WHAT MEAN AGGREGATION RETAINS IN GNNS: DEGREE, PRECISION, AND LEARNED READOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean aggregation in GNNs is non-injective, but this does not imply that it discards all information about neighborhood size. We characterize when degree remains recoverable from the aggregate and ask whether learned GNNs exploit that information. For categorical messages, the exact mean consists of empirical proportions (counts divided by degree), whose denominators constrain, and can sometimes identify, neighborhood size. Rounding or noise can remove this degree signal while preserving information about neighborhood composition. We use Fisher-information analysis to show how degree information depends on the message distribution: for Gaussian messages, the retained fraction vanishes at high degree even as information about the message mean remains. A broad class of regular readouts recovers asymptotically none of the degree information retained by exact categorical means; fitted neural readouts exhibit the same gap, and reducing numerical precision barely changes trained-GNN predictions. Supplying degree explicitly improves log loss and robustness under a low-degree shift but provides little benefit at high degree. Together, these results show that mean aggregation can retain degree information without learned GNNs making effective use of it.

1 INTRODUCTION

Message-passing graph neural networks repeatedly compress a variable-size neighborhood into a fixed-dimensional representation. Mean aggregation is a common choice because it is simple and stable across degrees, but it removes direct access to neighborhood size. Degree can carry structural information and indicate how much evidence supports the mean: the same value computed from two neighbors and from two hundred neighbors has different statistical reliability. This loss of explicit cardinality has motivated count-aware aggregation mechanisms (Xu et al., 2019; Zhang & Xie, 2020; Corso et al., 2020; Yu & Ha, 2026).

Repeating every incoming message leaves the mean unchanged, making the aggregation map many-to-one over multisets. Yet non-injectivity does not quantify the statistical information that remains. If a node has at most eight binary messages, an exact mean of $3/5$ identifies five neighbors through its reduced denominator. Rounding to one decimal place maps both $3/5$ and $4/7$ to 0.6. Exact arithmetic can therefore encode degree even when the aggregation map is non-injective, while limited precision can remove that information.

In this work, we study when mean aggregation preserves information about neighborhood size and when a GNN can use it. We distinguish realized degree N, the actual number of neighbors in a particular sampled neighborhood, from expected degree λ, the population-level parameter governing how neighborhood sizes vary across samples. We ask three questions: how much information about λ survives aggregation, how much of that information can be extracted from the aggregate by a learned readout function, and when explicit degree improves prediction. These questions separate three stages: the aggregate determines what information is available, the readout determines what is accessible, and the task determines its predictive value.

We take the complete neighborhood, including its count and individual messages, as the reference observation. We use Fisher information to compare how strongly the complete neighborhood and the mean respond to a small change in λ; their ratio measures the fraction of degree information preserved. To measure accessibility, we train auxiliary readouts to recover degree information from

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.

the mean. The amount recovered by each readout lower-bounds the information retained by the aggregate. We repeat this analysis for neighbor-category proportions and continuous-message centers. Finally, trained-GNN experiments test whether explicit degree affects predictions under distribution shift.

Our main contributions are:

- **Degree information in exact categorical means.** We prove that, with two or more one-hot categories, exact proportions asymptotically retain all local information about expected degree and composition. Some realized degrees remain ambiguous, but the information lost to this ambiguity vanishes as expected degree grows (Thm. 3).
- **Selective loss under numerical perturbations.** We prove that rounding or additive noise can erase expected-degree information encoded in categorical means while preserving information about neighborhood composition (Thm. 5; Cor. 10).
- **A gap between retained and recovered information.** We show that retaining degree information does not imply that a learned readout can recover it. At high expected degree, we prove that a broad class of regular readouts recovers an asymptotically vanishing fraction of the information retained by exact categorical means (Thm. 4). At finite expected degree, fitted neural readouts recover only a small fraction of the information retained by exact proportions (Fig. 3).
- **Dependence on the message distribution.** We show that, for Gaussian messages, degree retention vanishes at high expected degree while location information survives, and extend the result to fixed-depth Poisson trees (Thms. 6 and 7). In trained continuous-message GNNs, explicit degree helps under a low-degree shift, while reduced precision barely affects predictions.

Our central finding is that mean aggregation does not simply "discard degree." What it preserves depends on the message distribution and numerical representation, and a learned GNN may use far less information than the aggregate contains. Exact categorical means can preserve nearly all local degree information through their degree-dependent grid of possible proportions. Rounding or noise can blur this grid and erase the degree signal while preserving composition information. Regular readouts may also recover little of the degree information retained by the exact mean. For Gaussian messages, the retained fraction of degree information vanishes at high expected degree. Distinguishing retention, accessibility, and predictive use gives a more precise account of what mean aggregation loses and when explicit degree offers a more accessible count signal.

2 Related Work

Cardinality in neighborhood aggregation. Relevant works include Xu et al. (2019, Section 5.3), cardinality-preserving attention (CPA) (Zhang & Xie, 2020), Principal Neighbourhood Aggregation (PNA) (Corso et al., 2020), and Mass-Aware Attention (Yu & Ha, 2026). Xu et al. (2019) characterize which multisets mean aggregation maps to the same output. The others modify the aggregator so that cardinality survives explicitly; Yu & Ha (2026) also separates information accessibility from task utility. All treat non-injectivity as sufficient grounds for replacing the mean. None measures how much degree information a normalized aggregate retains under a graph-generating law.

Information-theoretic analysis of GNNs. Graph Information Bottleneck (Wu et al., 2020) and Fisher Information Embedding (Chen et al., 2023) use information criteria to *learn* representations, while oversquashing (Banerjee et al., 2022), message-passing optimality (Baranwal et al., 2023; Baranwal, 2026), and Fisher information flow (Weimar et al., 2025) analyze how signal degrades with depth. Graph shift benchmarks measure transfer empirically (Gui et al., 2022; Li et al., 2024). We instead hold the aggregation map fixed and measure what it retains, which separates statistical retention from decoder recovery.

Sample size from proportions. Lukaszewicz & Dennis (2024) estimate sample size from the variability of observed proportions around model proportions. We instead use their exact rational structure, in which the reduced denominator constrains the total directly. No prior work identifies when a perturbation erases that arithmetic signal while leaving composition intact.

3 PRELIMINARIES AND SETUP

Mean aggregation in GNNs. Let $\mathcal{N}(v)$ be the neighbors of node v and let $N_v = |\mathcal{N}(v)|$ be its degree. For $N_v > 0$, layer ℓ of a message-passing GNN computes

$$A_v^{(\ell)} = \frac{1}{N_v} \sum_{u \in \mathcal{N}(v)} M_{u \to v}^{(\ell)}, \qquad h_v^{(\ell+1)} = f_{\boldsymbol{\theta}}^{(\ell)}\big(h_v^{(\ell)}, A_v^{(\ell)}\big)\,, \tag{1}$$

where $h_v^{(\ell)}$ is the representation of node v at layer ℓ, $M_{u \to v}^{(\ell)} \in \mathbb{R}^{d_\ell}$ is the vector message from u to v, and $\boldsymbol{\theta}$ collects the learned GNN parameters. We ask what the mean $A_v^{(\ell)}$ reveals about degree and the distribution of these messages.

Probabilistic neighborhood model. We first analyze the aggregation at one node and one layer. We write N for the node degree and $M_1, \ldots, M_N \in \mathbb{R}^d$ for its incoming messages. The complete observation before aggregation is $O = (N, M_1, \ldots, M_N)$. For the one-layer analysis, we consider $N \sim \mathrm{Pois}(\lambda)$ and, conditional on N, draw the messages $M_1, \ldots, M_N$ independently from Q_ω. We set $\psi = (\lambda, \omega)$, where λ is expected degree and Q_ω is a distribution on $\mathbb{R}^d$. We write P_ψ for the resulting joint distribution of O. This marked Poisson model separates expected degree from the distribution of the messages. Sec. 5 studies categorical messages encoded as one-hot vectors and Sec. 6 studies Gaussian messages. Later sections consider two additional models: fixed-depth Poisson trees for repeated aggregation and finite contextual stochastic block models for the GNN experiments.

Score and Fisher information. For an observation $X \sim P_\psi^X$ with probability mass function or density p_ψ^X, the score and Fisher information are

$$s_\psi(X) \coloneqq \nabla_\psi \log p_\psi^X(X), \qquad F_X \coloneqq \mathrm{Cov}_\psi\big(s_\psi(X)\big)\,. \tag{2}$$

The score records how the likelihood of an observation changes under a small change in ψ. Fisher information measures the average squared sensitivity. For a direction v in parameter space, we write $F_X(v, v) = v^\top F_X v$. For symmetric matrices A and B of the same dimension, $A \preceq B$ means that $B - A$ is positive semidefinite, equivalently that $v^\top A v \le v^\top B v$ for every vector v. In this paper, F_O describes the information before aggregation, while F_H will describe the information in an aggregation output H.

Total variation. For distributions P and Q, their total variation distance is $\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)|$. It quantifies how well one observation can distinguish P from Q: with equal prior probabilities, the smallest testing error is $(1 - \mathrm{TV}(P, Q))/2$. We use this connection to study nearby expected degrees after perturbing an aggregate.

4 INFORMATION RETAINED BY AGGREGATION

This section asks how much statistical evidence about a neighborhood-model parameter survives aggregation. We treat aggregation as a fixed channel from the complete neighborhood O to an output H and compare their Fisher information. For expected degree λ, this asks whether nearby expected degrees remain as distinguishable from H as from O. Information is lost when neighborhoods mapped to the same output provide different evidence about ψ. We derive an exact expression for this loss, define a direction-specific retention ratio, and obtain lower bounds from discretizations and score probes. Secs. 5 and 6 apply the framework to categorical and Gaussian messages.

4.1 FIXED AGGREGATION CHANNEL

Let H be the output available after applying an aggregation channel to O,

$$H = T_{\boldsymbol{\theta}}(O, U)\,. \tag{3}$$

Here U represents any parameter-independent channel randomization, including the theoretical perturbations below, and is independent of O. Deterministic aggregation and rounding simply omit this variable. If T contains learned transformations, $\boldsymbol{\theta}$ denotes their GNN parameters. During each

comparison, T, $\boldsymbol{\theta}$, and the distribution of U remain fixed as ψ changes. Across the paper, H denotes the output specified in each result: an exact, rounded, or noisy categorical proportion; a Gaussian sample mean; a recursive mean at the root of a fixed-depth tree; or a layerwise mean or frozen GNN state in the finite-graph diagnostics.

The question we address is how well H distinguishes a small change in ψ, compared with the complete observation O. The following proposition characterizes the loss exactly: the score available from H is the conditional mean of the complete-neighborhood score, and the missing Fisher information is its expected conditional covariance (the proof is in App. A).

***Proposition* 1.** *Under differentiability in quadratic mean and a parameter-independent channel with regular support,*

$$s_\psi(H) = \mathbb{E}_\psi[s_\psi(O) \mid H]\,. \tag{4}$$

The corresponding Fisher information satisfies

$$F_H = \mathrm{Cov}_\psi(\mathbb{E}_\psi[s_\psi(O) \mid H])\,, \qquad F_O - F_H = \mathbb{E}_\psi[\mathrm{Cov}_\psi(s_\psi(O) \mid H)] \succeq 0\,. \tag{5}$$

To interpret the proposition, fix an output h and consider all neighborhoods mapped to it. Their average score is the evidence about ψ that remains in h, while the variation among their scores is the evidence lost during aggregation. If the neighborhoods mapped to each h have the same score, then H preserves all Fisher information even when the exact neighborhood remains ambiguous. Operationally, $F_H = F_O$ means that H distinguishes infinitesimal changes in ψ as well as O, to first order. Under standard regularity conditions, the two observations provide the same local asymptotic testing and estimation precision.

4.2 Retention Ratios and Probe Bounds

We call the fraction of the Fisher information in O that remains in H *information retention*. Since ψ may contain several parameters, we measure this fraction along a direction v, such as a change in expected degree. The corresponding *information-retention ratio* is

$$\eta_v(H) = \frac{F_H(v,v)}{F_O(v,v)} \quad \text{when } F_O(v,v) > 0\,. \tag{6}$$

When only λ varies, this ratio compares the local evidence for a change in expected degree available from H with the evidence in the complete neighborhood. $\eta_v(H) = 1$ gives equal first-order sensitivity to that change, while $\eta_v(H) = 0$ means that the distribution of H is locally insensitive along the direction. This quantity measures a change in the neighborhood law. To summarize all identifiable directions at once, we use the normalized matrix $E_H = F_O^{-1/2} F_H F_O^{-1/2}$. Its eigenvalues lie in $[0, 1]$, and $E_H = I$ represents full retention in every locally identifiable direction. App. A gives the coordinate-invariant formulation and handles directions with zero full-neighborhood information. We use F_H as the probe-independent reference and label it *canonical information* in figures and tables.

Computing F_H exactly can be difficult. We therefore use coarsenings of H and fitted score probes to obtain certified lower bounds. A *score probe* is a diagnostic model trained to predict $s_\psi(O)$ from H, in the sense of the probing classifiers used to interpret learned representations (Alain & Bengio, 2017; Belinkov, 2022). Fig. 1 follows one categorical mean from its neighborhood messages to the arithmetic degree channel, then shows two distinct limits: finite precision can merge rational grids, and canonical retention need not imply recovery by a regular score readout. The following proposition formalizes the lower bounds obtained from coarsenings and probes (the proof is in App. A).

***Proposition* 2.** *Let $g_1(H)$ be a measurable function of $g_2(H)$, and let $\phi(H)$ be any fixed probe feature map. Define $C = \mathrm{Cov}_\psi(\phi(H), s_\psi(O))$ and $V = \mathrm{Cov}_\psi(\phi(H))$. Then*

$$F_{g_1(H)} \preceq F_{g_2(H)} \preceq F_H, \qquad L_{\phi(H)} \coloneqq C^\top V^\dagger C \preceq F_H\,. \tag{7}$$

Equality in the probe bound requires $s_\psi(H)$ to lie in the closed linear span of the features.

The proposition gives two lower bounds on the information in H. A coarsening $g(H)$ can only discard information, so its Fisher information certifies how much information the full aggregate must contain. Likewise, a probe feature map $\phi(H)$ certifies how much score information can be recovered from those features. We call the remaining difference $F_H - L_{\phi(H)}$ the *probe approximation gap*.

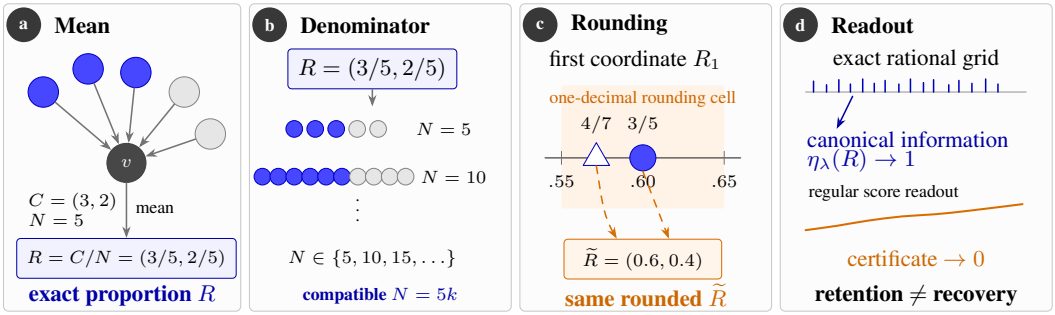


Figure 1: Where the arithmetic degree channel lives, and where access to it can fail. (a) Five categorical messages produce an exact proportion. (b) Its reduced denominator restricts compatible degrees without identifying the realized degree. (c) The values $4/7$ and $3/5$ fall in the same one-decimal rounding cell. (d) At high expected degree, exact proportions have asymptotically complete canonical degree information (Thm. 3), while uniformly regular score readouts an asymptotically vanishing fraction of this information (Thm. 4).

5 Precision Determines What Proportions Retain

We now apply the framework to categorical messages, whose mean gives the empirical category proportions. We show that exact proportions encode expected degree through their arithmetic structure, while regular readouts recover an asymptotically vanishing fraction of this information. We then identify a precision regime that erases this degree signal. In the same regime, the aggregate retains its ability to detect changes in the neighbor-category probabilities.

5.1 A Poisson Model Separating Degree from Composition

We assume that each neighboring node belongs to one of K categories and sends the corresponding one-hot message. Let N_b be the number of neighbors in category b, and let $N = \sum_b N_b$ be the total degree. We model these counts as independent Poisson variables,

$$C = (N_1, \dots, N_K), \qquad N_b \overset{\text{ind}}{\sim} \text{Pois}(\lambda p_b), \qquad p_b > 0, \quad \sum_{b=1}^{K} p_b = 1\,. \tag{8}$$

The Poisson intensity λ equals the expected degree, while p describes the neighborhood composition, i.e., the probabilities with which neighbors belong to the K categories. For a realized count vector $c = (c_1, \dots, c_K)$ with $n = \sum_b c_b$, the likelihood factorizes as $\Pr_{\lambda,p}(C = c) = \Pr_\lambda(N = n)\Pr_p(C = c \mid N = n)$. Thus, the distribution of total degree N depends only on λ and the conditional allocation $C \mid N$ depends only on p. Mean aggregation returns $R = C/N$ for $N > 0$ and a separate symbol for $N = 0$.

Under this model, the likelihood of the complete neighborhood O depends on (λ, p) only through C. Therefore, C contains all the Fisher information in O about these parameters. Writing ∂_λ for the parameter direction that varies λ while holding p fixed, and identifying composition directions with vectors $u, v \in \mathbb{R}^K$ satisfying $\mathbf{1}^\top u = \mathbf{1}^\top v = 0$, the Fisher information is

$$F_O(\partial_\lambda, \partial_\lambda) = \frac{1}{\lambda}, \quad F_O(u, v) = \lambda \sum_b \frac{u_b v_b}{p_b}, \quad F_O(\partial_\lambda, u) = 0\,. \tag{9}$$

Equation (9) is our benchmark. Its first two entries measure sensitivity to expected degree and composition, respectively. The zero cross-term separates the two directions locally. For $N > 0$, the mean aggregate is exactly $R = C/N$. Degree appears only through the exact fractional values in R, which can constrain N. The next subsections show when these values retain degree information and how numerical precision affects that signal. The derivation is in App. B.

5.2 Exact Proportions Preserve Degree Information at High Degree

Mean aggregation returns the category proportions $R = C/N$. The total degree affects which exact values of R are possible. We show that this arithmetic signal is surprisingly strong: for $K \geq 2$, R asymptotically preserves all local Fisher information about degree and composition as λ grows.

For example, $R = (3/5, 2/5)$ can arise at degrees $5, 10, 15, \ldots$, while degrees 6 and 7 are excluded. In general, express the entries of R using their smallest common denominator q. Every compatible degree must be a multiple of q, so the exact mean restricts degree to a sparse set of possible values.

We quantify the remaining ambiguity using Prop. 1. The degree score is $s_\lambda(O) = N/\lambda - 1$, so the proposition gives missing Fisher information $\mathbb{E}[\mathrm{Var}(N \mid R)]/\lambda^2$. Dividing by the raw degree information $F_O(\partial_\lambda, \partial_\lambda) = 1/\lambda$ yields

$$1 - \eta_\lambda(R) = \frac{\mathbb{E}[\mathrm{Var}(N \mid R)]}{\lambda}\,. \tag{10}$$

The right-hand side is the degree uncertainty remaining after aggregation, normalized by the original uncertainty $\mathrm{Var}(N) = \lambda$. We prove that this fraction vanishes.

Theorem 3. *Fix $K \geq 2$ and $\epsilon \in (0, 1/K)$. As $\lambda \to \infty$, the following convergences hold uniformly over all category-probability vectors p satisfying $p_b \geq \epsilon$ for every b:*

$$\frac{\mathbb{E}[\mathrm{Var}(N \mid R)]}{\lambda} \longrightarrow 0, \qquad \left\| F_O^{-1/2} F_R F_O^{-1/2} - I_K \right\|_{\mathrm{op}} \longrightarrow 0\,. \tag{11}$$

The first convergence gives $\eta_\lambda(R) \to 1$ for degree. The matrix convergence extends this conclusion to every local direction combining degree and composition. Thus, the exact mean distinguishes nearby values of λ and p as well as the complete count vector, to first order. Exact degree reconstruction can remain ambiguous, but that ambiguity has negligible statistical cost. The proof in App. B controls count vectors with unusually large common divisors.

The condition $K \geq 2$ is essential: with $K = 1$, every nonempty mean is identical, so the arithmetic degree channel disappears. By contrast, supplying N alongside R recovers $C = NR$ exactly and therefore restores the complete information at every finite λ (Prop. 9).

We next ask whether a regular learned readout can extract this retained degree information. Because R concentrates around p as λ grows, we standardize $R - p$ so that its typical fluctuations remain comparable across expected degrees.

Theorem 4 (Informal)***.*** *For fixed interior p and $K \geq 2$, any family of readouts applied to these standardized deviations recovers a vanishing fraction of the degree information in the exact proportions as $\lambda \to \infty$, provided the offsets and Lipschitz constants are uniformly bounded across the family.*

Thus, exact proportions can retain asymptotically all degree information while regular readouts recover asymptotically none of it. This includes fixed-width, fixed-depth networks with Lipschitz activations and uniformly bounded parameters. The formal statement, precise normalization, proof, and finite-degree rate are in App. C.

5.3 Small Perturbations Remove Degree-Shift Information

The degree signal above relies on exact arithmetic. We now show that small numerical perturbations can erase this signal while preserving all local information about composition. The two signals live at different scales: at degree N, changing one neighbor's category moves the proportion by $1/N$, whereas random composition fluctuations have scale $1/\sqrt{N}$. Noise at an intermediate scale blurs the degree-dependent grid and leaves the larger composition fluctuations visible,

$$\underbrace{1/N}_{\text{count grid spacing}} \ll \underbrace{\tau}_{\text{noise scale}} \ll \underbrace{1/\sqrt{N}}_{\text{composition fluctuations}}\,.$$

Because the entries R_b of $R = C/N$ satisfy $\sum_b R_b = 1$, the proportion vector has only $K - 1$ free coordinates. Let $L \in \mathbb{R}^{(K-1)\times K}$ have rank $K - 1$ and satisfy $L\mathbf{1} = 0$, so that LR uniquely determines R.

Let λ_0 be a reference expected degree and define the perturbed observation

$$H_{\lambda_0} = \perp \quad \text{if } N = 0, \qquad H_{\lambda_0} = LR + \tau_{\lambda_0} Z \quad \text{if } N > 0, \tag{12}$$

where Z is independent noise satisfying the regularity conditions in App. D. The intermediate scale above corresponds to

$$\lambda_0 \tau_{\lambda_0} \to \infty, \qquad \sqrt{\lambda_0}\tau_{\lambda_0} \to 0.$$

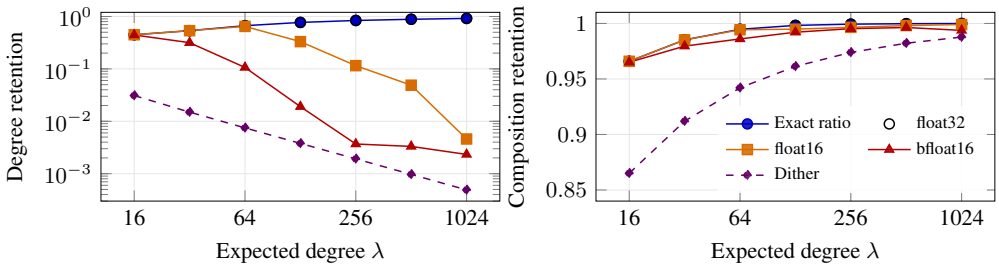


Figure 2: Information retained by exact, rounded, and dithered category proportions as expected degree varies: degree (left) and composition (right), for two Poisson neighbor types with $p_0 = 0.35$.

The first condition makes the perturbation large relative to the degree-dependent grid spacing, while the second makes it negligible relative to ordinary composition fluctuations.

Theorem 5 (Informal)**.** *Under this intermediate scaling, H_{λ_0} becomes asymptotically insensitive to local expected-degree shifts while preserving local composition shifts.*

The limiting experiment retains all local composition information (App. D.1). Consequently, distinguishing two nearby expected degrees from a single perturbed aggregate becomes asymptotically no better than random guessing. Subtractive-dithered quantization gives a concrete instance: a step size $\delta_{\lambda_0} = \lambda_0^{-\alpha}$ with $1/2 < \alpha < 1$ lies in this regime and asymptotically removes degree information while preserving composition (Cor. 11). The formal local alternatives, total-variation limit, testing result, and finite-λ bound are in App. D.

Finite precision selectively removes the degree signal. Fig. 2 shows how the numerical format changes the information retained by the same mean. At $K = 2$, $p_0 = .35$, and $\lambda = 1024$, exact ratios and the float32 baseline retain 91.85% of degree information. Float16 and bfloat16 retain only 0.46% and 0.23%, while every format shown retains above 98% of composition information. Reduced precision can therefore erase almost all degree information while leaving composition nearly unchanged. The exact outcome also depends on which values are stored and where they are cast (Apps. G and G.2).

The asymptotic perturbation result and finite-precision calculations reveal the same selectivity: degree is encoded at a much finer numerical scale than composition. Consequently, the amount of degree information retained by a mean depends strongly on numerical precision.

6 Continuous Messages and Repeated Aggregation

Sec. 5 established a precision-sensitive categorical degree channel and an asymptotic gap between retention and regular-readout recovery. Before studying learned readouts at finite degree in Sec. 7, we ask whether retention changes for continuous messages and repeated aggregation. For Gaussian vectors, we quantify information about expected degree and the message mean after one and any fixed number of layers.

One-layer Gaussian mean. We first analyze a single two-dimensional mean.

Theorem 6. *For $(\lambda, \theta) \in (0, \infty) \times \mathbb{R}^2$, let $N \sim \mathrm{Pois}(\lambda)$ and, conditionally on N, let $X_i \overset{\text{iid}}{\sim} \mathcal{N}_2(\theta, \Sigma)$, with fixed $\Sigma \succ 0$. The complete neighborhood observation is $O = (N, X_1, \ldots, X_N)$. Mean aggregation gives $H = \perp$ for $N = 0$ and $H = N^{-1} \sum_i X_i$ otherwise. Then the fractions of Fisher information retained by H, equivalently the generalized eigenvalues of (F^H, F^O), are*

$$\eta_\lambda = \frac{1 - e^{-\lambda}}{\lambda}, \qquad \eta_{\theta,1} = \eta_{\theta,2} = 1 - \mathbb{E}\left[\frac{N}{(N+1)^2}\right]. \tag{13}$$

Moreover, the λ–θ block of F^H is zero. Consequently, $E_H \to \mathrm{diag}(0, 1, 1)$ as $\lambda \to \infty$.

Thm. 6 gives the exact information retention from the complete neighborhood observation O to the mean aggregate H. The mean aggregate is asymptotically selective. It preserves nearly all local information about small changes in the expected message vector $\theta = \mathbb{E}[X_i]$. The fraction of

information it retains about expected degree decreases as $O(1/\lambda)$. Fig. 4(a) in App. F illustrates this separation, and App. E gives the proof.

Repeated mean aggregation. We next ask whether the one-layer separation persists through multiple GNN layers. To isolate aggregation, we use a depth-L Poisson tree: starting from the root, each internal node independently has $\mathrm{Pois}(\lambda)$ children, and Gaussian messages are placed at depth L and propagated to the root through successive means. At high degree, the root combines information from a number of leaf messages whose expectation is λ^L. Thm. 7 identifies the surviving information.

***Theorem* 7.** *Consider a depth-L single-type Poisson tree with mean offspring λ and iid leaf messages $\mathcal{N}_d(\theta, \Sigma)$. Let $H_L \in \mathbb{R}^d$ be the root output after L recursive means, using a fixed vector $a_\emptyset \in \mathbb{R}^d$ for empty neighborhoods, and let O_L contain the topology and leaf messages. The Fisher information in O_L is $I^O_{\lambda,L} = \lambda^{-1}\sum_{k=0}^{L-1}\lambda^k$ for λ and $F^O_{\theta,L} = \lambda^L\Sigma^{-1}$ for θ. Under*

$$\lambda = \lambda_0 + t/\sqrt{I^O_{\lambda,L}(\lambda_0)}, \qquad \theta = \theta_0 + \lambda_0^{-L/2}\Sigma^{1/2}u, \tag{14}$$

as $\lambda_0 \to \infty$, $\lambda_0^{L/2}\Sigma^{-1/2}(H_L - \theta_0)$ converges in total variation to $\mathcal{N}_d(u, I_d)$, uniformly for (t, u) in compact sets. The limit is the same for every t.

Thm. 7 shows that recursive mean separates message content from degree information at every fixed depth: as $\lambda_0 \to \infty$, local shifts in the expected message remain detectable, while the local expected-degree shifts considered in Eq. (14) become undetectable from the root output. In Eq. (14), this is reflected by the limit depending on u but not on t. Fig. 4(b) in App. F shows the same pattern at finite degree across depths two to four. The proof is in App. F, and the finite-degree calculation in App. H.

7 Experiments

We evaluate whether learned models recover and use degree information retained by mean aggregation. In Poisson neighborhoods, we compare the Fisher information in the aggregate with that recovered by neural readouts; App. I checks the same gap in finite CSBMs. We then ask whether explicit degree improves prediction under degree shift and reduces sensitivity to perturbations of the mean. We train GNNs at expected degree 24, freeze their parameters, and evaluate them at expected degrees 12 and 48, with and without explicit degree. We also round and perturb their layerwise means to measure the resulting changes in prediction.

Information recovery from categorical means. We quantify how much of the retained degree information a smooth learned readout can recover. At $\lambda = 256$ and $p = (.35, .65)$, we first compute the Fisher information contained in exact and quantized category proportions, independently of any readout. We call this probe-independent quantity *canonical information*. We compare it with lower bounds from three independently trained neural score probes and a fixed 12-bit output partition. Fig. 3 reports the canonical information and both lower bounds across output precisions. Each probe predicts the degree score from the aggregate. The partition and probe lower bounds follow from Prop. 2; Prop. 1 identifies the optimal score predictor from the aggregate.

In Fig. 3(a), we plot, as a function of output precision, the fraction of complete-neighborhood degree information retained by the aggregate (blue curve with circles) and the lower bounds obtained from the fixed 12-bit partition (orange dashed curve with squares) and the neural score probes (purple dotted curve with triangles). At exact output, the aggregate retains 83.9% of the complete-neighborhood degree information, the 12-bit discretization retains 14.1%, and the fitted probes recover only 0.09%–0.18%. A finer 16-bit discretization preserves 83.3% (App. G.1). These results isolate a large recovery gap: the aggregate retains a strong arithmetic degree signal, while the fitted smooth probes recover only a small fraction of it.

In Fig. 3(b), we plot the analogous comparison for composition information using the same visual encodings. The curves nearly coincide: the probes recover almost all composition information retained by the aggregate. Together, the two figures show that the same probes recover the smooth composition signal but miss almost all of the arithmetic degree signal. App. G.1 specifies the discretization and probe-training setup and verifies the numerical Fisher-information calculations.

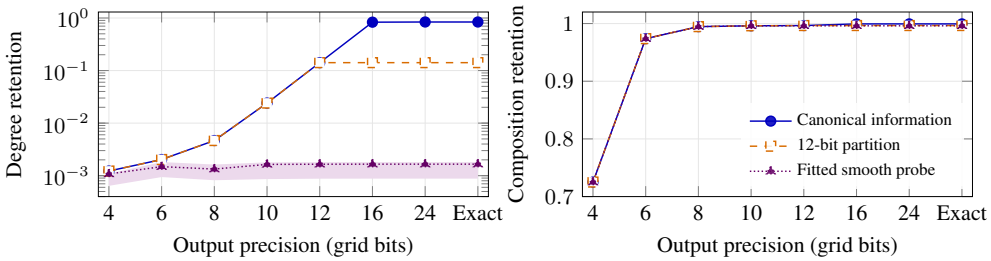


Figure 3: Fraction of complete-neighborhood information retained by the aggregate and a 12-bit partition, and recovered by fitted smooth probes, as output precision varies: degree (left) and composition (right).

Table 1: Frozen finite-CSBM transfer (log loss, nats/node); interval construction is in App. J.1.

Environment	Mean	Mean+degree	Sum	Degree advantage [range]
Source $\lambda = 24$	0.04491	0.04332	0.05092	$0.00159\ [0.00096, 0.00223]$
$\lambda = 12$	0.40839	0.26778	0.26922	$0.14061\ [0.11501, 0.16314]$
$\lambda = 48$	0.00737	0.01850	0.00322	$-0.01113\ [-0.03935, 0.00309]$

Explicit degree under degree shift. We next ask whether explicit degree improves generalization when the expected degree changes between training and evaluation. We train two-layer message-passing GNNs with width 64. The mean model uses mean aggregation, the mean-plus-degree model concatenates source-standardized log degree with the neighbor mean at both layers, and the sum model replaces mean aggregation with sum aggregation. Every model receives only node features and graph structure. In Table 1, we report log loss for the three models and the advantage of mean plus degree over mean, computed as the loss of mean minus the loss of mean plus degree. At expected degree 12, supplying degree reduces log loss by 0.14061 nats, or 34.4%, with a prespecified range of $[0.11501, 0.16314]$. Its effect on accuracy remains inconclusive. An exploratory error analysis indicates that the loss improvement comes largely from fewer highly confident mistakes: predictions that are wrong with at least .99 confidence account for 55.9% of the mean model's loss and 13.3% of the mean-plus-degree model's loss. At expected degree 48, mean plus degree has higher point-estimate loss than mean, and the reported range includes zero; the sum GNN achieves the lowest loss. Explicit degree therefore provides a large benefit under the low-degree shift, but this benefit does not extend uniformly across degree regimes. How a GNN incorporates neighborhood size also matters. App. J.1 gives the training and evaluation design, paired uncertainty and exploratory error analyses, matched controls, and the complete grid of additional shifts.

Sensitivity to aggregate precision and noise. In App. J.2, we show how trained GNN predictions respond to reduced precision and additive noise in their mean aggregates. We hold the weights and all other inputs fixed. Rounding each completed layerwise mean to float16 or bfloat16 changes average log loss by less than 4.3×10^{-6} nats per node across the source and both degree shifts. Reducing the precision of the mean aggregates therefore has no measurable effect on predictive performance. We then apply the same Gaussian perturbations to the first-layer means in both mean-based architectures. At expected degree 12, noise increases log loss by 0.03860 nats for mean and 0.01881 for mean plus degree, so explicit degree halves the increase. At expected degree 48, it provides no corresponding protection. Table 6 reports these contrasts. The appendix also gives the experimental design, uncertainty analysis, scale sweep, and uniform-noise check.

8 Conclusion

We showed that mean aggregation preserves more degree information than non-injectivity suggests. Exact categorical proportions asymptotically retain all local information about expected degree and composition. Rounding and noise can remove the degree signal while preserving composition. Gaussian means retain a vanishing fraction of degree information at high degree across fixed depth, but preserve information about the message mean. Our experiments expose a gap between preservation and use: tested readouts extract little from exact proportions, and reduced precision barely changes trained-GNN predictions. Adding degree improves log loss and robustness to aggregate noise under the low-degree shift, but does not help at high degree.

Limitations: We study controlled Poisson, Gaussian, and finite-CSBM models. Other structural properties and learned hidden states remain open.

Reproducibility statement

The appendices provide complete proofs, experimental designs, metrics, uncertainty procedures, and result tables. A supplementary source package containing code and resolved configurations is available here: https://anonymous.4open.science/r/what-mean-aggregation-retains-in-gnns-superseded-private-32DB/.

AI use statement

Generative AI assisted code, verification, literature organization, and editing. The authors checked the resulting derivations, analyses, citations, and prose and take responsibility for the paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2017.

Pradeep Kr. Banerjee, Kedar Karhadkar, Yu Guang Wang, Uri Alon, and Guido Montúfar. Oversquashing in GNNs through the lens of information contraction and graph expansion. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2022.

Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Optimality of message-passing architectures for sparse graphs. In *Advances in Neural Information Processing Systems*, 2023.

Aseem Raj Baranwal. The value of depth in message passing on sparse graphs: A Kesten–Stigum dichotomy. *arXiv preprint arXiv:2607.16676*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Dexiong Chen, Paolo Pellizzoni, and Karsten Borgwardt. Fisher information embedding for node and graph learning. In *International Conference on Machine Learning (ICML)*, 2023.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, 2022.

Xiner Li, Shurui Gui, Youzhi Luo, and Shuiwang Ji. Graph structure extrapolation for out-of-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2024.

Martyna Lukaszewicz and Brian Dennis. Determination of sample size for a multinomial model coupled with the phenology model. *Frontiers in Applied Mathematics and Statistics*, 10, 2024.

Joshua Rapp, Robin M. A. Dawson, and Vivek K. Goyal. Estimation from quantized gaussian measurements: When and how to use dither. *IEEE Transactions on Signal Processing*, 67(13): 3429–3443, 2019.

Leonard Schuchman. Dither signals and their effect on quantization noise. *IEEE Transactions on Communication Technology*, 12(4):162–165, 1964.

Maximilian Weimar, Lukas M. Rachbauer, Ilya Starshynov, Daniele Faccio, Linara Adilova, Dorian Bouchet, and Stefan Rotter. Fisher information flow in artificial neural networks. *Physical Review X*, 15:031072, 2025.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Minwoo Yu and Young-guk Ha. What softmax throws away: Mass-aware attention for evidence accumulation. *arXiv preprint arXiv:2607.22781*, 2026.

Shuo Zhang and Lei Xie. Improving attention mechanism in graph neural networks via cardinality preservation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Appendix overview. App. A proves the information-retention and probe results; App. B proves exact-arithmetic completeness; App. C gives the formal regular-readout result; App. D gives the formal perturbation result and its testing consequences; Apps. E–F cover continuous messages and repeated aggregation; and Apps. G–J give the finite-degree calculations and trained-GNN experiments.

A Proofs for Information Retention and Probe Bounds

An output's score must give the correct likelihood derivative for every function of that output. The first proof uses this requirement to identify the score as a conditional expectation. The second restricts that expectation to coarser observations or a fitted feature space, giving the lower bounds used to diagnose readout-recovery gaps.

Proof of Proposition 1. Let $K(\mathrm{d}h \mid o)$ be the parameter-independent Markov kernel of the channel. Differentiability in quadratic mean is preserved under a fixed Markov kernel, so the marginal experiment has an L^2 score s_ψ^H. For every bounded measurable scalar function f, differentiation of its expectation under the joint law $P_\psi(\mathrm{d}o)K(\mathrm{d}h \mid o)$ gives

$$\mathbb{E}_\psi[f(H)s_\psi^H(H)] = \nabla_\psi \mathbb{E}_\psi[f(H)] \tag{15}$$

$$= \mathbb{E}_\psi[f(H)s_\psi(O)] \tag{16}$$

$$= \mathbb{E}_\psi[f(H)\mathbb{E}_\psi[s_\psi(O) \mid H]]\,. \tag{17}$$

The defining property of conditional expectation identifies $s_\psi^H(H) = \mathbb{E}_\psi[s_\psi(O) \mid H]$ almost surely. Taking covariance and applying the law of total covariance gives the result. This argument also covers deterministic maps with continuous outputs, which need not admit a channel density against a common reference measure. □

Proof of Proposition 2. Conditional expectation is an orthogonal projection in L^2. Because $\sigma(g_1(H)) \subseteq \sigma(g_2(H)) \subseteq \sigma(H)$, projection onto these nested subspaces gives the two positive-semidefinite inequalities. The best linear projection of the score onto the centered features has coefficient $V^\dagger C$ and covariance $C^\top V^\dagger C$; projecting onto the full $\sigma(H)$ subspace can only increase that covariance. Equality holds exactly when the conditional score is in the closed linear feature span. □

Coordinate-invariant retention. On the identifiable tangent space, directional contraction is summarized by

$$F_H h_i = \eta_i F_O h_i, \qquad h_i^\top F_O h_j = \delta_{ij}. \tag{18}$$

The values $\eta_i \in [0, 1]$ are invariant under smooth reparameterization. Write $E_H = F_O^{-1/2} F_H F_O^{-1/2}$ on this space. Map its eigenvectors back to the original parameter space to interpret graph shifts. For a specified direction v, write $\eta_v(H) = F_H(v, v)/F_O(v, v)$ when the denominator is positive.

B Proof of Exact-Arithmetic Fisher Completeness

Once proportions are known, only their integer multiplier is missing from the count vector. We first show that this one unknown controls the entire missing Fisher matrix. We then decode the total using the denominator and bound the error separately on typical counts and Poisson tails. This is enough because conditional variance is no larger than the mean-squared error of any such decoder.

For $N > 0$, write

$$G = \gcd(N_1, \ldots, N_K), \qquad A = C/G, \qquad q = \sum_b A_b = N/G. \tag{19}$$

The exact rational vector R determines the primitive vector A and denominator q; only the integer multiplier G remains unknown.

To see why the hidden total controls both kinds of information, write the scores in terms of N and the observed proportions:

$$s_\lambda(C) = \frac{N}{\lambda} - 1, \qquad s_p(C; u) = \sum_b N_b \frac{u_b}{p_b} = N g_u(R), \quad g_u(r) = \sum_b r_b \frac{u_b}{p_b}. \tag{20}$$

***Proposition* 8** (Exact missing-information matrix)**.** *Conditional on R, every score component is affine in the same hidden total N. Hence*

$$F_O - F_R = \mathbb{E}\big[\mathrm{Var}(N \mid R)\, w(R)w(R)^\top\big], \tag{21}$$

where the intensity component of the covector $w(R)$ is $1/\lambda$ and its action on a composition tangent is $w(R)[u] = g_u(R)$. In particular,

$$1 - \eta_\lambda(R) = \frac{\mathbb{E}[\mathrm{Var}(N \mid R)]}{\lambda}, \tag{22}$$

$$F_O(u,u) - F_R(u,u) = \mathbb{E}[g_u(R)^2 \mathrm{Var}(N \mid R)]. \tag{23}$$

Proof of Proposition 8. Given R, the count vector is $C = NR$. The intensity score is $N/\lambda - 1$ and the score in a composition direction u is $Ng_u(R)$. Conditional covariance therefore equals $\mathrm{Var}(N \mid R)w(R)w(R)^\top$. Taking expectation and applying Proposition 1 gives all three displayed identities. □

Proof of Theorem 3. The exact-ratio support is the countable rational simplex together with the empty symbol. For every nonempty support point, define $q(r)$ through its unique primitive integer representation and define it arbitrarily off the support. This makes the primitive-denominator map measurable. At the empty symbol set the decoded count to zero; for nonempty observations at a fixed model point, consider

$$\widehat{N}_\lambda(R) = q\,\mathrm{round}(\lambda/q). \tag{24}$$

This decoder depends on λ. The pointwise conditional-variance bound permits that dependence because the conditional mean minimizes squared error among all measurable functions of R, including functions selected separately at each model point. It is enough to show $\mathbb{E}[(N - \widehat{N}_\lambda)^2] = o(\lambda)$.

The decoder succeeds whenever plausible multiples of q are sufficiently far apart. We make this precise by restricting each count to a central Poisson window. For $t \geq 1$, define

$$B_t = \left\{ \max_b |N_b - \lambda p_b| \leq t\sqrt{\lambda} \right\}. \tag{25}$$

On B_t, $|N - \lambda| \leq Kt\sqrt{\lambda}$. If $q > 2Kt\sqrt{\lambda}$, N/q is the unique integer within one half of λ/q, so $\widehat{N}_\lambda = N$. Otherwise $|N - \widehat{N}_\lambda| \leq 2Kt\sqrt{\lambda}$.

We therefore need to bound the probability of a small denominator inside that window. Count the compatible primitive vectors and their multipliers for $q \leq Q = 2Kt\sqrt{\lambda}$. Centrality implies

$$\left| \frac{N_b}{N} - p_b \right| \leq C_K \frac{t}{\sqrt{\lambda}}. \tag{26}$$

For a fixed q, the first $K-1$ coordinates of the primitive vector A therefore lie in integer intervals of length $O(tq/\sqrt{\lambda})$; the last coordinate is fixed by $\mathbf{1}^\top A = q$. Ignoring primitivity only enlarges the count, giving

$$\#\{A\} \leq C_K \left(1 + \frac{tq}{\sqrt{\lambda}}\right)^{K-1}. \tag{27}$$

For any such A, the central interval contains at most $1 + C_K t\sqrt{\lambda}/q$ compatible multipliers. Summing over $q \leq Q$ yields

$$S \leq C_K \sqrt{\lambda}(1+t)^{2K}(1 + \log\lambda). \tag{28}$$

Uniformly over $p_b \geq \epsilon$, the maximal joint Poisson atom is at most $C_{K,\epsilon}\lambda^{-K/2}$ by Stirling's bound at each marginal mode. Multiplying by (28),

$$\Pr(B_t, q \leq 2Kt\sqrt{\lambda}) \leq C_{K,\epsilon}(1+t)^{2K}(1+\log\lambda)\lambda^{-(K-1)/2}. \tag{29}$$

Consequently,

$$\frac{1}{\lambda}\mathbb{E}\Big[(N - \widehat{N}_\lambda)^2 \mathbf{1}_{B_t}\Big] \leq C_{K,\epsilon} t^2 (1+t)^{2K}(1+\log\lambda)\lambda^{-(K-1)/2}. \tag{30}$$

Outside the central window, exact decoding may fail, but the Poisson tail makes the contribution negligible. Nearest-integer rounding and $q \le N$ imply $(N-\widehat{N}_\lambda)^2 \le C(\lambda^2+N^2)$. Poisson Bernstein bounds give $\Pr(B_t^c) \le 2K\exp(-c_{K,\epsilon}t^2)$ when $t = o(\sqrt{\lambda})$, while $\mathbb{E}N^4 = O(\lambda^4)$. Cauchy–Schwarz therefore gives

$$\frac{1}{\lambda}\mathbb{E}\Big[(N-\widehat{N}_\lambda)^2\mathbf{1}_{B_t^c}\Big] \le C\lambda\exp(-c_{K,\epsilon}t^2/2). \tag{31}$$

Choose $t^2 = A\log\lambda$ with A sufficiently large. Both normalized terms vanish for $K \ge 2$, proving $\mathbb{E}[\mathrm{Var}(N \mid R)]/\lambda \to 0$.

The degree deficit now vanishes. To obtain the full matrix conclusion, use the same count-uncertainty bound for a raw-Fisher-unit composition tangent. Equation (9) gives $\lambda\sum_b u_b^2/p_b = 1$. Hence

$$|g_u(r)|^2 \le \sum_b \frac{u_b^2}{p_b^2} \le \frac{1}{\epsilon\lambda}. \tag{32}$$

Proposition 8 bounds every normalized composition deficit by $\epsilon^{-1}\mathbb{E}[\mathrm{Var}(N \mid R)]/\lambda$. The mixed entries vanish by positive-semidefinite Cauchy–Schwarz. In fixed dimension, the positive-semidefinite normalized missing matrix has operator norm bounded by its trace, which tends to zero. This proves the theorem. □

Explicit count provides a positive control for these losses. In the one-generation count model, cardinality-aware channels recover the complete observation, so their Fisher information is preserved exactly.

***Proposition* 9** (Exact cardinality restoration)**.** *In the fully observed count model (8), each of the following is sufficient for C: (i) sum aggregation of one-hot types; (ii) mean aggregation augmented with N; and (iii) any representation containing R and an injective function of N. Therefore $F_H = F_O$ in every identifiable direction.*

Proof. Sum returns C; (N, R) recovers $C = NR$; and an injective cardinality coordinate recovers N before the same reconstruction. □

C Proofs for Regular-Readout Separation

This section verifies why exact arithmetic completeness can coexist with vanishing recovery by uniformly regular functions. The first proof identifies the smooth limiting experiment; the second calculates its leading quadratic degree channel.

Formal statement of Thm. 4. Let $L \in \mathbb{R}^{(K-1)\times K}$ have rank $K-1$ and satisfy $L\mathbf{1} = 0$, set

$$V = L(\mathrm{diag}(p) - pp^\top)L^\top, \qquad X_\lambda = \sqrt{\lambda}V^{-1/2}L(R-p),$$

with $X_\lambda = 0$ on the empty neighborhood, and let $Z_\lambda = (N-\lambda)/\sqrt{\lambda}$. For fixed $A, M < \infty$, let $\mathcal{F}_{A,M}$ contain functions satisfying $|f(0)| \le A$ and $|f(x)-f(y)| \le M\|x-y\|$. Then, for fixed interior p and $K \ge 2$,

$$\inf_{f\in\mathcal{F}_{A,M}} \mathbb{E}[(Z_\lambda - f(X_\lambda))^2] \to 1,$$

and

$$\sup_{f\in\mathcal{F}_{A,M}} \{2\mathbb{E}[Z_\lambda f(X_\lambda)] - \mathbb{E}[f(X_\lambda)^2]\}_+ \to 0.$$

Moreover, with $Q_\lambda = \|X_\lambda\|^2$,

$$L_{(1,Q_\lambda)} = \frac{K-1}{2\lambda} + O(\lambda^{-2}).$$

Proof of Theorem 4. Let $\Sigma_p = \mathrm{diag}(p) - pp^\top$ and define

$$U_\lambda = \frac{C-\lambda p}{\sqrt{\lambda}}.$$

The multivariate Poisson central limit theorem gives $U_\lambda \Rightarrow U \sim \mathcal{N}(0, \mathrm{diag}(p))$. Since $Z_\lambda = \mathbf{1}^\top U_\lambda$ and $C - pN = \sqrt{\lambda}(U_\lambda - pZ_\lambda)$, on $N > 0$ we have

$$X_\lambda = V^{-1/2} L \frac{U_\lambda - pZ_\lambda}{1 + Z_\lambda/\sqrt{\lambda}}. \tag{33}$$

The empty event has probability $e^{-\lambda}$ and does not affect the limit. It follows that

$$(Z_\lambda, X_\lambda) \Rightarrow (Z, X), \qquad Z \sim \mathcal{N}(0,1), \quad X \sim \mathcal{N}(0, I_{K-1}).$$

The two limiting variables are independent. Indeed,

$$\mathrm{Cov}\{\mathbf{1}^\top U, U - p\mathbf{1}^\top U\} = p - p = 0,$$

and multiplication by $V^{-1/2}L$ preserves the zero cross-covariance.

We next show that the score correlation converges to zero uniformly over $\mathcal{F}_{A,M}$. Conditional multinomial fourth-moment formulas, split over $|N - \lambda| \le \lambda/2$ and its Poisson tail, give

$$\sup_{\lambda \ge \lambda_*} \mathbb{E}\|X_\lambda\|^4 < \infty, \qquad \sup_{\lambda \ge \lambda_*} \mathbb{E} Z_\lambda^4 < \infty \tag{34}$$

for every fixed interior p and sufficiently large λ_*. Suppose the uniform correlation convergence failed. Then some sequence $\lambda_j \to \infty$ and $f_j \in \mathcal{F}_{A,M}$ would keep the absolute correlation bounded away from zero. The functions are uniformly bounded and equicontinuous on each compact set. Arzelà–Ascoli and a diagonal argument therefore give a subsequence converging locally uniformly to a function $f \in \mathcal{F}_{A,M}$. Moreover, $|f_j(x)| \le A + M\|x\|$, so (34) makes $Z_{\lambda_j} f_j(X_{\lambda_j})$ uniformly integrable. Joint weak convergence, local uniform convergence, and truncation outside compact sets now yield

$$\mathbb{E}[Z_{\lambda_j} f_j(X_{\lambda_j})] \longrightarrow \mathbb{E}[Z f(X)] = 0,$$

contradicting the chosen sequence. If δ_λ denotes the resulting uniform absolute-correlation bound, then

$$1 - 2\delta_\lambda \le \inf_{f \in \mathcal{F}_{A,M}} \mathbb{E}[(Z_\lambda - f(X_\lambda))^2] \le 1,$$

where the upper bound uses $f = 0$. This proves the first limit. Finally,

$$2\mathbb{E}[Z_\lambda f(X_\lambda)] - \mathbb{E}[f(X_\lambda)^2] \le 2\,|\mathbb{E}[Z_\lambda f(X_\lambda)]|\,,$$

so the second limit follows. The canonical-information statement is exactly the degree part of Theorem 3. □

Proof of the quadratic-rate statement in Theorem 4. Write $d = K - 1$ and set $Q_\lambda = 0$ on the empty event. Conditional on $N = n > 0$, the counts are multinomial and

$$\mathbb{E}[Q_\lambda \mid N = n] = \frac{\lambda}{n} \mathrm{tr}(V^{-1/2} V V^{-1/2}) = \frac{d\lambda}{n}.$$

Because $\mathbb{E} Z_\lambda = 0$,

$$\mathrm{Cov}(Z_\lambda, Q_\lambda) = d\sqrt{\lambda}\left\{1 - e^{-\lambda} - \lambda \mathbb{E}[N^{-1}; N > 0]\right\} \tag{35}$$

$$= -\frac{d}{\sqrt{\lambda}} + O(\lambda^{-3/2}), \tag{36}$$

where the second line uses $\mathbb{E}[N^{-1}; N > 0] = \lambda^{-1} + \lambda^{-2} + O(\lambda^{-3})$. Conditional multinomial fourth moments similarly give

$$\mathrm{Var}(Q_\lambda) = 2d + O(\lambda^{-1});$$

equivalently, Q_λ converges with its second moment to a χ^2_d variable. Proposition 2, applied to the centered scalar feature Q_λ, therefore gives

$$L_{(1,Q_\lambda)} = \frac{\mathrm{Cov}(Z_\lambda, Q_\lambda)^2}{\mathrm{Var}(Q_\lambda)} = \frac{d}{2\lambda} + O(\lambda^{-2}).$$

The quadratic feature is smooth but not globally Lipschitz because its gradient grows linearly, so this rate does not contradict the fixed-Lipschitz limit. □

D PROOF OF THE PERTURBATION RESULT

An unsmoothed discrete law cannot approach a continuous Gaussian in total variation. We first add auxiliary jitter on each conditional count cell so that a multinomial–Gaussian comparison applies. The actual channel noise is larger than this jitter, which lets us remove the auxiliary step at negligible cost. Poisson concentration then replaces the random conditional covariance by its limit. The proof keeps these three errors separate.

Formal statement of Thm. 5. Consider the perturbation channel (12). Assume that Z is independent of the neighborhood, has finite first moment, and satisfies a local translation bound: for some $C, r > 0$,

$$\mathrm{TV}\{\mathcal{L}(Z+a), \mathcal{L}(Z)\} \le C\|a\|, \qquad \|a\| \le r.$$

At each λ_0, the contrast matrix L, noise scale τ_{λ_0}, and law of Z are held fixed as (λ, p) varies locally. Consider

$$\lambda = \lambda_0 + h\sqrt{\lambda_0}, \qquad p = p_0 + u/\sqrt{\lambda_0}, \qquad (h,u) \in \mathcal{K},$$

where $\mathbf{1}^\top u = 0$, p_0 has positive entries, and $\mathcal{K}$ is compact. Let

$$\Sigma_p = \mathrm{diag}(p) - pp^\top, \qquad V_0 = L\Sigma_{p_0}L^\top,$$

and define

$$Y_{\lambda_0} = \perp \quad \text{when } N = 0, \qquad Y_{\lambda_0} = \sqrt{\lambda_0}(H_{\lambda_0} - Lp_0) \quad \text{otherwise.}$$

If

$$\lambda_0\tau_{\lambda_0} \to \infty, \qquad \sqrt{\lambda_0}\tau_{\lambda_0} \to 0,$$

then, uniformly over $(h,u) \in \mathcal{K}$,

$$\mathrm{TV}\{\mathcal{L}_{\lambda,p}(Y_{\lambda_0}), \mathcal{N}(Lu, V_0)\} \to 0.$$

Proof of Theorem 5. Retain the first $d = K-1$ counts $X = (N_1, \dots, N_d)$ and write the columns of L as ℓ_b. Conditional on $N = n > 0$,

$$LR = \ell_K + M\frac{X}{n}, \qquad M = (\ell_1 - \ell_K, \dots, \ell_d - \ell_K), \tag{37}$$

where M is nonsingular. Let U be uniform on $(-1/2, 1/2)^d$ and independent. Ouimet's uniform approximation for interior cell probabilities gives

$$\mathrm{TV}(\mathcal{L}(X+U \mid N=n), \mathcal{N}(np_{1:d}, n\Sigma_{p,1:d})) \le Cn^{-1/2}. \tag{38}$$

Apply the affine ratio map, center and scale, and add $\sqrt{\lambda_0}\tau_{\lambda_0}Z$. Total variation contracts under parameter-independent Markov kernels, so the conditional jittered output is within $Cn^{-1/2}$ of

$$\mathcal{L}\Big(G_{n,p} + \sqrt{\lambda_0}\tau_{\lambda_0}Z\Big), \qquad G_{n,p} \sim \mathcal{N}\Big(\sqrt{\lambda_0}L(p-p_0), \frac{\lambda_0}{n}L\Sigma_pL^\top\Big). \tag{39}$$

We have compared a jittered observation with a Gaussian, but must return to the channel actually observed. Conditional on (X,U), scale invariance and the stated translation condition compare the two kernel translates whose means differ by $\sqrt{\lambda_0}MU/n$:

$$\mathrm{TV}\Big(\mathcal{L}(Y_{\lambda_0} \mid N=n), \mathcal{L}(G_{n,p} + \sqrt{\lambda_0}\tau_{\lambda_0}Z)\Big) \le C\left\{n^{-1/2} + (n\tau_{\lambda_0})^{-1}\right\}. \tag{40}$$

For bounded h, a Poisson variable with mean $\lambda_0 + h\sqrt{\lambda_0}$ lies in $[\lambda_0/2, 2\lambda_0]$ except with probability $Ce^{-c\lambda_0}$. Averaging (40) over this central event contributes

$$C\left\{\lambda_0^{-1/2} + (\lambda_0\tau_{\lambda_0})^{-1}\right\} + Ce^{-c\lambda_0}. \tag{41}$$

Under the local alternatives, the mean in (39) is exactly Lu. On the central event, the covariance of $G_{n,p}$ differs from V_0 by at most a constant times

$$\left|\frac{\lambda_0}{N} - 1\right| + \|p - p_0\|. \tag{42}$$

These matrices remain in a compact subset of the positive-definite cone, where Gaussian total variation is locally Lipschitz in covariance. Moreover,

$$\sup_{|h|\le C} \mathbb{E}\left[\left|\frac{\lambda_0}{N} - 1\right| \mathbf{1}_{\{\lambda_0/2 \le N \le 2\lambda_0\}}\right] \le C\lambda_0^{-1/2}. \tag{43}$$

Finally, the channel noise is negligible at the rescaled sampling scale. Let $G_0 \sim \mathcal{N}(Lu, V_0)$. Pinsker's inequality for equal-covariance Gaussians gives, for fixed z,

$$\mathrm{TV}(\mathcal{L}(G_0 + \epsilon z), \mathcal{L}(G_0)) \le C\epsilon\|z\|. \tag{44}$$

Convexity under mixing and $\mathbb{E}\|Z\| < \infty$ therefore bound removal of the actual kernel noise by $C\sqrt{\lambda_0}\tau_{\lambda_0}$. Combining these terms gives the explicit uniform bound

$$\sup_{(h,u)\in\mathcal{K}} \mathrm{TV}\{\mathcal{L}_{\lambda,p}(Y_{\lambda_0}), \mathcal{N}(Lu, V_0)\} \le C\Big\{\lambda_0^{-1/2} + (\lambda_0\tau_{\lambda_0})^{-1} + \sqrt{\lambda_0}\tau_{\lambda_0}\Big\} + Ce^{-c\lambda_0}. \tag{45}$$

If Z is centered Gaussian, retaining its covariance inside the Gaussian comparison and applying local covariance Lipschitzness replaces the third term by $C\lambda_0\tau_{\lambda_0}^2$. □

***Corollary* 10** (Degree testing)**.** *Under Thm. 5, the minimum equal-prior error for testing two bounded degree changes h_1, h_2 at the same composition change u satisfies*

$$e^*_{\lambda_0}(h_1, h_2; u) \to \tfrac{1}{2}.$$

Proof of Corollary 10. Both laws are within the uniform error $\varepsilon_{\lambda_0} \to 0$ in Theorem 5 of $\mathcal{N}(Lu, V_0)$. The triangle inequality therefore bounds their total variation by $2\varepsilon_{\lambda_0}$. For a test with rejection event A, its equal-prior error is $\frac{1}{2}\{1 - P^{\lambda_0}_{h_2,u}(A) + P^{\lambda_0}_{h_1,u}(A)\}$. Optimizing over tests gives $\{1 - \mathrm{TV}(P^{\lambda_0}_{h_1,u}, P^{\lambda_0}_{h_2,u})\}/2$, hence $1/2 - \varepsilon_{\lambda_0} \le e^*_{\lambda_0} \le 1/2$ uniformly. The statement concerns one aggregate; repeated observations require a separate comparison of the corresponding product laws. □

***Corollary* 11** (Subtractive-dithered quantization)**.** *Let Q_δ be a componentwise uniform quantizer without overload, and let $D \sim \mathrm{Unif}([-\delta/2, \delta/2)^{K-1})$ be independent of the input. Observing only $T_\delta(x) = Q_\delta(x + D) - D$ is equivalent to observing x plus independent uniform noise on a quantization cell (Schuchman, 1964; Rapp et al., 2019). Thus $\delta_{\lambda_0} = \lambda_0^{-\alpha}$, $1/2 < \alpha < 1$, yields a limiting local experiment that loses intensity and preserves composition.*

Proof of Corollary 11. Schuchman's condition makes $Q_\delta(x + D) - D - x$ product-uniform and independent of x. Its translation modulus satisfies $\mathrm{TV}\{k(\cdot - a), k\} \le \sum_j |a_j|$ locally, and it has finite first moment. Taking $\tau = \delta = \lambda_0^{-\alpha}$ with $1/2 < \alpha < 1$ satisfies both scale conditions. A separately retained empty symbol contributes relative intensity efficiency $\lambda/(e^\lambda - 1)$, which vanishes and leaves the limit unchanged. □

D.1 Composition Information in the Robust Limit

The robust limit retains a composition shift Lu with covariance V_0. To compare it with the raw counts, we pull the Gaussian Fisher form back to the simplex directions. The resulting form is exactly the categorical Fisher reference; only the degree direction is absent from the limit.

***Corollary* 12** (Composition isometry and intensity collapse)**.**

$$(Lu)^\top V_0^{-1}(Lv) = \sum_b u_b v_b / p_{0b}, \qquad E_{\text{limiting experiment}} = \mathrm{diag}(0, I_{K-1}). \tag{46}$$

Proof of Corollary 12. The covariance Σ_{p_0} has nullspace $\mathrm{span}\{\mathbf{1}\}$ and is positive definite on the simplex tangent space. Because $L\mathbf{1} = 0$ and L has rank $K - 1$, its restriction from the tangent space to $\mathbb{R}^{K-1}$ is an isomorphism. The categorical Fisher metric on a tangent u is $u^\top \mathrm{diag}(p_0)^{-1}u$. The inverse covariance form induced after the isomorphism $u \mapsto Lu$ is unique and equals

$$(Lu)^\top (L\Sigma_{p_0}L^\top)^{-1}(Lv) = u^\top \mathrm{diag}(p_0)^{-1}v, \tag{47}$$

which gives the result. □

E Proof of the Gaussian Mean-Aggregation Theorem

The information in the mean depends on how well its variance identifies the hidden count. In two dimensions the Gaussian density has a prefactor N, which cancels part of the Poisson factorial and yields an explicit posterior for N. Score projection then reduces the degree and location calculations to expectations under this posterior. This simplification is why the exact formulas are stated in dimension two.

Proof of Theorem 6. The output sample space is the disjoint union of an atom $\{\perp\}$ and $\mathbb{R}^2$, dominated by counting measure on the atom plus Lebesgue measure. For $h \in \mathbb{R}^2$, define

$$Q(h) = (h-\theta)^\top \Sigma^{-1}(h-\theta), \qquad \mu(h) = \lambda e^{-Q(h)/2}. \tag{48}$$

The conditional Gaussian density of H contributes a factor n in dimension two. Hence, for $n \geq 1$,

$$\Pr(N = n \mid H = h) \;\propto\; \frac{\lambda^n}{n!}\, n e^{-nQ(h)/2} = \frac{\mu(h)^n}{(n-1)!}. \tag{49}$$

Normalizing the series shows that

$$\begin{aligned} N \mid H = h \;&\stackrel{d}{=}\; 1 + \mathrm{Pois}(\mu(h)), \\ m(h) := \mathbb{E}[N \mid H = h] = 1 + \mu(h), \qquad & \mathrm{Var}(N \mid H = h) = \mu(h). \end{aligned} \tag{50}$$

At the empty symbol, $m(\perp) = 0$.

The raw scores are

$$s_\lambda(O) = N/\lambda - 1, \qquad s_\theta(O) = \Sigma^{-1}\sum_{i=1}^{N}(X_i - \theta), \tag{51}$$

which gives the stated F_O. Proposition 1 and (50) give, on the continuous component,

$$s_\lambda(H) = m(H)/\lambda - 1, \qquad s_\theta(H) = m(H)\Sigma^{-1}(H-\theta). \tag{52}$$

We next evaluate the covariance of these output scores. Introduce the scalar

$$a_\lambda = \mathbb{E}_{N\sim\mathrm{Pois}(\lambda)}\left[\frac{N}{(N+1)^2}\right], \tag{53}$$

the full Fisher matrices are

$$F_H = \begin{pmatrix} \dfrac{1-e^{-\lambda}}{\lambda^2} & 0 \\ 0 & \lambda(1-a_\lambda)\Sigma^{-1} \end{pmatrix}, \qquad F_O = \begin{pmatrix} \lambda^{-1} & 0 \\ 0 & \lambda\Sigma^{-1} \end{pmatrix}. \tag{54}$$

For the intensity entry, total variance gives

$$\mathrm{Var}(m(H)) = \mathrm{Var}(N) - \mathbb{E}[\mathrm{Var}(N \mid H)]. \tag{55}$$

Conditionally on $N = n > 0$, $nQ(H) \sim \chi^2_2$, and therefore

$$\mathbb{E}[e^{-Q(H)/2} \mid N = n] = \frac{n}{n+1}. \tag{56}$$

It follows that

$$\mathbb{E}[\mathrm{Var}(N \mid H)] = \lambda\mathbb{E}\left[\frac{N}{N+1}\right] = \lambda - 1 + e^{-\lambda}, \tag{57}$$

$$\mathrm{Var}(m(H)) = 1 - e^{-\lambda}. \tag{58}$$

Dividing by λ^2 proves the intensity formula.

For the location block, missing information is

$$F^O_{\theta\theta} - F^H_{\theta\theta} = \mathbb{E}\left[\mathrm{Var}(N \mid H)\Sigma^{-1}(H-\theta)(H-\theta)^\top\Sigma^{-1}\right]. \tag{59}$$

After whitening, rotational symmetry makes this matrix proportional to Σ^{-1}. Its coefficient is one half of its trace. Since

$$\mathbb{E}[Q(H)e^{-Q(H)/2} \mid N = n] = \frac{2n}{(n+1)^2}, \tag{60}$$

the missing block is $\lambda a_\lambda \Sigma^{-1}$, proving the location formula in (54). The intensity score is radial and even in $H - \theta$. The location score is odd, so parity makes the mixed block vanish. Finally, Taylor expansion of $N/(N+1)^2$ on $|N - \lambda| \leq \lambda/2$, together with an exponential Poisson-tail bound, gives $a_\lambda = \lambda^{-1} + O(\lambda^{-2})$. □

F Proofs for Fixed-Depth Score Geometry

This appendix proves Theorem 7. Conditional on a realized nonempty tree, the recursive mean is Gaussian with a random scalar variance factor. The proof shows that this factor concentrates after scaling by λ^L, bounds the probability of an empty internal node, and converts the variance control into total-variation convergence. It then checks that the complete tree and leaves retain the corresponding local degree and location changes.

Proof of Theorem 7. The complete-tree likelihood factorizes over internal vertices. Each observed offspring count has score $(N_v - \lambda)/\lambda$, conditional mean zero, and variance $1/\lambda$. The expected number of internal vertices at levels $0, \dots, L-1$ is $\sum_{k=0}^{L-1} \lambda^k$, so conditional orthogonality gives $I^O_{\lambda,L} = \lambda^{-1} \sum_{k=0}^{L-1} \lambda^k$. If Z_L is the number of leaves, the conditional Gaussian location information is $Z_L \Sigma^{-1}$; since $\mathbb{E} Z_L = \lambda^L$, this gives $F^O_{\theta,L} = \lambda^L \Sigma^{-1}$.

For a realized tree $\mathcal{T}$ with no empty internal node, define the scalar variance factor W_ℓ at depth ℓ by $\mathrm{Cov}(H_\ell \mid \mathcal{T}) = \Sigma W_\ell$. We show that $\lambda^\ell W_\ell$ concentrates near one. This replaces the random leaf weights by their leading variance effect. Empty internal nodes break the conditional Gaussian calculation, so we bound their probability separately. Define an auxiliary variance recursion from $R_0 = 1$, setting $R_\ell = 1$ at an empty node and using the following recursion when $N > 0$:

$$R_\ell = \frac{\lambda}{N} \left(\frac{1}{N} \sum_{i=1}^{N} R^{(i)}_{\ell-1} \right). \tag{61}$$

For $A = (\lambda/N)\mathbf{1}\{N > 0\} + \mathbf{1}\{N = 0\}$, a Poisson concentration split at $N = \lambda/2$ gives

$$\mathbb{E}(A-1)^2 = O(\lambda^{-1}), \quad \mathbb{E}A^2 = O(1), \quad \mathbb{E}[A^2 N^{-1} \mathbf{1}\{N > 0\}] = O(\lambda^{-1}). \tag{62}$$

Conditionally on N, the child variables in (61) are iid. Induction from $R_0 = 1$, using the conditional variance of their empirical mean, yields

$$\mathbb{E}(R_\ell - 1)^2 = O(\lambda^{-1}) \tag{63}$$

for every fixed ℓ. The same estimates hold uniformly when the Poisson mean is $\lambda + O(\sqrt{\lambda})$. On trees with no empty internal node, $R_\ell = \lambda^\ell W_\ell$. The auxiliary recursion can differ from the actual scaled variance only on the empty-node event, whose probability is bounded using the expected number of vertices at each level by $e^{-\lambda} \sum_{k=0}^{L-1} \lambda^k$.

Variance concentration now identifies the root's limiting observation law. Given the tree and outside the sentinel event, $Y_L = \lambda_0^{L/2} \Sigma^{-1/2} (H_L - \theta_0)$ is Gaussian with mean u and scalar covariance multiplier $\lambda_0^L W_L$. For $r \in [1/2, 2]$,

$$D_{\mathrm{KL}}\{\mathcal{N}(0, rI_d) \| \mathcal{N}(0, I_d)\} = \tfrac{d}{2}(r - 1 - \log r) \le C_d (r-1)^2. \tag{64}$$

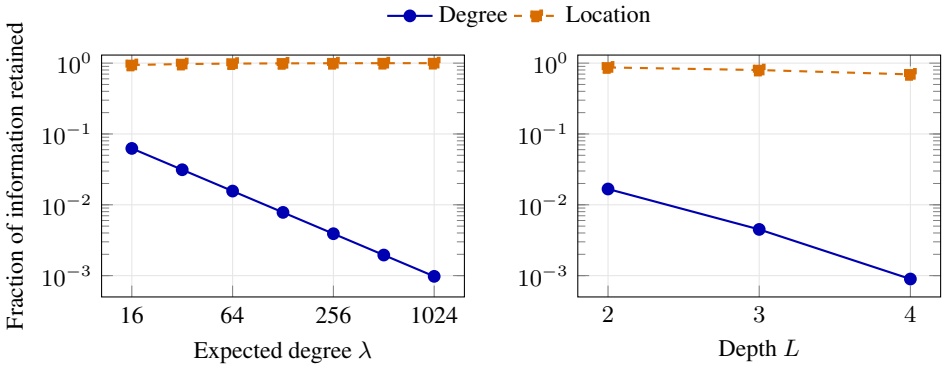
Pinsker's inequality, the L^2 concentration above, convexity of total variation under mixing, and the sentinel bound prove Theorem 7 uniformly on compact local alternatives. For comparison, if Z_L is the number of leaves, then $Z_L/\lambda_0^L \to 1$ in probability under the stated local alternatives. At fixed topology, the leaf-mark log-likelihood ratio for $\theta = \theta_0 + \lambda_0^{-L/2} \Sigma^{1/2} u$ against θ_0 is

$$u^\top \Delta_L - \frac{Z_L}{2\lambda_0^L} \|u\|^2, \qquad \Delta_L \mid \mathcal{T} \sim \mathcal{N}_d\left(0, \frac{Z_L}{\lambda_0^L} I_d\right) \quad \text{under } \theta_0. \tag{65}$$

Thus the raw leaf experiment converges to the full shift $\mathcal{N}_d(u, I_d)$, confirming that the location limit retained by mean is raw-Fisher complete at first order. □

G Categorical Precision and Readout Checks

The exact and perturbed theorems leave a finite-degree question: how much information survives a specific arithmetic operation? We answer it by grouping count states that produce the same recorded



(a) Continuous mean loses degree information as neighborhoods grow

(b) Location survives repeated averaging much better than degree

Figure 4: Selective retention for continuous and repeated mean aggregation. (a) The Gaussian mean preserves location information while its degree efficiency decreases with expected degree. (b) In the finite-state recursive calculation, leaf-location information remains substantially larger than degree information through depth four.

output. We also group exact outputs more coarsely to show how a simpler observation can certify information without reconstructing the full ratio.

For $K = 2$ and $p = (0.35, 0.65)$, we enumerate count states through at least 12 Poisson standard deviations and compute canonical Fisher information by exact conditional-score grouping. Each declared arithmetic channel observes only the scalar first-coordinate ratio. Writing cast_d for rounding to the stated IEEE-like format, the channel uses a distinguished empty symbol when $N = 0$ and otherwise returns, for $N = N_1 + N_2$,

$$T_d = \mathrm{cast}_d\left(\frac{\mathrm{cast}_d(N_1)}{\mathrm{cast}_d(N)}\right),$$

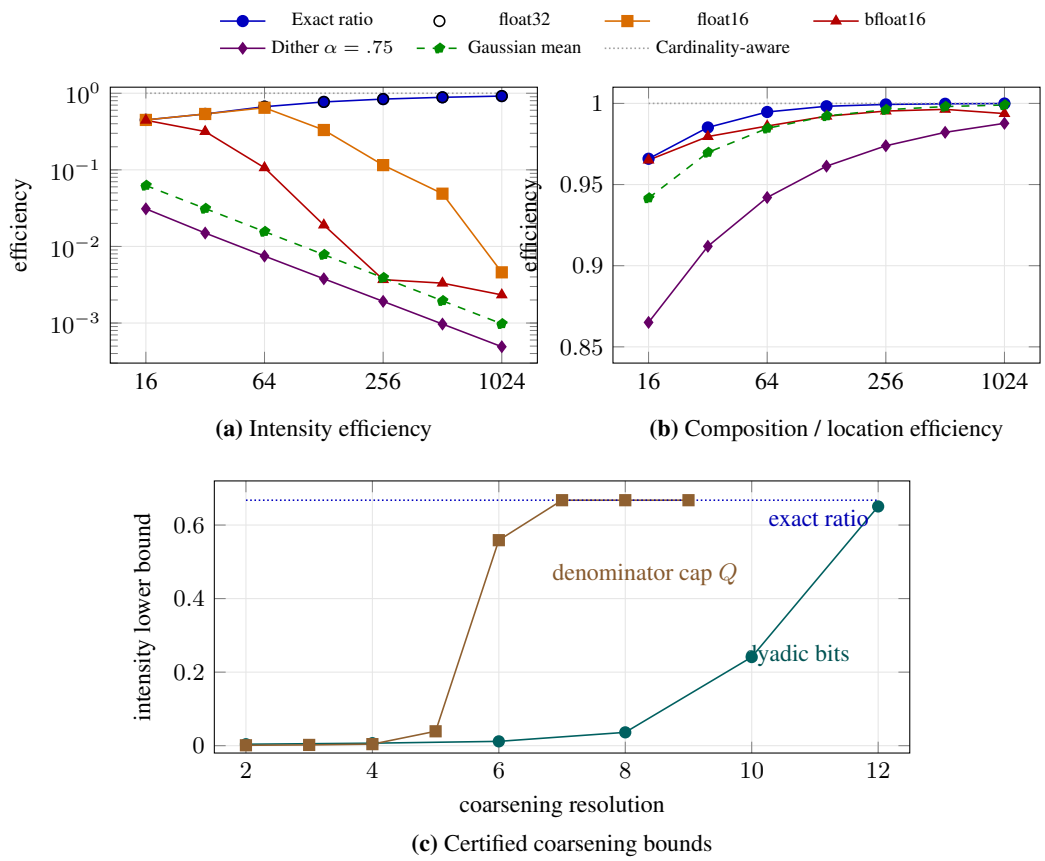
In exact arithmetic, N_1/N determines the full two-coordinate mean. Observing both coordinates after separately rounding them defines a different channel; Appendix G.2 evaluates that control. Figure 5(a) shows a precision transition: at $\lambda = 1024$, the intensity efficiencies of the exact ratio, float32, float16, bfloat16, and subtractive dither with $\delta = \lambda^{-3/4}$ are respectively 0.9185, 0.9185, 0.00459, 0.00233, and 0.000491. The exact Gaussian-mean intensity efficiency is 0.000977. Thus float32 still resolves the arithmetic encoding on this range. Lower precision and dither retain far less intensity efficiency on this finite range; the dither theorem gives asymptotic first-order erasure.

Figure 5(b) shows the complementary fact: at the same intensity, every displayed composition efficiency (or Gaussian location efficiency) is at least 0.9877. Panel (c) implements Proposition 2. At $\lambda = 64$, nested dyadic partitions raise the certified intensity lower bound from 0.00397 to 0.6506 as resolution increases from 2 to 12 bits. Revealing primitive ratios up to denominator cap Q raises it from 0.00109 at $Q = 4$ to the exact 0.66755 by $Q = 128$. Every nesting inequality and every contraction against the exact ratio is checked in positive-semidefinite order.

G.1 Canonical and Recovered Information Across Precision

This audit measures two distinct deficits under the same sampling law: information removed by the output channel and information a readout misses within that channel. The calculation holds the Poisson model fixed at $\lambda = 256$, $p = (.35, .65)$ and crosses output precision with readout resolution. The output is the separate empty symbol or $T_b = 2^{-b}\lfloor 2^b N_1/N \rfloor$, with $b \in \{4, 6, 8, 10, 12, 16, 24\}$; exact $R = N_1/N$ supplies a further channel. The endpoint $R = 1$ retains its own grid value. These deterministic channels are evaluated directly, separately from the additive-noise theorem.

Population information and constructive bounds. Let $Z = F_O^{-1/2} s_O$ denote raw-Fisher-normalized scores, using the degree coordinate $(N - \lambda)/\sqrt{\lambda}$ and composition coordinate $(N_1 - p_1 N)/\sqrt{\lambda p_1(1 - p_1)}$. Grouping the count probabilities and score moments by each output computes



(a) Intensity efficiency

(b) Composition / location efficiency

(c) Certified coarsening bounds

Figure 5: Finite-intensity comparisons; count channels use $K = 2$, $p_0 = 0.35$. (a) Canonical intensity efficiency across exact, emulated arithmetic, dithered, Gaussian-mean, and cardinality-aware channels. float32 markers coincide with exact ratios over this range. (b) Composition, or location for Gaussian mean, is largely retained; the vertical axis spans .84–1.01. (c) Exact measurable coarsenings at $\lambda = 64$ provide monotone certified lower bounds; the horizontal line is the exact-ratio value. IEEE-format results describe the declared rounding rule; other numerical implementations may define different channels.

its canonical Fisher matrix. For each output T_b, the histogram decoder observes only $g_j(T_b) = \lfloor 2^j T_b \rfloor$ and an empty indicator, with $j \in \{2, 4, 6, 8, 10, 12, 16, 24\}$. Population conditional scores for these nested partitions give certified lower bounds by Proposition 2. They require knowledge of the nominal law; no claim about learning these scores from a finite sample is made. On the exact output, the 8-, 12-, and 16-bit degree bounds are .004661, .141360, and .832611, against canonical .839257.

Fitted smooth probes. For each output channel, three independent seeds fit two-output regressors with two width-64 SiLU hidden layers. Each seed shares the same 24,000/6,000 independent training/validation count draws across channels. The scalar input is centered at .35 and divided by $\sqrt{.35 \cdot .65/256}$; a second coordinate indicates emptiness. Networks use float32 arithmetic, Adam with learning rate .001, batches of 1,024, and 800 updates. Validation occurs every 50 updates and selects minimum score mean-squared error. Architecture and optimization settings are fixed across the sweep.

For the fitted function m, population enumeration evaluates

$$B_m = \mathbb{E}[mZ^\top + Zm^\top - mm^\top] = E_T - \mathbb{E}[(m - \mathbb{E}[Z \mid T])(m - \mathbb{E}[Z \mid T])^\top] \preceq E_T. \tag{66}$$

Thus a gap measures the fitted function's score-reconstruction error. The displayed bounds have no Monte Carlo evaluation uncertainty; the ranges span three training seeds and are not confidence intervals. All values are reported without clipping. Smooth functions with different capacity or training procedures are outside this empirical comparison.

Quadratic-rate and resolution checks. A separate population sweep evaluates the feature $(1, Q_\lambda)$ at $\lambda \in \{32, 64, 128, 256, 512, 1024\}$ and $p_1 \in \{.2, .35, .5\}$. At $\lambda = 1024$, the values of $\lambda L_{(1,Q_\lambda)}$ range from .50018 to .50073, approaching the $1/2$ limit in Theorem 4. For $K = 3$ and $p = (.2, .3, .5)$, $\lambda L_{(1,Q_\lambda)} = 1.00526$ at $\lambda = 128$, close to its predicted limit one. An independent one-dimensional Poisson/binomial moment calculation matches all binary values to 1.15×10^{-13}. Dyadic partitions with $2 \log_2 \lambda$ bits recover at least 89.9% of canonical degree information at $\lambda = 32$, 97.0% at 64, and 99.1% for $\lambda \geq 256$ over the tested probabilities. This is a sufficient empirical resolution scale in these cells, not a universal threshold.

Table 2: Unified diagnostic at $\lambda = 256$, $p_1 = .35$. Smooth-probe values are population-evaluated achieved lower bounds; brackets span three training seeds. The partition uses 12 bits and population conditional scores; the composition column reports canonical information.

Output bits	Canonical degree	Partition degree	Smooth degree [range]	Composition
4	0.001225	0.001225	0.001078 [0.000641, 0.001094]	0.725188
6	0.002029	0.002029	0.001484 [0.000941, 0.001815]	0.973815
8	0.004661	0.004661	0.001335 [0.000819, 0.001629]	0.994670
10	0.024060	0.024060	0.001639 [0.000862, 0.001792]	0.996082
12	0.141360	0.141360	0.001652 [0.000876, 0.001805]	0.996539
16	0.832611	0.141360	0.001655 [0.000879, 0.001807]	0.999347
24	0.839257	0.141360	0.001656 [0.000879, 0.001807]	0.999366
exact	0.839257	0.141360	0.001656 [0.000879, 0.001807]	0.999366

Numerical controls. Enumeration truncates each independent Poisson marginal at an upper tail probability at most $10^{-14}/K$, normalizes the retained law, and centers its scores. Across the primary and arithmetic-sensitivity cells, the largest entrywise residual from the identity raw Fisher matrix is 4.46×10^{-13}. Tightening the tail cutoff to 10^{-16} in the primary sweep and two arithmetic control cells changes reported efficiencies by at most 3.09×10^{-13}. Conditional-score grouping and missing-score variance agree; all raw-information contractions, nested-partition inequalities, and fitted lower bounds pass at tolerance 10^{-8}. A finite-difference check uses fixed output groups under local degree and composition perturbations. A separate enumeration through a Poisson total and conditional binomial law reproduces the primary information, every partition's achieved bound, and the headline arithmetic controls within 3.01×10^{-13}. Format checks also verify the declared rounding semantics. These controls qualify the finite numerical calculations; population certification refers to the underlying conditional- expectation inequality.

G.2 Observed Coordinates and Arithmetic Change Retention

Exact proportions determine the omitted last coordinate because they sum to one. After rounding, observing that coordinate separately can provide extra information. We test this difference between reduced and full vectors across 13 model cells. For $K = 2$, the cells cross $\lambda \in \{64, 256, 1024\}$ with $p_1 \in \{.1, .35, .5\}$. For $K = 3$, they cross $\lambda \in \{16, 64\}$ with $p = (.2, .3, .5)$ and $(.05, .15, .8)$. Each cell uses float32, float16, and emulated bfloat16, and two pipelines:

$$T_d^{\text{out}} = \text{cast}_d(C/N), \qquad T_d^{\text{count}} = \text{cast}_d(\text{cast}_d(C)/\,\text{cast}_d(N))\,. \tag{67}$$

Both retain a separate empty symbol. The reduced observation contains the first $K - 1$ rounded coordinates; the full observation contains all K. bfloat16 uses ties-to-even rounding through float32, matching the earlier rounding rule. Format tests verify these conversion semantics. These pipelines do not model low-precision accumulation of individual messages.

Table 3: Expected-degree retention at $\lambda = 1024$, $p = (.35, .65)$. Output-only rounding and count-casting pipelines observe either the first coordinate or both rounded coordinates. Exact-ratio retention is .918490.

	Output rounding		Count casting	
Format	Scalar	Vector	Scalar	Vector
float32	0.918490	0.918490	0.918490	0.918490
float16	0.004592	0.007038	0.004592	0.007038
bfloat16	0.000902	0.001457	0.002331	0.027746

Table 4: Canonical directional Fisher efficiencies of the frozen quantized recursive mean. The headline grid is $\delta = 1/64$; controls are $1/32$ and $1/128$. Every leaf-location–intensity gap is positive.

δ	Depth	Intensity	Leaf location	Gap
$1/64$	2	0.016710	0.868914	0.852204
	3	0.004491	0.798501	0.794010
	4	0.000896	0.693854	0.692958
$1/32$	2	0.016448	0.866100	0.849652
	3	0.004242	0.775549	0.771307
	4	0.000635	0.573713	0.573078
$1/128$	2	0.016889	0.869332	0.852443
	3	0.004570	0.804291	0.799721
	4	0.000987	0.732976	0.731989

The full vector measurably increases information in several cells. At the headline point, float16 degree retention increases from .004592 to .007038 under either pipeline. For bfloat16, output-only retention increases from .000902 to .001457, whereas count casting increases it from .002331 to .027746. The format alone therefore does not specify the statistical channel. Count casting can distinguish proportional count vectors before normalization, so it need not be a coarsening of the exact ratio; only the output-rounding pipeline has that guaranteed contraction relation. Both pipelines obey contraction relative to the raw counts, and full-vector information dominates its reduced observation in every cell.

With three categories at $\lambda = 64$, full-vector bfloat16 output rounding retains .80055 and .84747 degree information at the two compositions, against exact-ratio .86733 and .87135. Finite-degree retention therefore also depends on category count and composition. These cells do not locate a universal precision threshold. The smallest composition-block efficiency across all tested full-vector arithmetic channels is .97446.

H Finite-Depth Verification for Recursive Mean

The Poisson-tree theorem describes a high-degree limit. To measure retention at finite depth without fitting a readout, we use a separate tree whose quantized outputs have finite support. Propagating each output probability and its derivatives then gives canonical Fisher information directly. Each internal vertex has $M = 16$ candidate children, independently present with $q(\lambda) = \lambda/16$. Leaves take values $X \in \{-1, +1\}$ with $\Pr_\theta(X = x) = (1 + x\theta)/2$. The fiducial point is $(\lambda_0, \theta_0) = (8, 0.2)$. A nonempty parent outputs the nearest-grid, ties-to-even quantization of its children's mean; an empty parent outputs zero. The quantizer is frozen under differentiation.

For each finite hidden value h, the recursion propagates its probability $p_r(h)$ and derivative $u_r(h) = \nabla p_r(h)$. Hence the representation Fisher matrix is computed directly as

$$F_r^H = \sum_{h:p_r(h)>0} \frac{u_r(h)u_r(h)^\top}{p_r(h)}. \tag{68}$$

Exact recursion over the complete finite support determines this calculation. Independently, for Bernoulli candidates with $d = \nabla q(\psi_0)$, the complete score obeys

$$S_r = -\frac{Md}{1-q} + \sum_{u \text{ present}} \left\{ \frac{d}{q(1-q)} + S_{r-1}^{(u)} \right\}, \tag{69}$$

which supplies both the raw Fisher recursion and the calibration-specific score-sum oracle.

All 12 depth–resolution cells satisfy the probability, derivative, Fisher-contraction, and oracle checks. Finite-difference Fisher error is at most 2.18×10^{-10}, and an independent 80-digit calculation reproduces the reported values. These results concern the declared finite recursive channel. Arbitrary trained deep GNNs remain outside their scope.

I FINITE-GRAPH INFORMATION BEYOND SMOOTH PROBES

To check whether a small smooth-probe estimate implies information loss, we compute a canonical lower bound from the aggregate's label-ratio coordinate. This coordinate is available in the diagnostic's label-bearing messages; it is not supplied to the ordinary node classifiers.

For an even, balanced two-class graph of size n, fix a root with label $+1$ and define

$$M_+ = n/2 - 1, \quad M_- = n/2, \qquad q_\pm = \frac{\lambda(1 \pm \rho)}{2M_\pm}. \tag{70}$$

Conditional on the labels, the incident counts satisfy $N_\pm \sim \mathrm{Bin}(M_\pm, q_\pm)$ independently. This is the exact finite-CSBM root-neighborhood marginal. Neighbor contexts obey $X_j \mid Y_j = s \sim \mathcal{N}(\mu s, 1)$, and every architecture receives the same parameter-independent message $x_j = (1, Y_j, X_j, Y_j X_j)$.

The fiducial environment is $(\rho_0, \mu_0) = (0.5, 0.75)$. For a count state $c = (n_+, n_-)$, let p_c be its exact product-binomial probability, $s_\lambda(c)$ its intensity score, and $g(c)$ the distinct empty symbol or the actual float32 value $(n_+ - n_-)/(n_+ + n_-)$. Direct grouping of the finite support gives

$$F^{\bar{Y}}_{\lambda\lambda} = \sum_r \frac{\left(\sum_{c:g(c)=r} p_c s_\lambda(c)\right)^2}{\sum_{c:g(c)=r} p_c}. \tag{71}$$

Because $\bar{Y}$ is measurable from the mean aggregate, $F^{A}_{\lambda\lambda} \geq F^{\bar{Y}}_{\lambda\lambda}$.

The neural protocol used $20{,}000/5{,}000$ architecture train/validation roots, $16{,}000/4{,}000$ nuisance fit/validation roots, and $25{,}000$ Fisher-evaluation roots per condition and seed. Its original checks verified raw-score simulation, sufficient controls, and internal numerical consistency. The neural protocol quantifies sampling variation around the fitted smooth regressor; mean-channel approximation error remains an unquantified source of downward bias. Consequently, those neural outputs are smooth-probe diagnostics; the exact grouping calculation supplies the canonical lower bound.

Table 5 reports the exact bounds and smooth-probe medians.

Table 5: Finite-CSBM decoder audit: exact label-ratio and smooth-probe degree retention.

n	λ	$F^{O}_{\lambda\lambda}$	$\eta_\lambda(\bar{Y})$	Smooth probe
512	16	0.065063	0.4532	0.0828
512	64	0.018603	0.6998	0.0210
2048	16	0.063118	0.4521	0.0892
2048	64	0.016264	0.6778	0.0213

J TRAINED-GNN PREDICTION UNDER DEGREE SHIFT

This section verifies the two predictive claims in Section 7: whether explicit degree improves frozen-model transfer across expected-degree shifts, and whether it protects predictions from perturbations of the mean. The first subsection gives the data-generating model, training design, estimand, and uncertainty procedure for Table 1; the second gives the paired intervention protocol for Table 6.

J.1 CSBM TRANSFER PROTOCOL AND UNCERTAINTY

The controlled information calculations use observable categories or known message laws. This experiment instead trains GNNs for node classification with latent labels and asks whether explicit degree improves their target predictions. Arithmetic precision is held fixed, so its loss contrasts test the predictive value of explicit degree in isolation. Every finite Bernoulli CSBM has exactly balanced labels and $n = 4{,}096$ nodes. For environment $\psi = (\lambda, \rho, \mu)$, same- and cross-class edges occur independently with probabilities

$$p_{\mathrm{same}} = \frac{\lambda(1+\rho)}{n-2}, \qquad p_{\mathrm{cross}} = \frac{\lambda(1-\rho)}{n}, \tag{72}$$

and node features satisfy $X_i = \mu Y_i e_1 + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, I_8)$. Models observe only features and graph structure. Thus ρ changes edge probabilities while μ changes class separation in the features, giving them distinct roles here. The source environment is $\psi_0 = (24, 0.60, 0.35)$.

The experiment uses eight independent training blocks. Within each block, all reported architectures share six source training graphs and two source validation graphs; each evaluation condition has eight independent target graphs. The target graphs are independent across blocks and conditions.

At both message-passing layers, the mean-plus-degree model concatenates the neighbor mean with source-standardized log degree,

$$z_i = \frac{\log(1 + d_i) - m_{\text{train}}}{s_{\text{train}}}, \tag{73}$$

where m_{train} and the unbiased s_{train} are computed from the six source training graphs and then fixed. Mean, mean plus degree, and sum use two width-64 layers. The first two differ by 1.36% in parameter count. All models use dropout 0.1, binary cross-entropy, and full-batch AdamW with learning rate 0.003, weight decay 10^{-4}, and gradient clipping at 5. Training runs for 50–300 epochs; early stopping uses only the source validation graphs, with patience 30 and minimum improvement 10^{-5}. Fitted models are frozen before target evaluation.

Fresh evaluation graphs use the source expected degree 24 and the two target degrees 12 and 48, holding $(\rho, \mu) = (.60, .35)$ fixed. Metrics are first averaged within graph. The primary contrast is mean loss minus mean-plus-degree loss. A paired nested bootstrap resamples training blocks and then target graphs within blocks, using 10,000 replicates. The shifted-degree ranges are the prespecified 1.25th–98.75th percentiles.

Table 1 reports the complete three-environment comparison. At $\lambda = 12$, all eight blocks favor explicit degree and the loss advantage is 0.14061 nats per node. Its accuracy interval includes zero. At $\lambda = 48$, the loss-advantage interval includes zero, while sum has the lowest loss. These comparisons establish a low-degree benefit without a degree-independent advantage or a causal link to the exact arithmetic channel.

J.2 Precision and Noise Interventions on Frozen CSBM Predictors

This follow-up tests predictive sensitivity to aggregate perturbations in the previously trained classifiers. Its protocol was fixed after the original transfer results were known and before generating the new intervention targets. The eight training blocks and their 16 frozen mean and mean-plus-degree models are reused, with no retraining or target-based selection.

Graphs and pairing. Each training block receives eight new independent $4{,}096$-node graphs at each of $\lambda \in \{12, 24, 48\}$, with $(\rho, \mu) = (.60, .35)$ and the generator in Appendix J.1. This gives 192 graphs. The source, low-degree, and high-degree conditions use independent random streams, separate from all original training, validation, and target pools. All models and interventions within a block share each graph. The original transfer table is preserved; the new contrasts compare perturbed and native predictions on the new graphs only.

Intervention channels. The primary channel adds Gaussian noise to the eight-dimensional first-layer neighbor mean, before concatenation with the self state and explicit degree. Writing that mean as a_i, the perturbed coordinates are

$$\widetilde{a}_i = a_i + c\, 24^{-3/4} D_0 Z_i, \qquad D_0 = \operatorname{diag}\Big(\sqrt{1 + .35^2}, 1, \ldots, 1\Big), \quad Z_i \sim \mathcal{N}_8(0, I_8). \tag{74}$$

D_0 contains the unconditional source feature standard deviations. It is common across architectures and frozen across target environments. We use $c \in \{.25, .5, 1\}$, with $c = 1$ primary. Its noise standard deviation is about .0922 in the seven nonsignal coordinates. This scale is motivated by the theoretical window, but the continuous, correlated graph messages and finite degrees do not instantiate the discrete Poisson limit experiment.

A uniform-noise control replaces each Gaussian coordinate by an independent $\mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ variable at $c = 1$, giving the same variance. Four noise realizations per graph are paired across architectures; the Gaussian scale sweep shares its standard-normal draws. Noise is omitted at

empty neighborhoods, whose means remain zero. Explicit degree, self-state inputs, all weights, and subsequent arithmetic remain unchanged.

Four deterministic channels round completed means to float16 or bfloat16 and back to float32, either at the first layer or at both layers. These channels change output rounding only; the accumulation and count-division arithmetic follow Section 5. Together with native float32 and the four random channels, there are nine settings and 8,064 graph–architecture–channel–replicate evaluations.

Estimands and frozen decision. For architecture a, let $\Delta\mathcal{L}_a$ be its perturbed minus native log loss, averaged over noise realizations within each graph. The protection contrast is

$$D = \Delta\mathcal{L}_{\text{mean}} - \Delta\mathcal{L}_{\text{degree}}. \tag{75}$$

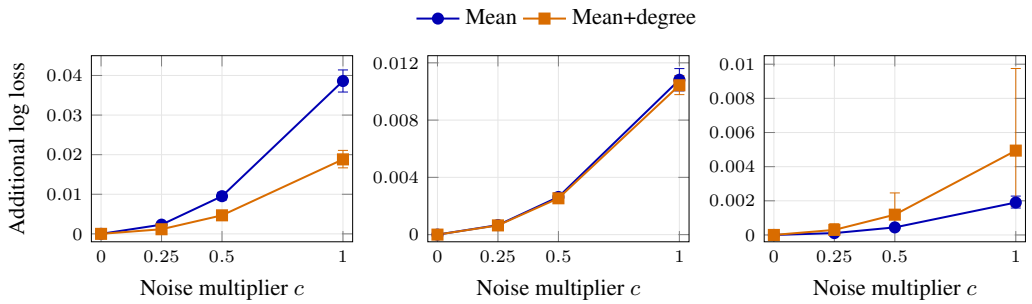
Positive D means less additional harm in the degree model. The primary conjunction requires both $\Delta\mathcal{L}_{\text{mean}} > .001$ and $D > .001$ at each of the two degree shifts under Gaussian $c = 1$. Each of these four tests uses a one-sided 98.75% lower bootstrap bound, allocating $.05/4$ to each test. Paired nested resampling selects training blocks and then graphs within each selected block, with 10,000 replicates. The paired noise draws are averaged within graph, not treated as independent observations. Reported primary ranges use the 1.25th–98.75th percentiles.

The source, remaining channels, accuracy, and Brier results are descriptive and use ordinary 95% intervals. A bootstrap over block means alone supplies a sensitivity check. Coverage remains approximate with eight training blocks.

Table 6: Additional log loss from the primary Gaussian intervention, in nats per node. The difference is mean's additional loss minus mean-plus-degree's; positive values indicate less harm in the degree model. Brackets use paired 1.25th–98.75th bootstrap percentiles.

λ	Mean	Mean+degree	Difference [interval]
12	0.03860	0.01881	$0.01979\ [0.01732, 0.02205]$
48	0.00190	0.00494	$-0.00304\ [-0.00845, 0.00015]$

Degree protects predictions under the low-degree shift. The primary mean-harm bounds exceed .001 at both shifts. Low-degree protection also passes, with $D = .019785\ [.017325, .022052]$ and positive point estimates in all eight blocks. High-degree protection is $-.003039\ [-.008446, .000150]$. Thus the experiment supports protection at low degree but not at high degree. Block-only intervals give the same qualitative conclusion. Figure 6 shows the full Gaussian sweep, including the matched uniform-noise control.



(a) $\lambda = 12$ **(b)** $\lambda = 24$ (source) **(c)** $\lambda = 48$

Figure 6: Additional log loss under first-layer Gaussian noise in frozen classifiers. The panels use different vertical scales and losses are in nats per node. Error bars are paired nested-bootstrap intervals: 97.5% for the primary mean-harm tests at $c = 1$ and shifted degrees, and 95% otherwise. Degree reduces additional low-degree loss; high-degree protection remains inconclusive.

Across both architectures and all rounding settings, the largest absolute average log-loss change is below 4.24×10^{-6} nats per node. Every associated descriptive 95% interval lies within $\pm 2.46 \times 10^{-5}$ nats. Uniform noise yields low-degree protection $.020723\ [.018433, .023057]$ and high-degree protection $-.003100\ [-.007645, .000072]$. The primary Gaussian intervention reduces low-degree

accuracy by .646 percentage points for mean and .615 points for mean-plus-degree; the corresponding high-degree reductions are .0615 and .1905 points.

These findings establish predictive sensitivity conditional on the frozen models. The two architectures have separately learned continuous representations; their loss differences neither identify erased count information nor estimate canonical Fisher information. The observed rounding insensitivity therefore bounds the practical interpretation of the arithmetic diagnostic in these classifiers, while attribution to its rational-denominator mechanism remains unresolved.