Known-truth validation. Left: deterministic finite-t LLC references. Right: SGLD step-size refinement toward schedule-specific references; error bars show mean $\pm$ standard error over five trials.

6 Experiments

In this section, we study three questions through experiments. **RQ1**: Do the model-agnostic valid localization recover a known target LLC while the predicted weak- and strong-localization failures appear? **RQ2**: Do LLC trajectories reveal reproducible geometric changes beyond predictive metrics? **RQ3**: Which observations do persist across localization strength, sampler randomness, and step-size refinement? Appendix E gives full settings and the retained fixed-horizon step-size checks. Code is available at `https://anonymous.4open.science/r/vloc-3C32/`.

Known-truth localization boundaries (RQ1). We use $K(w) = (w_1 + 1)^2(w_1 - 1)^4/16 + w_2^4$ with a Gaussian localizer centered at $w^* = (-1, 0)$. The target LLC is $3/4$, whereas the remote zero $(1, 0)$ has LLC $1/2$. For $n \in \{10^6, 10^7, 10^8\}$, $t = n/\log n$, and $C \in \{1, \ldots, 5\}$, Fig. 3 compares $\gamma = C$, $C \log t \log\log t$, $C(\log t)^2$, and Ct. The two polylogarithmic schedules satisfy Theorem 2, while the constant and linear controls violate its lower and upper growth conditions, respectively. Deterministic quadrature shows the admissible schedules moving toward $3/4$ as n increases, whereas the two controls approach different finite-scale regimes. At $n = 10^7$ and $C = 1$, decreasing the SGLD step size approaches the corresponding deterministic references for each schedule. Therefore, the separation between schedules is not explained by the tested discretization error.

Grokking (RQ2, RQ3). We examine the OmniGrok MNIST setting (Liu et al., 2023) with a depth-3 width-200 ReLU MLP trained on 1,000 examples. Fig. 4 (a)(b) shows delayed generalization: training accuracy saturates early while test accuracy improves later. At each checkpoint, $t = 1000/\log 1000 = 144.76$ and we compare the primary $\gamma = \log t \log\log t = 7.98$, the stronger admissible $\gamma = (\log t)^2 = 24.75$, and the polynomial control $\gamma = t = 144.76$. All use $\epsilon = 2.5 \times 10^{-6}$, 50,000 SGLD updates, and five sampler seeds on the same training trajectory.

Fig. 4 (c) shows a broad LLC elevation during delayed generalization followed by a late decline under both admissible schedules. From 20k to 100k, their cross-seed means decrease from 351.6 to 176.3 and from 244.4 to 89.6, respectively, while the polynomial control is much more strongly suppressed, from 74.2 to 9.5. The different absolute levels are compatible with the asymptotic theory because admissible schedules are not required to agree at one finite t. More importantly, the late training-aligned decline is shared by both admissible schedules. We observe that four step sizes preserve the localization ordering, and the final halving changes cross-seed means by at most 2.7% across all audited checkpoints and schedules (Appendix E.3). We therefore treat the absolute values as finite-horizon measurements and interpret the shared late decline, rather than a particular finite-t level, as the robust geometric signal.

Multi-stage learning rate (RQ2, RQ3). We train CIFAR-10 ResNet20 (He et al., 2016) with learning rates 0.1, 0.01, and 0.001 over steps 0–31,999, 32,000–47,999, and 48,000–64,000, and evaluate LLC at 32k, 36k, 48k, and 52k. We set $t = n/\log n = 1000/\log 1000 = 144.76$ and compare the weak control $\gamma = 1$, the primary $\gamma = \log t \log\log t = 7.98$, the stronger admissible $\gamma = (\log t)^2 = 24.75$, and the polynomial control $\gamma = t$. All four use the same $\epsilon = 5 \times 10^{-7}$, 50,000

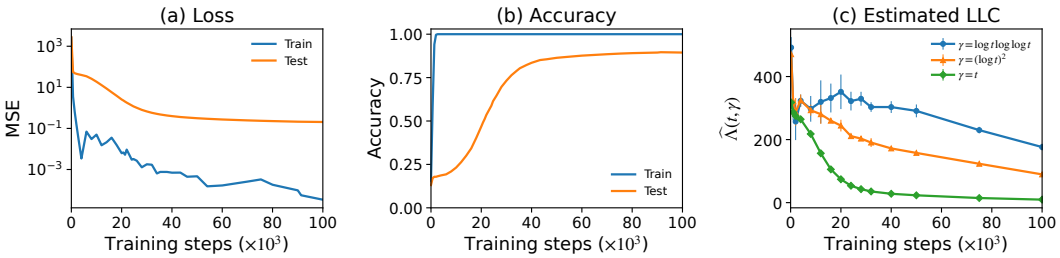


Figure 4: Grokking on MNIST. (a) MSE, (b) accuracy, and (c) estimated LLC under two admissible schedules and a polynomial control. Points in (c) are means over five sampler seeds and error bars show one standard deviation.

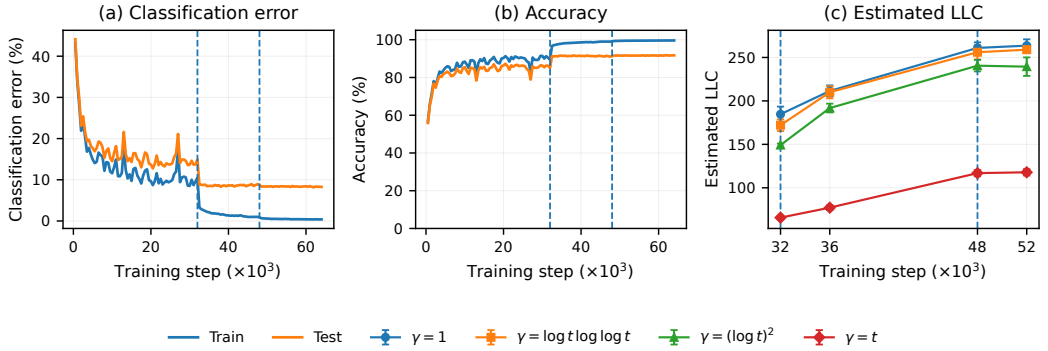


Figure 5: Multi-stage learning-rate experiment on CIFAR-10 ResNet20. (a) Classification error, (b) accuracy, and (c) estimated LLC. All four localization conditions use the same SGLD step size and horizon and reproduce the increase-to-plateau pattern, while the absolute finite-t scale depends on localization strength.

updates, and five sampler seeds, so their finite-t trajectories are directly comparable under the same numerical protocol.

Fig. 5 shows a common increase from 36k to 48k followed by a plateau across all four localization conditions, while stronger localization lowers the absolute scale. The primary estimate rises from 210.0 to 256.0, about 22%, although test error changes only from 8.47% to 8.71% over the same interval. The matched weak control and second admissible schedule show the same increase-to-plateau pattern, indicating that the trajectory is not specific to the primary choice of γ. At 36k and 48k, halving ϵ and doubling the updates at fixed $H = \eta T$, where η is stepsize and T is the number of steps, changes the primary means by only -1.4% and 2.6%; the 36k–48k contrast remains positive for all five sampler seeds before and after refinement (Appendix E.4). The LLC therefore continues to change after classification performance is already nearly stable, and the observed stagewise reorganization persists across localization choices and the tested step-size refinement.

7 Conclusion

In this study, we showed that localization strength is part of a local model complexity target estimated by sampling. Theoretically, simultaneous resolution of the loss and localizer yields a model-agnostic sufficient condition that suppresses remote low-loss regions without imposing a polynomial local scale, characterizes the limits of uniform guarantees without model-specific geometry, and separates localization, empirical, and numerical error. Empirically, controlled models verify the predicted failure regimes, while experiments on DNNs show that localization choice can affect the structural signals revealed by local model complexity. These results provide a foundation for future work on practical finite-scale certificates and extensions to broader localizers and model classes toward more reliable local model complexity estimation.

AI DISCLOSURE

In this work, we used generative AI tools to assist in the writing of proofs and to design or provide feedback on research methodology and experiments. We also used them to draft and edit parts of the manuscript, summarize and analyze existing literature, identify relevant literature, and format references. All AI-assisted mathematical arguments, methodological and experimental suggestions, manuscript text, and references were reviewed by the authors, with relevant literature and references checked against the original sources. The authors made all final decisions regarding the mathematical claims, proofs, methodology, experiments, interpretation, and presentation of the work. We take responsibility for the final content of this work, including text, claims, proofs, experimental procedures, references, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We state the assumptions and formal results for valid localization in Section 4 and provide complete proofs and supporting derivations in the appendices. Section 5 specifies the finite-target construction, localization diagnostics, and numerical qualification criteria used throughout the experiments. The experimental section and appendices report the model, data, training, localization, sampling, and evaluation settings needed to interpret and reproduce the reported results. We provide the code used for the experiments and figures, together with reproduction instructions, in an anonymous repository at `https://anonymous.4open.science/r/vloc-3C32/`.

REFERENCES

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41, 2016.

Miki Aoyagi. Learning coefficient in bayesian estimation of restricted boltzmann machine. *Journal of Algebraic Statistics*, 4(1):30–57, 2013.

Miki Aoyagi. Consideration on the learning efficiency of multiple-layered neural networks with linear units. *Neural Networks*, 172:106132, 2024.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.

Michael F. Atiyah. Resolution of singularities and division of distributions. *Communications on Pure and Applied Mathematics*, 23(2):145–150, 1970.

Garrett Baker, George Wang, Jesse Hoogland, Vinayak Pathak, and Daniel Murfet. Structural inference: Interpreting small language models with susceptibilities. In *Proceedings of the fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

Edward Bierstone and Pierre D. Milman. Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant. *Inventiones Mathematicae*, 128(2):207–302, 1997.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1877–1901, 2020.

Yingjia Cai. Bypassing minimization bias: A shift-invariant variance estimator for off-equilibrium local learning coefficients, 2026. arXiv preprint arXiv:2606.22389.

Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers, 2025. arXiv preprint arXiv:2501.17745.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007.

Zhongtian Chen and Daniel Murfet. Modes of sequence models and learning coefficients, 2025. arXiv preprint arXiv:2504.18048.

Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus bayesian phase transitions in a toy model of superposition, 2023. arXiv preprint arXiv:2310.06301.

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 *Proceedings of the ninth International Conference on Learning Representations (ICLR)*, 2021.

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. Grokking as a phase transition between competing basins: A singular learning theory approach, 2026. arXiv preprint arXiv:2603.01192.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318, 2018.

Chris Elliott and Daniel Murfet. Linear response estimators for singular statistical models, 2026. arXiv preprint arXiv:2605.07970.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2nd edition, 1999.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8803–8812, 2018.

Andrew Gordon, Garrett Baker, George Wang, William Snell, Stan van Wingerden, and Daniel Murfet. Towards spectroscopy: Susceptibility clusters in language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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 (CVPR)*, pp. 770–778, 2016.

Max Hennick and Stijn De Baerdemacker. Almost bayesian: Dynamics of SGD through singular learning theory. In *Proceedings of the fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero: I, II. *Annals of Mathematics*, 79:109–203; 205–326, 1964.

Rohan Hitchcock and Jesse Hoogland. From global to local: A scalable benchmark for local posterior sampling, 2025. arXiv preprint arXiv:2507.21449.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *Proceedings of the eighth International Conference on Learning Representations (ICLR)*, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Philipp Alexander Kreer, Wilson Wu, Maxwell Adam, Zach Furman, and Jesse Hoogland. Bayesian influence functions for hessian-free data attribution. In *Proceedings of the fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Yuki Kurumadani. Learning coefficients in semiregular models I: Properties. *Japanese Journal of Statistics and Data Science*, 8:1051–1079, 2025a.

Yuki Kurumadani. Learning coefficients in semiregular models II: Extensions. *Japanese Journal of Statistics and Data Science*, 9:249–283, 2025b.

Yuki Kurumadani. Upper bounds for local learning coefficients of three-layer neural networks. *Neural Networks*, 205:109374:1–109374:13, 2026.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 244–252, 2025.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism, 2020. arXiv preprint arXiv:2003.02218.

Derek Lim, Theo Moe Putterman, Robin Walters, Haggai Maron, and Stefanie Jegelka. The empirical impact of neural parameter symmetries, or lack thereof. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 28322–28358, 2024.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *Proceedings of the eleventh International Conference on Learning Representations (ICLR)*, 2023.

Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press, 1995.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. arXiv preprint arXiv:2201.02177.

Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Cambridge University Press, 2nd edition, 2014.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Tejas Pradeep Shirodkar. Dead directions: Geometric singular learning, 2026. arXiv preprint arXiv:2606.05957.

Tejas Pradeep Shirodkar and P. J. Narayanan. Algebraic dead directions in LayerNorm transformers: A forward-pass-only diagnostic at LLM scale, 2026a. arXiv preprint arXiv:2606.19491.

Tejas Pradeep Shirodkar and P. J. Narayanan. Dead-direction signatures: A cheap spectral reading of singular complexity, 2026b. arXiv preprint arXiv:2606.21158.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732, 2021.

Yee Whye Teh, Alexandre H. Thiery, and Sebastian J. Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

Einar Urdshals and Jasmina Urdshals. Structure development in list-sorting transformers, 2025. arXiv preprint arXiv:2501.18666.

Einar Urdshals, Edmund Lau, Jesse Hoogland, Stan van Wingerden, and Daniel Murfet. Compressibility measures complexity: Minimum description length meets singular learning theory, 2025. arXiv preprint arXiv:2510.12077.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, 16(2):667–718, 2021.

Sebastian J. Vollmer, Konstantinos C. Zygalakis, and Yee Whye Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *Proceedings of the thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.

Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10473–10486, 2023.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 681–688, 2011.

Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, 2003.

Keisuke Yamazaki and Sumio Watanabe. Algebraic geometry and stochastic complexity of hidden markov models. *Neurocomputing*, 69(1–3):62–84, 2005.

A NOTATION

In this section, we summarize the notation used in the main text and appendices. Vectors are columns, $\|\cdot\|_2$ is the Euclidean norm, and vector inequalities are coordinatewise unless stated otherwise.

Table 1: Main notation.

	Symbol	Description
Data	X, Y	Random input and output variables. Lowercase x, y denote realizations.
	D_n	Dataset $\{(x_i, y_i)\}_{i=1}^n$.
Model	w, w^*	Parameter and target learned solution.
	$\mathcal{W}, \mathcal{W}_0, \mathcal{V}$	Parameter space, zero excess loss set, and target neighborhood.
Loss	ℓ, L, L_n	Per example, population, and empirical losses.
	K, K_n	Population and empirical excess losses relative to w^*.
SLT	$\lambda(\mathcal{V})$	LLC associated with a fixed neighborhood $\mathcal{V}$.
	$\lambda(w^*)$	Pointwise LLC obtained as $\mathcal{V}$ shrinks to w^*.
	$Z_{\mathcal{A}}(t)$	Unlocalized partition function on a measurable region $\mathcal{A} \subset \mathcal{W}$.
Localization	$\omega(w)$	Fixed background density in the SLT integral, distinct from the Gaussian localizing prior ϕ_γ.
	$q(w)$	Target-dependent nonnegative localizer.
	t, γ, γ_t	Effective inverse temperature scale, localization strength, and schedule.
	$\mathcal{G}_{\mathrm{MA}}$	Model-agnostic sufficient schedule class.
	$Z_{\mathcal{A}}(t,\gamma), \pi^{\mathcal{A}}_{t,\gamma}$	Localized partition function and restricted Gibbs target on $\mathcal{A}$. Full-space quantities use $Z(t,\gamma)$ and $\pi_{t,\gamma}$.
Observable	$\Lambda(t,\gamma)$	Finite temperature local complexity $t\mathbb{E}_{\pi_{t,\gamma}}[K]$.
	$Q(t,\gamma)$	Localizer load $\gamma\mathbb{E}_{\pi_{t,\gamma}}[q]$.
	R_t, R_γ	Temperature and localization responses of Λ.
	P_{eff}	Effective homogeneity diagnostic when the exact virial identity applies.
Error	$B_{\mathrm{num}}, B_{\mathrm{emp}}, B_{\mathrm{loc}}$	Numerical error, empirical error, and population-level localization discrepancy.
Geometry	$\mathcal{M}, U_\alpha, u$	Resolved manifold, local chart, and resolved coordinates.
	$M_\xi(u)$	Monomial shorthand $\prod_j u_j^{\xi_j}$.
	$k_\alpha, r_\alpha, d_\alpha$	Loss, localizer, and transformed volume exponent vectors.
	$\mathcal{S}_\alpha(a)$	Feasible contraction-rate polyhedron for the fixed-power diagnostic.
	$\psi_\alpha(a), \chi(a)$	Chartwise localized and combined decay exponents.
Finite scale	κ_P	Ratio t/γ^P when single homogeneity applies.
	$\mathcal{D}_\epsilon, D^{(2)}_{\mathrm{end}}, D^{(2)}_{\mathrm{sup}}, D^{(4)}_{\mathrm{sup}}$	Compact window partition ratio defects.
	C_2	Symmetric dyadic non-affinity defect.

B SINGULAR LEARNING THEORY AND RESOLVED LOCAL GEOMETRY

In this section, we summarize the singular learning theory (SLT) concepts used in the proof and clarify how definition of the local learning coefficient (LLC) relates to scalable LLC estimation. As in Section 3, we take $w \in \mathcal{W} \subset \mathbb{R}^{d_w}$, define the zero excess loss set by $\mathcal{W}_0 \triangleq \{w \in \mathcal{W} \mid K(w) = 0\}$, and consider a relatively compact target neighborhood $\mathcal{V} \subset \mathcal{W}$ that contains w^*. The essential progression is from a fixed neighborhood coefficient $\lambda(\mathcal{V})$ to the pointwise quantity $\lambda(w^*)$, followed by a resolved coordinate representation in which singular orders become explicit.

Regular and singular local models. We call a model regular at a target $w^* \in \mathcal{W}$ when (i) the parameterization is locally identifiable, so nearby distinct parameters represent distinct model distributions, and (ii) the Hessian $H \in \mathbb{R}^{d_w \times d_w}$ of K at w^* is positive definite. For $h \in \mathbb{R}^{d_w}$ near zero, this gives $K(w^* + h) = h^\top H h/2 + o(\|h\|_2^2)$, so the local partition function decays as $t^{-d_w/2}$. For a well specified likelihood model, this Hessian agrees with the Fisher information $I(w^*) \triangleq \mathbb{E}[\nabla_w \log p(Y \mid X; w^*) \nabla_w \log p(Y \mid X; w^*)^\top]$ (Watanabe, 2009; 2013). A model is singular at the target when at least one of these conditions fails. Such violations can produce a zero excess loss set with positive dimension or higher order degeneracy (Watanabe, 2001; 2009; 2018). SLT replaces ordinary quadratic dimension by invariants of this local analytic geometry.

Definition B.1 (Local learning coefficient)**.** *Let $K(w) \geq 0$ for every $w \in \mathcal{V}$, where K is real analytic on the relatively compact neighborhood $\mathcal{V}$ of w^*, and let ω be positive and smooth there. Under the standard SLT conditions, we have*

$$Z_{\mathcal{V}}(t) = c(\mathcal{V}) t^{-\lambda(\mathcal{V})} (\log t)^{m(\mathcal{V})-1} (1 + o(1)), \tag{4}$$

where $c(\mathcal{V}) > 0$, $\lambda(\mathcal{V}) > 0$, and $m(\mathcal{V}) \in \mathbb{N}$. We call $\lambda(\mathcal{V})$ LLC on $\mathcal{V}$. For the Euclidean ball $B_r(w^*) \triangleq \{w \in \mathcal{W} \mid \|w - w^*\|_2 \leq r\}$, we define the pointwise LLC by $\lambda(w^*) \triangleq \lim_{r \downarrow 0} \lambda(B_r(w^*))$ when the limit is taken within the parameter space (Watanabe, 2009; Lau et al., 2025).

Remark. This is the local asymptotic coefficient used by scalable LLC studies, written explicitly as a neighborhood to point limit (Lau et al., 2025). No additional factor that involves the zero excess loss set is required in the definition. The set $\mathcal{W}_0 = \{w \in \mathcal{W} \mid K(w) = 0\}$ is nevertheless essential because different points or strata of $\mathcal{W}_0$ can have different local coefficients. Localization is necessary precisely when an integral over a larger neighborhood can include such competing regions. In a tempered posterior with inverse temperature β, the corresponding large scale parameter is $t = n\beta$. Ordinary Bayesian inference uses $\beta = 1$, whereas the widely applicable Bayesian criterion (WBIC) uses $\beta = 1/\log n$ (Watanabe, 2013).

A smaller LLC means that more weighted parameter volume remains near zero excess loss as the tolerance shrinks. The multiplicity $m(\mathcal{V})$ controls only the logarithmic correction.

Zeta and sublevel volume views. We define the local zeta function by $\zeta_{\mathcal{V}}(z) \triangleq \int_{\mathcal{V}} K(w)^z \omega(w)\, dw$ and the weighted sublevel volume by $V_{\mathcal{V}}(\epsilon) \triangleq \int_{\mathcal{V}} \mathbb{1}\{K(w) \leq \epsilon\} \omega(w)\, dw$. The largest pole of the former is $-\lambda(\mathcal{V})$, while the leading power of the latter is $\epsilon^{\lambda(\mathcal{V})}$ up to logarithmic factors (Atiyah, 1970; Watanabe, 2009). The Laplace form is most directly connected to estimation by sampling.

Before introducing the local coordinate formulas, Fig. 6 summarizes the geometric role of resolution. Fig. 6 (a) shows the singular zero loss geometry in the original parameters. Fig. 6 (b) shows the resolved manifold covered by overlapping charts, and Fig. 6 (c) shows one such chart in local coordinates. The singular set can be difficult to describe in the original parameters, but resolution replaces it locally by coordinate hyperplanes. The loss and transformed weighted mass then have explicit monomial orders in the same coordinates, which is the form used by LLC asymptotics below.

Normal crossing representation. For real analytic functions, resolution of singularities gives local coordinates in which vanishing orders become monomial (Hironaka, 1964; Atiyah, 1970; Watanabe, 2009; 2013).

Theorem B.2 (Normal crossing representation after resolution (Hironaka, 1964; Atiyah, 1970; Watanabe, 2009; 2013))**.** *Under the standard analytic SLT conditions, after a proper analytic resolution map and finite chart decomposition, each local chart $U_\alpha \subset \mathbb{R}^{d_w}$ admits analytic functions*

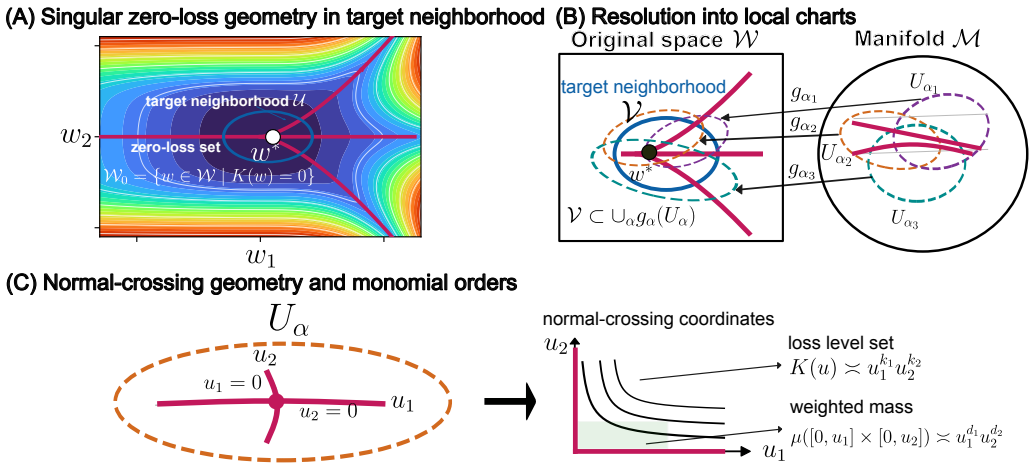


Figure 6: Picture of resolution of singularity (Hironaka, 1964). (a) Singular zero loss geometry in the original target neighborhood $\mathcal{V}$. (b) Resolution replaces this by a nonsingular resolved manifold $\mathcal{M}$ covered by overlapping local charts whose images cover $\mathcal{V}$. The oblique appearance of the red components is schematic. (c) In a chart U_α, local coordinates turn the pulled back zero loss set into coordinate hyperplanes, and the loss and transformed weighted mass take monomial form up to smooth positive units. The analytic localizer used in Section 4 is specific to our localization analysis and is therefore omitted from this background figure.

$A_\alpha, C_\alpha : U_\alpha \to \mathbb{R}_{>0}$ *such that*

$$K \circ g_\alpha(u) = A_\alpha(u) \prod_{j=1}^{d_w} u_j^{k_{\alpha,j}}, \quad \omega(g_\alpha(u))|\det Dg_\alpha(u)| = C_\alpha(u) \prod_{j=1}^{d_w} |u_j|^{d_{\alpha,j}-1},$$

where $k_{\alpha,j} \in \mathbb{N}_0$, $d_{\alpha,j} > 0$, *and* A_α *and* C_α *are bounded above and below by positive constants on compact chart supports.*

Remark. Theorem B.2 supplies the normal crossing forms for K and the weighted measure. Assumption 1 adds two ingredients. First, standard principalization refines the resolved charts so that the analytic localizer q also becomes a monomial times a positive analytic unit in the same coordinates (Hironaka, 1964; Bierstone & Milman, 1997). Second, it assumes positive separation of q outside the target neighborhood. The latter is not a consequence of resolution alone and encodes the requirement that the localizer distinguish the target neighborhood from remote regions.

This normal crossing representation is the standard SLT consequence of resolution of singularities. The vector $k_\alpha = (k_{\alpha,1}, \ldots, k_{\alpha,d_w})$ records the vanishing orders of the loss in chart α, while $d_\alpha = (d_{\alpha,1}, \ldots, d_{\alpha,d_w}) \in \mathbb{R}^{d_w}_{>0}$ records the powers contributed by the transformed density and Jacobian.

For each chart, the chartwise LLC is the power-law exponent governing how quickly its weighted low loss volume shrinks as the tolerance decreases. The standard normal crossing calculation (Atiyah, 1970; Watanabe, 2009) gives

$$\lambda_\alpha \triangleq \min_{s \in \mathbb{R}^{d_w}_{\geq 0}} \left\{ d_\alpha^\top s \mid k_\alpha^\top s \geq 1 \right\}. \tag{5}$$

The constraint $k_\alpha^\top s \geq 1$ requires enough coordinate contraction to make the loss of order t^{-1}, while $d_\alpha^\top s$ is the corresponding power-law cost in weighted volume. Therefore, λ_α is the slowest decay exponent available in chart α, that is, its chartwise contribution to LLC. The pointwise LLC is the minimum over charts whose images intersect arbitrarily small neighborhoods of w^*. Ties between contraction directions generate the logarithmic multiplicity. Section 4 adds the second localizer constraint and studies how its movement changes the minimizing face, that is, the set of contraction patterns attaining the minimum volume cost.

C PROOFS OF MODEL-AGNOSTIC LOCALIZATION

This appendix proves Theorem 2. Inside the target neighborhood $\mathcal{V}$, localization must preserve the partition-function power law $Z_{\mathcal{V}}(t) \asymp t^{-\lambda(w^*)}$ up to logarithmic factors from Eq. (4). Outside $\mathcal{V}$, it must suppress competing low-loss mass. The localized partition function $Z_{\mathcal{A}}(t, \gamma)$ in Eq. (1) contains both scales, so the proof must control them simultaneously.

The resolution and normal-crossing representation used below are standard SLT ingredients from Theorem B.2 (Atiyah, 1970; Watanabe, 2009; 2013). Linear programming (LP) duality and polyhedral sensitivity are also standard (Bertsimas & Tsitsiklis, 1997). Our localization-specific step is to resolve q with K, obtain a second moving constraint, and prove that the resulting two-scale geometry remains in the LLC-preserving region for the model-agnostic schedule of Theorem 2.

Initial localization region. For a resolved chart α from Assumption 1 and Theorem B.2, Section 4 associates the fixed-power family $\gamma = t^a$ with the feasible contraction-rate polyhedron $\mathcal{S}_\alpha(a)$ and chartwise localized decay exponent $\psi_\alpha(a)$ in Eq. (2). The corresponding primal linear program minimizes the transformed volume cost $d_\alpha^\top s$ over $\mathcal{S}_\alpha(a)$, equivalently subject to the loss constraint $k_\alpha^\top s \geq 1$ and the localization constraint $r_\alpha^\top s \geq a$. Its standard LP dual is formulated as

$$\psi_\alpha(a) = \max_{\mu,\nu \geq 0} \{\mu + a\nu \mid \mu k_{\alpha,i} + \nu r_{\alpha,i} \leq d_{\alpha,i},\ \forall i\}. \tag{6}$$

The primal problem chooses coordinate contraction rates. In the dual problem, μ and ν are the marginal values assigned to meeting the loss and localization constraints. The inequalities $\mu k_{\alpha,i} + \nu r_{\alpha,i} \leq d_{\alpha,i}$ state that their combined gain in coordinate i cannot exceed its available volume cost. Therefore, each feasible dual point defines an affine lower bound $\mu + a\nu$ on the primal minimum, and the dual optimum chooses the largest such bound (Bertsimas & Tsitsiklis, 1997). The dual feasible set in (μ, ν) is a fixed polyhedron defined by finitely many linear inequalities and is independent of a. Whenever the optimum is finite, the linear objective $\mu + a\nu$ is attained on an optimal vertex or face, and only finitely many vertices can occur. Therefore, $\psi_\alpha(a)$ is piecewise affine. Geometrically, each affine piece corresponds to a fixed active-constraint pattern and hence to the same combinatorial type of minimizing face of the primal polyhedron. A breakpoint is a value of a where this pattern can change.

For each target-relevant chart $\alpha \in \mathcal{I}_*$, let $a_{0,\alpha} \in (0, \infty]$ be its first positive breakpoint, or ∞ if none exists. The initial active-constraint pattern is fixed on $0 < a < a_{0,\alpha}$. Finiteness of $\mathcal{I}_*$ gives a common $a_0 > 0$ within all these intervals.

Proof of Proposition 1. On every compact interval $I \subset [0, \infty)$, Eq. (6) is the maximum of finitely many affine functions $\mu + a\nu$. Therefore, each ψ_α in Eq. (2) is piecewise affine in a, and the finite minimum $\chi(a) = \min_{\alpha \in \mathcal{I}_*} \psi_\alpha(a)$ is also piecewise affine. At $a = 0$, $\psi_\alpha(0) = \lambda_\alpha$. Among the dual vertices with intercept $\mu = \lambda_\alpha$, the one with the largest slope ν maximizes $\mu + a\nu$ for all sufficiently small $a > 0$. Therefore, $\psi_\alpha(a) = \lambda_\alpha + \beta_\alpha a$ on $(0, a_{0,\alpha})$.

By the definition of $\mathcal{I}_*$ in Section 4 and the pointwise LLC characterization in Appendix B, $\min_{\alpha \in \mathcal{I}_*} \lambda_\alpha = \lambda(w^*)$. Because $\mathcal{I}_*$ is finite, every $\alpha \in \mathcal{I}_*$ with $\lambda_\alpha > \lambda(w^*)$ has a fixed positive intercept gap, which an $O(a)$ slope term cannot close for all sufficiently small a. Therefore, on a nonempty common interval, $\chi(a) = \lambda(w^*) + \beta_0 a$, where $\beta_0 = \min\{\beta_\alpha : \alpha \in \mathcal{I}_*,\ \lambda_\alpha = \lambda(w^*)\}$. On any differentiable affine region $\chi(a) = c_\chi + \beta a$, we have $\chi'(a) = \beta$. Therefore, we have $c_\chi = \chi(a) - a\chi'(a)$. □

Logarithmic contour geometry. Now, we connect the smooth resolved integrand to the polyhedra. We fix a resolved chart α as in Assumption 1 and Theorem B.2, and omit its subscript only in this local calculation. After rescaling the compact chart support, we take $u_i \in (0, 1]$ and set $u_i = \exp(-s_i)$ with $s = (s_1, \ldots, s_{d_w}) \in \mathbb{R}^{d_w}_{\geq 0}$, $\ell \triangleq \log t$, and $\eta \triangleq \log \gamma$. Assumption 1 gives $K \circ g_\alpha(u) = A_\alpha(u) M_{k_\alpha}(u)$ and $q \circ g_\alpha(u) = B_\alpha(u) M_{r_\alpha}(u)$. Because $M_k(u) = \exp(-k^\top s)$ and $M_r(u) = \exp(-r^\top s)$, we obtain $tK \circ g_\alpha(u) = A_\alpha(u) \exp(\ell - k^\top s)$ and $\gamma q \circ g_\alpha(u) = B_\alpha(u) \exp(\eta - r^\top s)$. Temporarily setting the positive analytic units A_α and B_α to one gives the smooth exponential factors $\exp(-\exp(\ell - k^\top s))$ and $\exp(-\exp(\eta - r^\top s))$, while the transformed weighted measure contributes $\exp(-d^\top s)\, \mathrm{d}s$. We call $k^\top s = \ell$ and $r^\top s = \eta$ the order-one contours because they

correspond to $tK = 1$ and $\gamma q = 1$, where $\exp(-tK)$ and $\exp(-\gamma q)$ transition from order-one weight to strong exponential suppression. On the side satisfying both $k^\top s \geq \ell$ and $r^\top s \geq \eta$, neither exponential factor is strongly suppressed.

We define the corresponding threshold integral, feasible contraction polyhedron, minimum volume cost, and minimizing face by

$$H(\ell,\eta) \triangleq \int_{s\in\mathcal{S}_{\ell,\eta}} \exp(-d^\top s)\,\mathrm{d}s, \quad \mathcal{S}_{\ell,\eta} \triangleq \left\{ s \in \mathbb{R}^{d_w}_{\geq 0} \mid k^\top s \geq \ell,\ r^\top s \geq \eta \right\}, \tag{7}$$
$$\mathcal{E}(\ell,\eta) \triangleq \min_{s\in\mathcal{S}_{\ell,\eta}} d^\top s, \quad \mathcal{F}_{\ell,\eta} \triangleq \left\{ s \in \mathcal{S}_{\ell,\eta} \mid d^\top s = \mathcal{E}(\ell,\eta) \right\}.$$

Therefore, $\mathcal{S}_{\ell,\eta}$ collects contraction rates for which both exponential penalties remain unsuppressed at leading order. Among them, $\mathcal{F}_{\ell,\eta} = \arg\min_{s\in\mathcal{S}_{\ell,\eta}} d^\top s$ is the set of contraction patterns that lose the least transformed weighted volume. It is the minimizing face of the polyhedron and determines the dominant contribution because $\exp(-d^\top s)$ is largest there. The threshold integral H is not yet the actual smooth resolved-chart contribution to the localized partition function; it exposes the same leading geometry in a form accessible to linear optimization.

The scaling relation follows directly from the linear constraints. For every $\tau > 0$, $\mathcal{S}_{\tau\ell,\tau\eta} = \tau\mathcal{S}_{\ell,\eta}$, so $\mathcal{E}(\tau\ell,\tau\eta) = \tau\mathcal{E}(\ell,\eta)$. Taking $\tau = 1/\ell$ gives $\mathcal{E}(\ell,\eta) = \ell\mathcal{E}(1,\eta/\ell) = \ell\psi_\alpha(\eta/\ell)$, where ψ_α is the chartwise localized decay exponent from Eq. (2), equivalently the minimum transformed volume cost on $\mathcal{S}_\alpha(a)$. Let $a = \eta/\ell$. Within an affine interval of $\psi_\alpha(a)$, the same active-constraint pattern in Eq. (7) and the same dimension of $\mathcal{F}_{\ell,\eta}$ persist. In the initial affine interval, Proposition 1 gives $\psi_\alpha(a) = \lambda_\alpha + \beta_\alpha a$. Therefore, $\mathcal{E}(\ell,\eta) = \ell\psi_\alpha(\eta/\ell) = \lambda_\alpha\ell + \beta_\alpha\eta$, so there are constants $\xi, \zeta \geq 0$ such that

$$\mathcal{E}(\ell,\eta) = \xi\ell + \zeta\eta, \quad 0 < \eta < a_{0,\alpha}\ell, \tag{8}$$

where $\xi = \lambda_\alpha$ is the loss-scaling intercept and $\zeta = \beta_\alpha$ is the slope contributed by the localization threshold in Proposition 1.

The smooth exponential factors $\exp(-tK)$ and $\exp(-\gamma q)$ do not depend only on the single levels $tK = 1$ and $\gamma q = 1$; as Eq. (12) will show, they average nearby contour levels. In logarithmic coordinates these levels appear as bounded shifts $(\ell + h_1, \eta + h_2)$ of the thresholds in Eq. (7). Before returning to the actual smooth resolved-chart integral, we therefore need to show that these bounded shifts remain in the same affine regime and preserve the same active-constraint pattern. Intuitively, the distance to a breakpoint grows with ℓ and η, whereas the shifts being averaged are only $O(1)$.

For a convex set S, $\mathrm{aff}(S)$ denotes its affine hull, the smallest affine subspace containing S, and $\mathrm{relint}(S) \triangleq \{x \in S \mid \exists r > 0,\ B_r(x) \cap \mathrm{aff}(S) \subset S\}$ denotes its relative interior. Intuitively, $\mathrm{relint}(S)$ consists of points that are interior when movement is restricted to the affine subspace containing S. An inequality that defines $\mathcal{S}_{\ell,\eta}$ is active on a face when it holds with equality there. Its slack is the positive distance, in the corresponding linear functional, from equality. For a positive-dimensional convex set S, we define its relative inradius by

$$\rho(S) \triangleq \sup\{r > 0 \mid \exists x \in \mathrm{relint}(S),\ B_r(x) \cap \mathrm{aff}(S) \subset S\}.$$

This is the largest ball that fits inside S when distance is measured within its own affine hull (Schneider, 2014). It measures how much room remains inside the face before a perturbation reaches its relative boundary. We define $f_{\ell,\eta} \triangleq \dim(\mathrm{aff}(\mathcal{F}_{\ell,\eta}))$ and, when $f_{\ell,\eta} > 0$, $\rho_{\ell,\eta} \triangleq \rho(\mathcal{F}_{\ell,\eta})$.

Lemma C.1 (Interior geometry of the LLC-preserving region)**.** *We consider a sequence $\ell \to \infty$ with $\eta = \eta(\ell) \to \infty$ and $\eta/\ell \to 0$. For every fixed $D < \infty$, all $(\ell + h_1, \eta + h_2)$ with $|h_1|, |h_2| \leq D$ eventually remain in the same affine region. If $f_{\ell,\eta} > 0$, $\rho_{\ell,\eta} \geq c\min\{\ell,\eta\}$ holds for some $c > 0$. If $f_{\ell,\eta} = 0$, every inactive constraint has slack at least $c\min\{\ell,\eta\}$.*

Proof. Within a fixed affine region, the vertices of $\mathcal{S}_{\ell,\eta}$, the slacks of inactive constraints, and the distances to the finitely many boundaries where the active face changes are affine and positively homogeneous in (ℓ,η). Each relevant positive margin therefore has the form $A\ell + B\eta$ for fixed real coefficients A, B. Positivity throughout the open initial cone implies either $A > 0$, in which case $\eta/\ell \to 0$ gives $A\ell + B\eta \geq (A/2)\ell$ eventually, or $A = 0$ and then $B > 0$, in which case the margin equals $B\eta$. Hence every such margin is bounded below by $c\min\{\ell,\eta\}$ after taking the minimum

over finitely many constraints and boundaries. The same finite-dimensional affine description gives the relative inradius bound when the minimum-cost set has positive dimension. Therefore, every bounded shift (h_1, h_2) eventually remains within the same affine region. □

The next issue is that $\mathcal{F}_{\ell,\eta}$ can be a vertex, a line segment, or a higher-dimensional polytope. Its ordinary d_w-dimensional volume is zero whenever $f_{\ell,\eta} < d_w$, so the asymptotic prefactor must measure its size within its own affine hull. For a set U, we define $\operatorname{diam}(U) \triangleq \sup_{x,y\in U} \|x - y\|_2$. For a Hausdorff-measurable set $S \subset \mathbb{R}^{d_w}$, the f-dimensional Hausdorff measure is defined by

$$\mathcal{H}^f(S) \triangleq c_f \liminf_{\delta\to 0} \left\{ \sum_i (\operatorname{diam} U_i)^f \mid S \subset \bigcup_i U_i,\ \operatorname{diam} U_i \le \delta \right\}.$$

The normalization c_f is chosen so that $\mathcal{H}^f$ equals ordinary Euclidean f-volume on an f-dimensional affine plane (Mattila, 1995). Therefore, it reduces to length, area, or volume when $f = 1, 2, 3$. We use it below for polyhedral minimizing faces and write $\operatorname{Vol}_f(S) \triangleq \mathcal{H}^f(S)$.

The LP determines the minimum exponential cost $\mathcal{E}$, but H also accumulates mass along every point of $\mathcal{F}_{\ell,\eta}$, where $d^\top s = \mathcal{E}(\ell, \eta)$ has the same minimum value. If the minimizer is a line segment or higher-dimensional polytope rather than a single point, integrating along it contributes a factor proportional to its length, area, or higher-dimensional volume. The next lemma separates these effects: $\exp(-\mathcal{E}(\ell, \eta))$ gives the dominant exponential decay, while $\operatorname{Vol}_f(\mathcal{F}_{\ell,\eta})$ measures the f-dimensional size of the minimizing face within its own affine hull. This is a standard local face-and-cone decomposition for polyhedra (Schneider, 2014). Here, we apply it uniformly to the moving thresholds $k^\top s \ge \ell$ and $r^\top s \ge \eta$ in Eq. (7).

Lemma C.2 (Minimum-set tube asymptotic)**.** *Under Lemma C.1, if $f_{\ell,\eta} > 0$, there exists $C_F > 0$, independent of (ℓ, η) in the affine region, such that*

$$H(\ell, \eta) = \exp(-\mathcal{E}(\ell, \eta)) C_F \operatorname{Vol}_{f_{\ell,\eta}}(\mathcal{F}_{\ell,\eta}) \left[1 + O\left(\frac{\log \rho_{\ell,\eta}}{\rho_{\ell,\eta}} \right) \right]. \tag{9}$$

If $f_{\ell,\eta} = 0$, there exists $C_v > 0$ such that $H(\ell, \eta) = \exp(-\mathcal{E}(\ell, \eta)) C_v (1 + o(1))$. Both statements are uniform under bounded shifts of (ℓ, η) because Lemma C.1 gives margins of order $\min\{\ell, \eta\}$, whereas bounded shifts are only $O(1)$.

Proof. First, we take $f_{\ell,\eta} > 0$ and abbreviate $\mathcal{F} = \mathcal{F}_{\ell,\eta}$ and $\mathcal{S} = \mathcal{S}_{\ell,\eta}$. Let $L_F \triangleq \operatorname{span}(\mathcal{F} - \mathcal{F})$ and choose $s_F \in \operatorname{relint}(\mathcal{F})$. The space L_F consists of directions parallel to the affine hull of the minimizing face. The feasible tangent cone $T_{\mathcal{S}}(s_F) \triangleq \overline{\{a(s - s_F) \mid a \ge 0,\ s \in \mathcal{S}\}}$ collects directions in which one can move infinitesimally from s_F while remaining feasible. We write $L_F^\perp \triangleq \{z \in \mathbb{R}^{d_w} \mid z^\top v = 0 \text{ for all } v \in L_F\}$ for directions orthogonal to the face and define the transverse cone by $\mathcal{C}_F \triangleq T_{\mathcal{S}}(s_F) \cap L_F^\perp$. It removes directions along the face and retains feasible displacements $z_\perp$ that move away from it, which are the directions whose decay must be integrated. Because $\mathcal{F} = \arg\min_{s\in\mathcal{S}} d^\top s$ is the full set of minimizers, every nonzero $z_\perp \in \mathcal{C}_F$ satisfies $d^\top z_\perp > 0$; otherwise the minimizer set would extend transversely beyond $\mathcal{F}$. The set $\mathcal{C}_F \cap \{z : \|z\|_2 = 1\}$ is compact, and the continuous function $d^\top z$ is strictly positive there, so it has a positive minimum c_F. Therefore, $d^\top z_\perp \ge c_F \|z_\perp\|_2$ and $C_F \triangleq \int_{z_\perp \in \mathcal{C}_F} \exp(-d^\top z_\perp)\, \mathrm{d}z_\perp$ is finite and positive. It is the exponentially decaying transverse mass per unit f-dimensional volume along the minimizing face.

We define the relative boundary by $\partial_{\mathrm{rel}}\mathcal{F} \triangleq \mathcal{F} \setminus \operatorname{relint}(\mathcal{F})$; it is the endpoints, edges, or higher-dimensional boundary of $\mathcal{F}$ viewed within its own affine hull. For $x \in \operatorname{aff}(\mathcal{F})$, define $\operatorname{dist}_{\operatorname{aff}(\mathcal{F})}(x, \partial_{\mathrm{rel}}\mathcal{F}) \triangleq \inf_{y\in\partial_{\mathrm{rel}}\mathcal{F}} \|x - y\|_2$. The R-interior core $\mathcal{F}^{\mathrm{core}}(R) \triangleq \{x \in \mathcal{F} \mid \operatorname{dist}_{\operatorname{aff}(\mathcal{F})}(x, \partial_{\mathrm{rel}}\mathcal{F}) \ge R\}$ is the part of the minimizing face left after removing an intrinsic boundary layer of width R. By the definition of $\rho_{\ell,\eta}$, there is a center $x_{\ell,\eta} \in \operatorname{relint}(\mathcal{F})$ such that an intrinsic ball $B_{\rho_{\ell,\eta}}(x_{\ell,\eta}) \cap \operatorname{aff}(\mathcal{F})$ lies in $\mathcal{F}$. The homothetic contraction $x \mapsto x_{\ell,\eta} + (1 - R/\rho_{\ell,\eta})(x - x_{\ell,\eta})$ moves the face away from its relative boundary by at least R, so its image lies in $\mathcal{F}^{\mathrm{core}}(R)$. Its f-dimensional volume is scaled by $(1 - R/\rho_{\ell,\eta})^f$, and $1 - (1 - x)^f \le fx$ gives $\operatorname{Vol}_f(\mathcal{F} \setminus \mathcal{F}^{\mathrm{core}}(R)) / \operatorname{Vol}_f(\mathcal{F}) \le fR/\rho_{\ell,\eta}$ (Schneider, 2014). Over this core, no additional polyhedral inequality becomes active within transverse distance R, so the feasible set has the local product geometry of the face and $\mathcal{C}_F$ up to that

distance. The omitted transverse cone tail is $O(\exp(-c_F R))$. Choosing $R = c_F^{-1} \log \rho_{\ell,\eta}$ balances the two errors: the cone tail becomes $O(\rho_{\ell,\eta}^{-1})$ and the boundary-layer error is $O(\log \rho_{\ell,\eta}/\rho_{\ell,\eta})$, giving Eq. (9). If $f_{\ell,\eta} = 0$, Lemma C.1 gives an expanding neighborhood of the minimizing vertex with a fixed tangent cone, and the complementary cone tail is exponentially small. Bounded threshold shifts preserve the same affine region, so the bounds are uniform. □

Lemma C.2 shows that bounded threshold shifts change the dominant exponential cost according to Eq. (8), while the intrinsic face-volume ratio approaches one. The next lemma records the resulting ratio. This is exactly the stability needed before the smooth averaging step.

Lemma C.3 (Hard polyhedral ratio)**.** *We consider $\ell \to \infty$ with $\eta = \eta(\ell) \to \infty$ and $\eta(\ell)/\ell \to 0$. For every fixed $D < \infty$, we have*

$$\lim_{\ell \to \infty} \sup_{|h_1|,|h_2| \le D} \left| \frac{H(\ell + h_1, \eta + h_2)}{H(\ell, \eta)} - \exp(-\xi h_1 - \zeta h_2) \right| = 0. \tag{10}$$

Proof. Eq. (8) gives $\mathcal{E}(\ell + h_1, \eta + h_2) - \mathcal{E}(\ell, \eta) = \xi h_1 + \zeta h_2$. If the minimum-cost set is a vertex, Lemma C.2 immediately gives the stated ratio. Otherwise, bounded threshold shifts move $\mathcal{F}_{\ell,\eta}$ by bounded Hausdorff distance, while Lemma C.1 gives $\rho_{\ell,\eta} \to \infty$. Standard inner and outer parallel-body bounds for fixed-dimensional polytopes then give $\mathrm{Vol}_f(\mathcal{F}_{\ell+h_1,\eta+h_2})/\mathrm{Vol}_f(\mathcal{F}_{\ell,\eta}) \to 1$ uniformly for bounded (h_1, h_2) (Schneider, 2014). Substituting this ratio and the cost difference into Eq. (9) yields Eq. (10). □

Exact transfer to the smooth chart integral. The threshold geometry above identifies the minimizing geometry and shows that nearby thresholds remain in the same affine regime, but the actual resolved-chart contribution to $Z(t, \gamma)$ still contains the smooth exponential factors $\exp(-tK)$ and $\exp(-\gamma q)$. We now return exactly to that smooth chart integral rather than treating the threshold problem as an approximation. After temporarily setting the bounded analytic units to one, we define its monomial form by

$$J(\ell, \eta) \triangleq \int_{s \in \mathbb{R}^{d_w}_{\ge 0}} \exp\left(-\exp(\ell - k^\top s) - \exp(\eta - r^\top s) - d^\top s\right) \mathrm{d}s. \tag{11}$$

Eq. (11) is the exact monomial resolved-chart partition mass in logarithmic coordinates: the first double-exponential factor is the loss suppression, the second is the localization suppression, and $\exp(-d^\top s)$ is the transformed weighted-volume factor. We define $g(z) \triangleq \exp(z - \exp(z))$. Because $\int_{z \in [x,\infty)} g(z)\,\mathrm{d}z = \int_{y \in [\exp(x),\infty)} \exp(-y)\,\mathrm{d}y = \exp(-\exp(x))$ under $y = \exp z$, each smooth factor is an exact weighted average of nearby threshold levels. Applying the identity to both factors and using Tonelli's theorem (Folland, 1999) gives

$$J(\ell, \eta) = \int_{(z_1,z_2) \in \mathbb{R}^2} g(z_1) g(z_2) H(\ell - z_1, \eta - z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2. \tag{12}$$

Therefore, the shifts studied above are not arbitrary perturbations. They are precisely the nearby contour levels averaged by the original smooth exponential factors. To pass the hard ratio limit through this exact average, we need a uniform integrable bound.

Lemma C.4 (Threshold relaxation domination)**.** *There exist constants $C_k, C_r > 0$ such that, for all real z_1, z_2 and admissible (ℓ, η), we have*

$$\frac{H(\ell - z_1, \eta - z_2)}{H(\ell, \eta)} \le \exp(C_k (z_1)_+ + C_r (z_2)_+). \tag{13}$$

Proof. For the resolved chart under consideration, the exponent vectors $k, r \in \mathbb{R}^{d_w}_{\ge 0}$ are nonzero, so we can choose $p_k, p_r \in \mathbb{R}^{d_w}_{\ge 0}$ with $k^\top p_k \ge 1$ and $r^\top p_r \ge 1$. For $z_1, z_2 \ge 0$, translation by $y = z_1 p_k + z_2 p_r$ maps the relaxed feasible set $\mathcal{S}_{\ell - z_1, \eta - z_2}$ in Eq. (7) into $\mathcal{S}_{\ell,\eta}$. Therefore, $H(\ell, \eta) \ge \exp(-d^\top y) H(\ell - z_1, \eta - z_2)$. We define $C_k \triangleq d^\top p_k$ and $C_r \triangleq d^\top p_r$. These constants depend only on the fixed exponent vectors, not on ℓ or η. A negative z_i makes the corresponding threshold stricter and cannot increase the integral. Eq. (13) follows. □

Lemma C.4 supplies the domination required to average Lemma C.3. Therefore, the next lemma transfers the same leading ratio from the stable contour geometry to the exact smooth monomial integral.

Lemma C.5 (Smooth monomial ratio)**.** *Inside the initial affine region with $\eta \to \infty$ and $\eta/\ell \to 0$, we have*

$$\lim_{\ell\to\infty} \frac{J(\ell,\eta)}{H(\ell,\eta)} = \Gamma(\xi+1)\Gamma(\zeta+1), \tag{14}$$

where, for $a > 0$, $\Gamma(a) \triangleq \int_{x\in(0,\infty)} x^{a-1}\exp(-x)\,\mathrm{d}x$ is the Gamma function. For every fixed $D < \infty$, we obtain

$$\lim_{\ell\to\infty} \sup_{|h|\le D} \left| \frac{J(\ell+h,\eta)}{J(\ell,\eta)} - \exp(-\xi h) \right| = 0. \tag{15}$$

Proof. For fixed (z_1, z_2), Lemma C.3 gives $H(\ell - z_1, \eta - z_2)/H(\ell,\eta) \to \exp(\xi z_1 + \zeta z_2)$. Lemma C.4 gives an integrable dominating function after multiplication by $g(z_1)g(z_2)$, because g decays exponentially as $z \to -\infty$ and superexponentially as $z \to \infty$. Dominated convergence in Eq. (12) then applies. For $c > -1$, $\int_{z\in\mathbb{R}} \exp(cz)g(z)\,\mathrm{d}z = \int_{z\in\mathbb{R}} \exp((c+1)z - \exp(z))\,\mathrm{d}z = \int_{x\in(0,\infty)} x^c \exp(-x)\,\mathrm{d}x = \Gamma(c+1)$ under $x = \exp(z)$. This gives Eq. (14). The same dominated-convergence argument after the bounded shift $\ell \mapsto \ell + h$ gives Eq. (15), where the Gamma factors cancel. □

Analytic units and finite chart aggregation. The preceding lemmas treat the monomial factors exactly. Assumption 1 also contains positive analytic units $A_\alpha, B_\alpha, C_\alpha : U_\alpha \to \mathbb{R}_{>0}$ bounded above and below on compact chart supports. A standard partition of unity handles overlapping chart images and yields finitely many compactly supported chart contributions. On sufficiently small supports, uniform continuity and positivity give, for every $\varepsilon > 0$, constants $A_0, B_0, C_0 > 0$ such that $A_0\exp(-\varepsilon) \le A_\alpha(u) \le A_0\exp(\varepsilon)$, $B_0\exp(-\varepsilon) \le B_\alpha(u) \le B_0\exp(\varepsilon)$, and $C_0(1-\varepsilon) \le C_\alpha(u) \le C_0(1+\varepsilon)$ (Watanabe, 2009). The first two bounds shift ℓ and η by $\log A_0 + O(\varepsilon)$ and $\log B_0 + O(\varepsilon)$, while the C_α bound changes only a multiplicative prefactor. Thus Lemma C.5 transfers to each analytic chart. Partition weights are nonnegative, bounded, and may vanish at support boundaries, so only boundedness is used for upper bounds. On a dominant resolved stratum, some weight is positive and hence uniformly positive on a smaller neighborhood, preserving one leading lower-bound contribution. Other vanishing weights only suppress contributions. Finiteness of the cover then aggregates the chartwise ratios to the neighborhood ratio.

These lemmas provide the chartwise asymptotic control needed to prove Theorem 2. Eq. (7) identifies the dominant contour geometry, Lemma C.3 keeps its active-constraint pattern stable across the nearby levels required by Eq. (12), and Lemma C.5 returns to the exact smooth chart integral. It remains to combine the charts contributing the target pointwise LLC, remove mass outside $\mathcal{V}$, and pass from the partition ratio to the energy observable.

Proof of Theorem 2. First, we control the contribution inside $\mathcal{V}$. For each target-relevant chart $\alpha \in \mathcal{I}_*$, let λ_α be the chartwise loss exponent in Eq. (5). The minimum defining the pointwise LLC is taken over precisely these charts; no constraint involving q enters this loss-only definition. We denote this minimum by λ_*. By Theorem B.2 and Eq. (5), shrinking the covered neighborhood to w^* gives $\lambda_* = \lambda(w^*)$ (Atiyah, 1970; Watanabe, 2009). The localizer q affects the localization slope within each chart but not this LLC intercept.

For a target-relevant chart with $\lambda_\alpha > \lambda_*$, Eq. (8) and Proposition 1 give the exponential cost $\lambda_\alpha\ell + \beta_\alpha\eta$, while face-volume and logarithmic-multiplicity factors contribute only $O(\log\ell)$ to its logarithm. Therefore, the initial-region exponent has the form $\lambda_\alpha\ell + \beta_\alpha\eta + O(\log\ell)$. The same argument gives $\lambda_*\ell + \beta_{\alpha_*}\eta + O(\log\ell)$ for a chart $\alpha_* \in \mathcal{I}_*$ with $\lambda_{\alpha_*} = \lambda_*$. Therefore, their ratio is of order $\exp(-(\lambda_\alpha - \lambda_*)\ell + O(\eta + \log\ell))$. The upper condition in Eq. (3) gives $\eta/\ell \to 0$, so the fixed positive intercept gap dominates and every nonminimal target-relevant chart is negligible. All target-relevant charts attaining λ_* have the same loss-scaling exponent, even when their initial localization slopes differ.

For $\alpha \in \mathcal{I} \setminus \mathcal{I}_*$, target isolation and compact chart support give a chartwise constant $q_\alpha > 0$ with $q \geq q_\alpha$ on the support. Since there are finitely many such charts, their total contribution is suppressed by $\exp(-q_{\text{off}}\gamma_t)$ for some $q_{\text{off}} > 0$ and is negligible relative to the target-chart polynomial mass under $\gamma_t/\log t \to \infty$. Therefore, for every compact interval $C = [c_-, c_+] \subset (0, \infty)$, we have

$$\lim_{t\to\infty} \sup_{c\in C} \left| \log \frac{Z_{\mathcal{V}}(ct, \gamma_t)}{Z_{\mathcal{V}}(t, \gamma_t)} + \lambda(w^*) \log c \right| = 0. \tag{16}$$

This is locally uniform convergence of the logarithmic partition ratio inside the target neighborhood. Here, $Z_{\mathcal{V}}(t, \gamma) = \int_{\mathcal{V}} \exp\left(-tK(w) - \gamma q(w)\right)\omega(w)\,\mathrm{d}w$ is the localized partition function restricted to $\mathcal{V}$, as defined in Eq. (1).

Next, we control the complement. Let $\bar{\gamma}$ be the fixed constant in Assumption 1. Since $\gamma_t \to \infty$, eventually $\gamma_t \geq \bar{\gamma}$. Because $q(w) \geq q_0$ for $w \notin \mathcal{V}$ and $K(w) \geq 0$ on $\mathcal{W}$, we then have

$$Z_{\mathcal{V}^c}(t, \gamma_t) \leq \exp\left(-q_0(\gamma_t - \bar{\gamma})\right) \int_{\mathcal{W}} \exp(-\bar{\gamma} q(w))\omega(w)\,\mathrm{d}w. \tag{17}$$

For a target-relevant chart attaining λ_*, Lemma C.2 and the analytic-unit comparison above give fixed constants $C_0, C_1, C_2, C_3 > 0$, independent of ℓ and η, such that

$$Z_{\mathcal{V}}(t, \gamma_t) \geq C_0 \exp\left(-C_1\ell - C_2\eta\right)(1 + \ell)^{-C_3}. \tag{18}$$

Eq. (8) poses the exponential cost. Lemma C.2 poses the intrinsic face-volume factor, which is at most polynomial in (ℓ, η) because the vertices vary affinely within the fixed initial region. The bounded units $A_\alpha, B_\alpha, C_\alpha$ only change fixed constants and bounded threshold shifts. Replacing $\ell = \log t$ and $\eta = \log \gamma_t$ and combining Eqs. (17) and (18) gives fixed constants $C, p, q > 0$, independent of t, γ_t, and c on any fixed compact interval, such that

$$\frac{Z_{\mathcal{V}^c}(ct, \gamma_t)}{Z_{\mathcal{V}}(ct, \gamma_t)} \leq C t^p \gamma_t^q \exp\left(-q_0\gamma_t\right). \tag{19}$$

The lower condition $\gamma_t/\log t \to \infty$ makes the final exponential dominate every fixed polynomial prefactor. Therefore, the outside-to-inside mass ratio in Eq. (19) converges to zero locally uniformly in c.

For $c > 0$, we define $R_t(c) \triangleq Z_{\mathcal{V}^c}(ct, \gamma_t)/Z_{\mathcal{V}}(ct, \gamma_t)$. Because $Z = Z_{\mathcal{V}} + Z_{\mathcal{V}^c}$, we have

$$\log \frac{Z(ct, \gamma_t)}{Z(t, \gamma_t)} = \log \frac{Z_{\mathcal{V}}(ct, \gamma_t)}{Z_{\mathcal{V}}(t, \gamma_t)} + \log[1 + R_t(c)] - \log[1 + R_t(1)]. \tag{20}$$

For every compact interval $C \subset (0, \infty)$, Eqs. (19) and (20), together with $0 \leq \log(1 + x) \leq x$, give

$$\sup_{c\in C} \left| \log \frac{Z(ct, \gamma_t)}{Z(t, \gamma_t)} + \lambda(w^*) \log c \right| \leq \sup_{c\in C} \left| \log \frac{Z_{\mathcal{V}}(ct, \gamma_t)}{Z_{\mathcal{V}}(t, \gamma_t)} + \lambda(w^*) \log c \right| + \sup_{c\in C} R_t(c) + R_t(1).$$

The first term tends to zero by Eq. (16). The second tends to zero uniformly by Eq. (19), and the same bound at $c = 1$ gives $R_t(1) \to 0$. Therefore, we have

$$\lim_{t\to\infty} \sup_{c\in C} \left| \log \frac{Z(ct, \gamma_t)}{Z(t, \gamma_t)} + \lambda(w^*) \log c \right| = 0. \tag{21}$$

This proves locally uniform convergence of the logarithmic partition ratio on $(0, \infty)$. If $C \subset (0, \infty) \setminus \{1\}$, then $|\log c|$ is bounded away from zero on C, and dividing Eq. (21) by $-\log c$ gives $\lim_{t\to\infty} \sup_{c\in C} |\lambda_t^{\text{ration}}(c) - \lambda(w^*)| = 0$.

It remains to pass from the ratio limit to $\Lambda(t, \gamma_t)$. We define $g_t(c) \triangleq \log Z(ct, \gamma_t) - \log Z(t, \gamma_t)$. Direct differentiation gives $g_t'(c) = -t\mathbb{E}_{\pi_{ct,\gamma_t}}[K]$ and $g_t''(c) = t^2 \operatorname{Var}_{\pi_{ct,\gamma_t}}(K) \geq 0$, so g_t is convex. Eq. (21) gives $g_t(c) \to -\lambda(w^*) \log c$ locally uniformly. For $0 < h < 1$, convexity sandwiches the derivative at $c = 1$ between the two secant slopes as

$$\frac{g_t(1) - g_t(1 - h)}{h} \leq g_t'(1) \leq \frac{g_t(1 + h) - g_t(1)}{h}.$$

We first take $t \to \infty$ using the locally uniform limit in both secants and then let $h \downarrow 0$. Both bounds converge to $-\lambda(w^*)$. Because $g_t'(1) = -\Lambda(t, \gamma_t)$, the squeeze gives $\Lambda(t, \gamma_t) \to \lambda(w^*)$. □

Corollary 3 rewrites Eq. (3) as growth faster than $\log t$ but slower than t^a for every fixed $a > 0$. The word fixed is essential. The statement does not require a bound uniform over t-dependent exponents $a = a_t$.

Proof of Corollary 3. The first condition in Eq. (3) is exactly $\log t = o(\gamma_t)$. If $\log \gamma_t / \log t \to 0$, for every fixed $a > 0$, we eventually have $\log \gamma_t < (a/2) \log t$, so $\gamma_t / t^a < t^{-a/2} \to 0$. Conversely, if $\gamma_t = o(t^a)$ for every fixed $a > 0$, then $\limsup_{t\to\infty} \log \gamma_t / \log t \le a$ for every $a > 0$. Because the lower condition implies $\gamma_t \to \infty$, the ratio is eventually nonnegative. Letting $a \to 0$ forces both limsup and liminf to zero. For $\gamma_t = (\log t)^p$ with $p > 1$, both conditions follow directly. □

Boundary counterexamples. Theorem 2 gives a sufficient schedule class that is uniform over the stated analytic models. Proposition 4 addresses whether either asymptotic condition can be removed while retaining such a uniform guarantee. The next two lemmas answer this by explicit analytic constructions. They are representative counterexamples, not claims that every schedule outside the sufficient class fails for every model.

Lemma C.6 (Remote-minimum lower-boundary counterexample)**.** *We suppose that $\gamma_t / \log t$ does not diverge. Then, there exist $m \in \mathbb{N}$, $\delta > 0$, and a one-dimensional real analytic model with $K_{m,\delta}(x) \triangleq x^{2m}(x-\delta)^2$, $q_\delta(x) \triangleq \frac{(x-\delta)^2}{2}$, such that the target is $x^* = \delta$ with pointwise LLC $\lambda(x = \delta) = 1/2$, while the remote zero at $x = 0$ dominates the localized partition function along an unbounded sequence $t_j \to \infty$. Consequently, $\Lambda(t, \gamma_t)$ does not necessarily converge to the target pointwise LLC.*

Proof. Because $\gamma_t / \log t$ does not diverge, there exist a constant $C_\gamma < \infty$ and an unbounded sequence $t_j \to \infty$ such that $\gamma_{t_j} \le C_\gamma \log t_j$. Let $Z_{\mathrm{tar}}(t, \gamma)$ and $Z_{\mathrm{rem}}(t, \gamma)$ denote contributions from fixed disjoint small neighborhoods of δ and 0, respectively. Near the target, we write $x = \delta + y$. Then, we have $K_{m,\delta}(\delta + y) = \delta^{2m} y^2 [1 + o(1)]$ and $q_\delta(\delta + y) = y^2/2$. Therefore, standard one-dimensional quadratic Laplace scaling gives $Z_{\mathrm{tar}}(t, \gamma) \asymp t^{-1/2}$ when $\gamma = O(\log t)$ (Atiyah, 1970; Watanabe, 2009).

Near the remote zero, $K_{m,\delta}(x) = \delta^2 x^{2m}[1 + o(1)]$ and $q_\delta(x) = \delta^2/2 + o(1)$ on the $t^{-1/(2m)}$ loss scale. Because $\gamma = O(\log t)$, the variation of γq_δ across that scale is $o(1)$. Monomial Laplace scaling therefore gives $Z_{\mathrm{rem}}(t, \gamma) \asymp t^{-1/(2m)} \exp(-\delta^2 \gamma / 2)$. Therefore, we have

$$\frac{Z_{\mathrm{rem}}(t, \gamma)}{Z_{\mathrm{tar}}(t, \gamma)} \asymp t^{1/2 - 1/(2m)} \exp(-\delta^2 \gamma / 2). \tag{22}$$

Along t_j, the bound $\gamma_{t_j} \le C_\gamma \log t_j$ gives $\exp(-\delta^2 \gamma_{t_j}/2) \ge t_j^{-C_\gamma \delta^2/2}$. Therefore, the ratio is bounded below in power-law order by $t_j^{1/2 - 1/(2m) - C_\gamma \delta^2/2}$. We choose m large and then δ small so that this exponent is positive. Then, the remote contribution dominates along the stated unbounded sequence $t_j \to \infty$. Moreover, standard monomial Laplace scaling in the remote neighborhood gives $t\,\mathbb{E}[K_{m,\delta} \mid \mathrm{rem}] \to 1/(2m)$, whereas the target neighborhood gives $t\,\mathbb{E}[K_{m,\delta} \mid \mathrm{tar}] \to 1/2$. The remaining region is exponentially small in t because $K_{m,\delta}$ is bounded away from zero there. Hence $Z_{\mathrm{rem}} / Z_{\mathrm{tar}} \to \infty$ implies $\Lambda(t_j, \gamma_{t_j}) \to 1/(2m) \neq 1/2$, which proves failure of convergence to the target pointwise LLC. □

The lower-bound construction shows why the first condition in Eq. (3) is necessary for a uniform guarantee. To test the upper condition, we use a single target zero so that failure can only come from localization overwhelming the loss scale.

Lemma C.7 (Localizer-dominated upper-boundary counterexample)**.** *We suppose that $\limsup_{t\to\infty} \log \gamma_t / \log t > 0$. Then, there exists $m \in \mathbb{N}$ such that the one-dimensional real analytic model $K_m(x) \triangleq x^{2m}$, $q(x) \triangleq \frac{x^2}{2}$ has pointwise LLC $\lambda(x = 0) = 1/(2m)$ but satisfies $t\mathbb{E}[K_m] \to 0$ along an unbounded sequence under the localized target.*

Proof. There exist $\epsilon > 0$ and an unbounded sequence $t_j \to \infty$ such that $\gamma_{t_j} \ge t_j^\epsilon$. Under $x = \gamma^{-1/2} y$, the localized exponent becomes $t x^{2m} + \gamma x^2/2 = \kappa_m(t, \gamma) y^{2m} + y^2/2$, where $\kappa_m(t, \gamma) \triangleq \frac{t}{\gamma^m}$. We choose an integer $m > 1/\epsilon$. Then, $\kappa_m(t_j, \gamma_{t_j}) \le t_j^{1 - m\epsilon} \to 0$. The unnormalized density satisfies

$\exp(-\kappa_m y^{2m} - y^2/2) \leq \exp(-y^2/2)$, and the corresponding moment integrand is bounded by $|y|^{2m}\exp(-y^2/2)$; both dominating functions are integrable. Therefore, the dominated convergence theorem (Folland, 1999) gives convergence of the normalized y-density to the Gaussian density proportional to $\exp(-y^2/2)$ and convergence of $\mathbb{E}[y^{2m}]$ to the corresponding Gaussian moment. Consequently, $t\mathbb{E}[K_m] = t\gamma^{-m}\mathbb{E}[y^{2m}] = \kappa_m(t,\gamma)\mathbb{E}[y^{2m}] \to 0$ along t_j. This differs from the target pointwise LLC $1/(2m)$. □

Lemmas C.6 and C.7 match the two conditions in Proposition 4. The first gives an unbounded sequence $t_j \to \infty$ along which remote mass dominates when the lower-growth condition fails. The second gives an unbounded sequence along which the localizer controls the contraction scale when the upper subpolynomial condition fails.

Proof of Proposition 4. If $\gamma_t/\log t$ does not diverge, Lemma C.6 and Eq. (22) give an analytic model whose remote contribution dominates along an unbounded sequence $t_j \to \infty$, so the localized observable does not converge to the target pointwise LLC. If $\limsup_{t\to\infty} \log\gamma_t/\log t > 0$, Lemma C.7 gives $\kappa_m(t_j, \gamma_{t_j}) \to 0$ when we set $\kappa_m(t,\gamma) \triangleq \frac{t}{\gamma^m}$, and the localized observable converges to 0 rather than $1/(2m)$. These constructions prove the two existential statements in Proposition 4. □

D FINITE SCALE PHASE GEOMETRY AND ERROR DECOMPOSITION

In this Appendix, we organize the finite-t consequences of the valid localization theory. Theorem 2 identifies an asymptotically valid schedule class, but it does not imply that all admissible schedules have the same finite bias or that a stable response certifies LLC. The relevant finite geometry is model-dependent. Therefore, we use exact identities rather than assume a universal expansion. The results below serve three purposes. They expose how loss and localization compete at finite scale, derive exact finite-scale response quantities, and separate local geometry error from remote region contamination. The decomposition here concerns localization error itself, separating local resolved geometry from mass outside the target neighborhood. The results apply broadly to analytic targets, with stronger exact formulas available for homogeneous or polynomial local models.

Free energy profile. The partition function $Z(t, \gamma)$ already combines the loss scale t and localization scale γ. Taking its negative logarithm converts multiplicative power laws into additive slopes, so derivatives of this quantity separate the response to temperature from the response to localization. This is why free energy is useful here. We define $F(\ell, \eta) \triangleq -\log Z(e^\ell, e^\eta)$, where $\ell = \log t$ and $\eta = \log \gamma$. The fixed power family in Section 4 follows the line $\eta = a\ell$, which is equivalent to $\gamma = t^a$. We define the pulled back free energy by $\widetilde{F}(\ell, a) \triangleq F(\ell, a\ell)$. In other words, $\widetilde{F}$ restricts the two scale surface $F(\ell, \eta)$ to one fixed power path. Changing ℓ then moves along that path, whereas changing a at fixed ℓ changes the localization strength relative to the temperature.

The exact response identities follow from the chain rule. First, we have

$$\frac{\partial F}{\partial \ell}(\ell, \eta) = -\frac{1}{Z(e^\ell, e^\eta)} \frac{\partial Z(e^\ell, e^\eta)}{\partial \ell} = e^\ell \mathbb{E}_{\pi_{e^\ell, e^\eta}}[K] = \Lambda(e^\ell, e^\eta),$$
$$\frac{\partial F}{\partial \eta}(\ell, \eta) = -\frac{1}{Z(e^\ell, e^\eta)} \frac{\partial Z(e^\ell, e^\eta)}{\partial \eta} = e^\eta \mathbb{E}_{\pi_{e^\ell, e^\eta}}[q] = Q(e^\ell, e^\eta).$$

Because $\widetilde{F}(\ell, a) = F(\ell, a\ell)$, another application of the chain rule gives

$$\frac{\partial \widetilde{F}}{\partial \ell} = \frac{\partial F}{\partial \ell}(\ell, a\ell) + a \frac{\partial F}{\partial \eta}(\ell, a\ell) = \Lambda(e^\ell, e^{a\ell}) + aQ(e^\ell, e^{a\ell}), \tag{23}$$

$$\frac{1}{\ell} \frac{\partial \widetilde{F}}{\partial a} = \frac{\partial F}{\partial \eta}(\ell, a\ell) = Q(e^\ell, e^{a\ell}). \tag{24}$$

Proposition 1 shows that the resolved leading exponent is affine on intervals of a. We suppose that one such interval contains the chosen a in its interior and write $\widetilde{F}(\ell, a) = \chi(a)\ell + R(\ell, a)$, where $R(\ell, a) = o(\ell)$. To differentiate this asymptotic representation, we additionally require $\partial R/\partial \ell = o(1)$ and $\ell^{-1} \partial R/\partial a = o(1)$ locally uniformly in a on a compact subinterval of the same affine region. Under this derivative control, Eqs. (23)–(24) give $Q \to \chi'(a)$ and $\Lambda \to \chi(a) - a\chi'(a)$. This gives the asymptotic interpretation of remaining away from a breakpoint of the resolved geometry described in Section 4 and Appendix C. Therefore, LLC is the intercept of the initial affine regime, not the slope or the total free energy exponent. At finite ℓ, the derivative identities remain exact even when the asymptotic approximation is not yet accurate.

Single homogeneity scaling. To make the finite-scale loss–localizer competition explicit, we first consider the special case in which a nonnegative local loss term $K_0 : \mathbb{R}^{d_s} \to \mathbb{R}_{\geq 0}$ obeys $K_0(cz) = c^{2P} K_0(z)$ for $c > 0$ and $P > 0$. This radial scaling defines degree $2P$ homogeneity and includes ordinary quadratic behavior at $P = 1$ and pure higher-order terms at $P > 1$. Let $z \in \mathbb{R}^{d_s}$ denote the local perturbation coordinates and let $q_0(z) = z^\top G z/2$ with $G \succ 0$. The next proposition shows that the finite-scale competition then collapses to one dimensionless ratio.

Proposition D.1 (Homogeneous finite scale coordinate)**.** *Under these conditions, with perturbation domain $\mathbb{R}^{d_s}$ and a positive constant background weight with respect to Lebesgue measure, there are $C_G > 0$ and $\Phi : (0, \infty) \to (0, \infty)$ such that*

$$Z_0(t, \gamma) = C_G \gamma^{-d_s/2} \Phi\left(\frac{t}{\gamma^P}\right). \tag{25}$$

Therefore, we define the homogeneous loss and localizer competition coordinate by $\kappa_P \triangleq t/\gamma^P$.

Proof. With $z = \gamma^{-1/2}y$, homogeneity gives $K_0(z) = \gamma^{-P}K_0(y)$ and $\gamma q_0(z) = y^\top Gy/2$. The Jacobian contributes $\gamma^{-d_s/2}$, so the remaining integral is proportional to $\Phi(\kappa) \triangleq \int_{y\in\mathbb{R}^{d_s}} \exp\left(-\kappa K_0(y) - y^\top Gy/2\right) \mathrm{d}y$ with $\kappa = t/\gamma^P$. Because $K_0 \geq 0$ and $G \succ 0$, this integral is finite for $\kappa > 0$. Therefore, the dependence on (t, γ) is only through $\kappa_P = t/\gamma^P$ apart from the prefactor $\gamma^{-d_s/2}$, which gives Eq. (25) after absorbing the constant background weight into C_G. □

Remark. Large κ_P is loss dominated and small κ_P is localizer dominated. Proposition D.1 is a finite-scale special case rather than an assumption of the model-agnostic theorem. The model-agnostic localization schedule does not require P, and a general analytic model can contain several degrees and therefore several competition coordinates.

Mixed homogeneous local models. A local loss can contain several homogeneous degrees. For the exact identities in this paragraph, we assume the perturbation domain is $\mathbb{R}^{d_w}$, the background weight with respect to Lebesgue measure is constant, and the boundary terms below vanish. We further assume an exact finite decomposition $K(u) = \sum_{j=j_{\min}}^{j_{\max}} K_j(u)$ with $K_j(cu) = c^j K_j(u)$, where $w = w^* + u$, together with the quadratic localizer $q(u) = \|u\|_2^2/2$. The substitution $u = \gamma^{-1/2}v$ gives

$$Z(t,\gamma) = \gamma^{-d_w/2}\Psi\left(\{\kappa_j\}_{j=j_{\min}}^{j_{\max}}\right), \quad \kappa_j \triangleq t\gamma^{-j/2},$$

where Ψ denotes the remaining dimensionless integral after the Gaussian rescaling. Its arguments κ_j record the relative strength of each homogeneous loss component against the quadratic localizer. Therefore, several finite scale competition coordinates can be active simultaneously. Individual K_j can be signed even when their sum is nonnegative.

We define the homogeneous component energies and localizer load by $\Lambda_j \triangleq t\mathbb{E}[K_j]$ and $Q \triangleq \gamma\mathbb{E}[q]$. Under the same whole-space, constant-background, and boundary assumptions, Gaussian integration by parts gives

$$\frac{d_w}{2} = Q + \frac{1}{2}\sum_{j=j_{\min}}^{j_{\max}} j\Lambda_j. \tag{26}$$

We define the total loss contribution by $\Lambda \triangleq \sum_j \Lambda_j$. When $\Lambda \neq 0$, we define the effective radial homogeneity diagnostics by

$$r_{\text{eff}} \triangleq \frac{d_w - 2Q}{\Lambda}, \quad P_{\text{eff}} \triangleq \frac{r_{\text{eff}}}{2}.$$

For a pure degree-$2P$ loss, $P_{\text{eff}} = P$. For a mixed polynomial it is only a finite-scale radial descriptor, because signed homogeneous components prevent interpretation as a positive weighted average. Eq. (26) is exact only under the whole-space, constant-background, and boundary assumptions stated above. Moreover, P_{eff} becomes ill-conditioned as $\Lambda \to 0$ and is therefore meaningful only away from $\Lambda = 0$.

General resolved models. The homogeneous formulas are special cases of the resolved chart geometry in Section 4. A general analytic or polynomial model can have several candidate resolved geometries and therefore several breakpoints as the relative scale of t and γ changes. The model-agnostic theorem avoids estimating these transition locations. At finite scales, the exact response identities below characterize sensitivity to t and γ without assuming a universal finite-t rate.

Exact response identities. For an integrable observable f with no explicit γ dependence, differentiation under the localized Gibbs density gives $\partial\mathbb{E}[f]/\partial\log\gamma = -\gamma\,\mathrm{Cov}(f, q)$. Therefore, we define the exact response quantities by

$$R_\gamma(t,\gamma) \triangleq \frac{\partial\Lambda}{\partial\log\gamma} = -t\gamma\,\mathrm{Cov}(K, q), \quad R_t(t,\gamma) \triangleq \frac{\partial\Lambda}{\partial\log t} = \Lambda - t^2\,\mathrm{Var}(K),$$
$$C_q(t,\gamma) \triangleq \frac{\partial Q}{\partial\log\gamma} = Q - \gamma^2\,\mathrm{Var}(q).$$

These identities quantify finite-scale sensitivity of the localized target. They do not establish that the target is already in the pointwise LLC regime.

Table 2: Finite-scale response quantities associated with the localized target.

Quantity	Exact meaning	Interpretation
$\Lambda = t\mathbb{E}[K]$	Local complexity energy	Finite-t localized complexity observable.
$Q = \gamma\mathbb{E}[q]$	Localizer load	Amount of radial confinement supplied by the localizer.
R_γ	$\partial\Lambda/\partial\log\gamma$	Localization sensitivity.
R_t	$\partial\Lambda/\partial\log t$	Temperature sensitivity at fixed γ.
C_q	$\partial Q/\partial\log\gamma$	Sensitivity of the localizer load.
P_{eff}	Virial-derived quantity when Eq. (26) applies	Effective radial homogeneity under the stated homogeneous-model assumptions.
$D^{(r)}_{\sup}$	Compact ratio defect	Departure from a reference power law over a multiplicative window.

Compact window ratio defects. For any $r > 1$, we define the compact window ratio defect by

$$D^{(r)}_{\sup}(t,\gamma;\lambda) \triangleq \sup_{c\in[r^{-1},r]} \left|\log\frac{Z(ct,\gamma)}{Z(t,\gamma)} + \lambda\log c\right|.$$

For the dyadic endpoints, we define $\mathfrak{F}(t,\gamma) \triangleq -\log Z(t,\gamma)$, $a_+ \triangleq [\mathfrak{F}(2t,\gamma) - \mathfrak{F}(t,\gamma)]/\log 2$, $a_- \triangleq [\mathfrak{F}(t,\gamma) - \mathfrak{F}(t/2,\gamma)]/\log 2$, $\bar{a} \triangleq (a_+ + a_-)/2$, and $C_2 \triangleq \mathfrak{F}(2t,\gamma) + \mathfrak{F}(t/2,\gamma) - 2\mathfrak{F}(t,\gamma)$. This notation distinguishes the free energy in the original variables (t,γ) from $F(\ell,\eta)$ above. We define the endpoint defect $D^{(2)}_{\text{end}}$ by the maximum of the absolute ratio errors at $c = 1/2$ and $c = 2$.

Proposition D.2 (Dyadic endpoint defect decomposition)**.** *For any reference exponent λ, we obtain*

$$D^{(2)}_{\text{end}}(t,\gamma;\lambda) = (\log 2)|\bar{a} - \lambda| + \frac{|C_2|}{2}. \tag{27}$$

Moreover, $D^{(2)}_{\text{end}} \le D^{(2)}_{\sup}$.

Proof. The two endpoint errors are $(\log 2)|a_+ - \lambda|$ and $(\log 2)|a_- - \lambda|$. The identity $\max\{|x+y|, |x-y|\} = |x| + |y|$, with $x = \bar{a} - \lambda$ and $y = (a_+ - a_-)/2$, gives Eq. (27) because $C_2 = (\log 2)(a_+ - a_-)$. The inequality follows because the endpoint set is contained in $[1/2, 2]$. □

The first term is a reference-exponent mismatch and the second measures endpoint non-affinity. A nearly affine endpoint response can therefore still disagree with the reference exponent. The full compact window defect $D^{(2)}_{\sup}$ is stronger because it also controls intermediate ratios.

Proof of Proposition 5. We define g by $g(c) \triangleq \log Z(ct,\gamma) - \log Z(t,\gamma)$. We have $g'(1) = -\Lambda(t,\gamma)$ and $g''(c) = t^2 \operatorname{Var}_{\pi_{ct,\gamma}}(K) \ge 0$. Convexity gives

$$\frac{g(1) - g(1-\epsilon)}{\epsilon} \le g'(1) \le \frac{g(1+\epsilon) - g(1)}{\epsilon}. \tag{28}$$

By the definitions of $\mathcal{D}_\epsilon$ and $\kappa(\epsilon)$ in Section 4, each secant differs by at most $\mathcal{D}_\epsilon/\epsilon$ from the corresponding secant of $-\lambda\log c$. The quantity $\kappa(\epsilon)$ is the deterministic discrepancy between the derivative $-\lambda$ at $c = 1$ and the forward or backward ϵ-secant of $-\lambda\log c$. Combining these two discrepancies with Eq. (28) gives the stated localization-error bound. □

Local and remote contributions. The compact ratio defect can fail for two conceptually different reasons. The finite-scale ratio $a = \log\gamma/\log t$ may still lie near a breakpoint where the active resolved constraints change, or a remote low loss region may still carry nonnegligible mass. We write $Z = Z_{\mathcal{V}} + Z_{\mathcal{V}^c}$ and $R(t,\gamma) \triangleq Z_{\mathcal{V}^c}(t,\gamma)/Z_{\mathcal{V}}(t,\gamma)$. The identity $\log Z = \log Z_{\mathcal{V}} + \log(1+R)$ and $0 \le \log(1+R) \le R$ give

$$\mathcal{D}_\epsilon(t,\gamma;\lambda) \le \mathcal{D}_{\epsilon,\mathcal{V}}(t,\gamma;\lambda) + 2\sup_{c\in[1-\epsilon,1+\epsilon]} R(ct,\gamma).$$

Appendix C gives $R(ct, \gamma) \leq Ct^p \gamma^q e^{-q_0 \gamma}$ for finite geometry dependent constants. The lower growth condition in Theorem 2 removes this remote term asymptotically without estimating p or q. The remaining local defect depends on the distance to breakpoints where the active resolved constraints change and has no universal finite t rate. This is the main finite scale limitation of Theorem 2 and the reason that admissible schedules can differ substantially before reaching their common asymptotic LLC.

E EXPERIMENTAL SETTINGS AND STEP-SIZE CHECKS

This appendix gives the reproduction-critical settings for Section 6 and the fixed-horizon step-size checks used to separate localization effects from SGLD discretization.

E.1 COMMON DNN MEASUREMENT PROTOCOL

At each checkpoint, w^* is frozen and SGLD targets $K_n(w) = L_n(w) - L_n(w^*)$ with $q(w) = \|w - w^*\|_2^2/2$. Comparisons use the same anchor, data bank, and t, with $\widehat{\Lambda} = tM^{-1}\sum_{m=1}^{M} K_n(w_m)$. Variability is across SGLD seeds conditional on the same trained network.

E.2 KNOWN-TRUTH LOCALIZATION BOUNDARIES

We use $K(w) = (w_1+1)^2(w_1-1)^4/16 + w_2^4$ and $q(w) = \|w-(-1,0)\|_2^2/2$, whose target and remote LLCs are $3/4$ and $1/2$. For $n \in \{10^6, 10^7, 10^8\}$, deterministic quadrature evaluates $C \in \{1,\dots,5\}$ for $\gamma = C$, $C\log t\log\log t$, $C(\log t)^2$, and Ct, with $t = n/\log n$. At $n = 10^7$ and $C = 1$, five SGLD trials at the step sizes in Fig. 3 approach the schedule-specific deterministic references.

E.3 GROKKING

We follow the MNIST OmniGrok setting (Liu et al., 2023): a depth-3 width-200 ReLU MLP with 199,210 parameters is trained on the first 1,000 examples for 100,000 AdamW steps using learning rate 10^{-3}, weight decay 0.01, mini-batch size 200, and initialization scale $\alpha = 8$. LLC is evaluated at 200, 1k, 2k, 4k, 8k, 12k, 16k, 20k, 24k, 28k, 32k, 40k, 50k, 75k, and 100k with $t = 144.76$ and $\gamma \in \{7.98, 24.75, 144.76\}$. Each condition uses five SGLD seeds, $\epsilon = 2.5\times10^{-6}$, 50,000 updates, and $H = 0.125$.

At 20k, 50k, and 100k, we repeat all schedules at fixed $H = 0.125$ with $\epsilon \in \{5, 2.5, 1.25, 0.625\}\times 10^{-6}$. Table 3 summarizes the final refinement; the localization ordering is unchanged at every tested step size.

Table 3: Size audit on the Grokking experiment. Entries are relative changes under the final halving $\epsilon : 1.25\times10^{-6} \to 0.625\times10^{-6}$. The schedule ordering is unchanged at all four tested step sizes.

Step	$\log t\log\log t$	$(\log t)^2$	t
20k	0.88%	2.52%	0.98%
50k	2.68%	1.01%	0.86%
100k	0.25%	2.31%	1.77%

E.4 MULTI-STAGE LEARNING-RATE EXPERIMENT

We train CIFAR-10 ResNet20 (He et al., 2016) with learning rates 0.1, 0.01, and 0.001 over steps 0–31,999, 32,000–47,999, and 48,000–64,000. LLC is measured at 32k, 36k, 48k, and 52k on a fixed bank of 1,000 training examples with $t = 144.76$, $\gamma \in \{1, 7.98, 24.75, 144.76\}$, mini-batch size 200, and five SGLD seeds. All four conditions use $\epsilon = 5\times10^{-7}$ and 50,000 updates.

At 36k and 48k, the primary schedule and polynomial control are repeated with $\epsilon = 2.5\times10^{-7}$ and 100,000 updates. Table 4 shows that the pointwise changes are small while the stage contrast remains positive for every sampler seed.

Table 4: Fixed-$H = 0.025$ step-size refinement from $(\epsilon, M) = (5\times10^{-7}, 50\text{k})$ to $(2.5\times10^{-7}, 100\text{k})$. C is the symmetric 36k–48k LLC contrast.

γ	Δ_{36k}	Δ_{48k}	C: base → refined	$C > 0$
$\log t\log\log t$	$-1.43 \pm 4.37\%$	$+2.63 \pm 2.45\%$	$19.75 \to 23.77\%$	$5/5 \to 5/5$
t	$+1.40 \pm 2.30\%$	$-1.56 \pm 2.56\%$	$40.96 \to 38.11\%$	$5/5 \to 5/5$