THE STATISTICAL VALUE OF RESOLUTION IN HEAVY-TAILED DEPENDENT RESIDUAL FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured predictors often return a residual field with many coordinates per observation, but higher resolution need not provide proportionally more statistical evidence. We ask a basic question: *under heavy tails and structured dependence, when does an additional residual coordinate actually reduce uncertainty?* For a valid dependency graph on the residual field and a finite centered q-th moment, $q \in (1, 2]$, we derive a profile-aware evidence count $N_{\mathrm{eff},q}(\mathcal{C}) = \left(nm/\sum_\ell N_\ell^{1/q}\right)^{q/(q-1)}$ that controls generalized Catoni risk estimation. Fractional independent-set covers give a graph-level refinement that can increase the evidence count, subject to support-size and confidence costs, while constant-probability lower bounds on canonical coordinate and space–time block families show that the dependence penalty is unavoidable up to logarithmic factors. We expose the graph assumption through safe supergraphs, independent-pilot transfer, conservative complete-graph fallbacks, and approximate-independence extensions. Experiments support the predicted finite-q scaling trend, show severe undercoverage when the dependence/color factor is undercounted, and reveal a $10.49\times$ average uncertainty inflation on bounded CIFAR-10 residual maps. Decision audits further show that the value of dependence information is target-specific: it can change uncertainty dramatically without changing point rankings, and frozen dependence-aware risk-control rules can still abstain at small sample sizes. The resulting framework separates observed resolution from decision-relevant evidence.

1 INTRODUCTION

A modern learning system rarely produces one scalar error per example. Image restoration yields residual maps, medical reconstruction yields voxel-wise discrepancies, dense prediction yields per-location losses, and multichannel sensors return coordinate-indexed error vectors. If n observations each contain m residual coordinates, the loss table has nm entries. Treating those entries as nm independent observations can be dramatically overconfident; collapsing each field to one scalar, on the other hand, can discard genuine information when coordinates are only partially dependent.

Heavy tails create a second difficulty. Robust mean estimators and robust empirical risk minimization can recover high-probability guarantees under weak moments (Catoni, 2012; Audibert and Catoni, 2011; Brownlees et al., 2015; Hsu and Sabato, 2014; 2016; Lugosi and Mendelson, 2019; Chen et al., 2021; Bhatt et al., 2022), but robustification of magnitudes does not by itself decide how many correlated residual coordinates may be counted as distinct evidence. Dependency-graph learning instead quantifies dependence among random variables or examples (Janson, 2004; Ralaivola et al., 2009; Zhang et al., 2019; Zhang and Amini, 2024); existing heavy-tailed dependent-data theory primarily addresses temporal or sample-level dependence (Roy et al., 2021; Kengne and Wade, 2025). We study the intersection: heavy-tailed losses whose *within-observation residual coordinates* are dependent.

Our scientific question is therefore:

What determines the statistical value of resolution when residual coordinates are heavy-tailed and dependent?

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 answer should distinguish three regimes: essentially independent new coordinates should make resolution valuable; dependent blocks should contribute only one unit per block; and complete within-observation dependence should make increasing m asymptotically useless.

We formalize this using a dependency graph on the nm residual entries. A valid coloring $\mathcal{C} = \{D_\ell\}_{\ell=1}^{L}$ with class sizes N_ℓ induces the evidence count

$$N_{\mathrm{eff},q}(\mathcal{C}) := \left(\frac{nm}{\sum_{\ell=1}^{L} N_\ell^{1/q}} \right)^{q/(q-1)}, \qquad q \in (1, 2]. \tag{1}$$

It is a *scale-equivalent evidence count*: the number of independent residual entries that would yield the same finite-q stochastic-error scale, up to constants and logarithmic factors. Balanced color classes recover nm/L; the exact profile can be sharper. The question "when does resolution count?" is therefore equivalent to asking how quickly $N_{\mathrm{eff},q}$ grows with m. Fractional independent-set covers provide a further refinement.

A central concern with any graph-based guarantee is where the graph comes from. We therefore treat graph information as an explicit interface rather than a hidden oracle. Valid structure can be supplied by a controlled data-generating mechanism or acquisition design; any externally certified supergraph remains safe; and an independent pilot can transfer a supergraph with an explicit failure-probability price. If observations are independent but no within-observation structure is defensible, the complete within-observation coordinate graph gives the conservative fallback $N_{\mathrm{eff},q} = n$; if dependence on both axes is unknown, the fully complete space–time graph gives the trivial count $N_{\mathrm{eff},q} = 1$. Empirical correlation alone is *not* used to certify independence. Table 1 separates theorem-valid structure from descriptive diagnostics on natural images.

Contributions.

1. **Evidence counting under heavy tails.** We prove a generalized-Catoni excess-risk bound for valid space–time dependency graphs and introduce the profile functional equation 1. We also derive an exact fractional independent-set refinement.
2. **Sharp canonical resolution laws.** Coordinate-block and joint space–time constant-probability lower bounds match the polynomial dependence on (n, m, r, g) of the upper bounds at fixed confidence, up to logarithmic factors on the stated canonical families. They yield phase laws that separate useful resolution, partial redundancy, and complete redundancy.
3. **Assumption-aware empirical validation.** Pre-specified experiments support the finite-q scaling trend and quantify the cost of evidence-count misspecification; bounded CIFAR-10 residuals exhibit strong dependence, high kurtosis, and large bootstrap uncertainty inflation. Decision audits show an equally important boundary: better uncertainty accounting need not improve every ranking or deployment rule.

2 Related work and positioning

Heavy-tailed robust learning. Catoni estimators yield exponential-style deviations under weak moments, with robust ERM and median-of-means extensions to learning (Catoni, 2012; Audibert and Catoni, 2011; Brownlees et al., 2015; Hsu and Sabato, 2014; 2016; Lugosi and Mendelson, 2019). Chen et al. (2021) treat finite q-th moments for $q \in (1, 2)$, while Bhatt et al. (2022) address infinite-variance adaptation; truncation-based heavy-tailed learning includes Zhang and Zhou (2018); Xu et al. (2023). Our novelty is not scalar robust mean estimation, but how its stochastic scale changes when residual *coordinates* form a dependency graph.

Dependent data and graph covers. Dependency-graph bounds decompose variables into independent sets, with fractional covers used in graph-dependent PAC-Bayes analysis (Janson, 2004; Ralaivola et al., 2009; Zhang et al., 2019; Zhang and Amini, 2024). Heavy-tailed learning under weak temporal dependence has also been studied through mixing (Roy et al., 2021; Kengne and Wade, 2025). Those theories principally count examples or time indices; here the graph spans the full $n \times m$ residual field, making coordinates per observation part of sample complexity.

Structured errors versus statistical evidence. Low-rank, sparse, and decomposable models describe the *shape* of structured errors or parameters (Candès et al., 2011; Negahban et al., 2012). Our

Table 1: How graph information enters the guarantee. A graph is never inferred from pairwise correlation inside the theorem.

Source of structure	Guarantee	Price
Designed blocks / acquisition	Valid by construction if nonadjacent residual collections are independent	Potentially tight count
Certified supergraph	Validity is preserved	Smaller $N_{\mathrm{eff},q}$
Independent pilot	Conditional transfer when pilot returns a valid supergraph	Failure probability $+\delta_0$
No trusted coordinate structure (independent observations)	Complete within-observation coordinate graph	$N_{\mathrm{eff},q} = n$
No trusted structure on either axis	Complete space–time graph	$N_{\mathrm{eff},q} = 1$
Approximate class independence	Total-variation transfer (Appendix D)	Failure $+\sum \epsilon$

complementary question is how many residual coordinates constitute statistically distinct evidence for a risk target; the graph is therefore a dependence assumption for uncertainty, not a representation model of the residual itself.

Positioning. Treating all coordinates as independent can replace n by nm despite common modes, while collapsing each observation to one scalar discards partial independence. We introduce an intermediate profile-sensitive evidence unit compatible with heavy-tailed robust estimation, conservative fallbacks, and canonical fixed-confidence lower bounds; Table 1 summarizes how graph information enters the guarantee.

3 Residual fields, graph structure, and robust estimation

3.1 Target Risk and Valid Graphs

Residual field. Let $Z_1, \dots, Z_n$ denote observations, not necessarily identically distributed, and let $\theta \in \Theta$. For coordinate $s \in [m]$, define the nonnegative residual loss

$$X_{t,s}(\theta) = \ell_\theta(Z_t, s).$$

We refer to the indexed collection

$$\big(X_{t,s}(\theta)\big)_{(t,s)\in[n]\times[m]}$$

as the corresponding $n \times m$ *residual field*. Throughout, we allow its entries to exhibit structured dependence across coordinates, observations, or both. Any dependence information used by our guarantees is encoded through a valid dependency graph $G_{n,m}$, defined below.

We study the finite-sample average expected risk

$$R_{n,m}(\theta) = \frac{1}{nm}\sum_{t=1}^{n}\sum_{s=1}^{m}\mathbb{E}X_{t,s}(\theta), \qquad R^\star_{n,m} = \inf_{\theta\in\Theta} R_{n,m}(\theta). \tag{2}$$

Under i.i.d. or stationary observations, this is the usual average-coordinate population risk. For general non-identically distributed observations, $R_{n,m}$ is instead a finite-design average of marginal expected risks over the indexed field; without additional sampling or stationarity assumptions, we do not identify it with an unseen-time deployment risk.

Definition 1 (Valid space–time dependency graph)**.** A graph $G_{n,m}$ on $[n] \times [m]$ is valid if, for every fixed θ, any two disjoint vertex sets with no edge between them index independent residual collections. A valid coloring $\mathcal{C} = \{D_1, \dots, D_L\}$ consists of nonempty color classes, with $N_\ell = |D_\ell| \geq 1$.

By concavity, equation 1 gives $N_{\mathrm{eff},q}(\mathcal{C}) \geq nm/L$, with equality for balanced classes; the exact count retains the class-size profile. Independent residual entries yield $N_{\mathrm{eff},q} = nm$, while independent observations with a complete within-observation coordinate graph give $N_{\mathrm{eff},q} = n$.

A *certified supergraph* is externally guaranteed—by the data-generating mechanism, acquisition design, or a separate certification procedure—to contain a valid graph; empirical correlation alone is not a certificate. Supergraphs remain valid, with a smaller count. An *independent* pilot that returns such a supergraph with probability at least $1 - \delta_0$ transfers the analysis guarantee with failure probability increased by at most δ_0 (Appendix B.1); otherwise the candidate set, coloring/cover, q, and moment upper bound are fixed before residual inspection, and same-sample adaptation requires uniformization or sample splitting.

3.2 Finite-q generalized Catoni estimation

Set $a_q = (q-1)/q$. Following Chen et al. (2021), use the odd strictly increasing score

$$\psi_q(x) = \operatorname{sign}(x) \log\left(1 + |x| + \frac{|x|^q}{q}\right), \tag{3}$$

which satisfies, for every $x \in \mathbb{R}$, the generalized Catoni envelope

$$-\log\left(1 - x + \frac{|x|^q}{q}\right) \le \psi_q(x) \le \log\left(1 + x + \frac{|x|^q}{q}\right),$$

used in Chen et al. (2021). Because ψ_q is continuous and strictly increasing, the estimating score is continuous and strictly decreasing in the location parameter and changes sign between $-\infty$ and $+\infty$; hence each estimating equation below has a unique root. For color class D_ℓ, define

$$R_\ell(\theta) = \frac{1}{N_\ell} \sum_{(t,s)\in D_\ell} \mathbb{E} X_{t,s}(\theta)$$

and assume the centered-moment condition

$$\sup_{\theta\in\Theta} \max_{\ell \le L} \frac{1}{N_\ell} \sum_{(t,s)\in D_\ell} \mathbb{E}\big|X_{t,s}(\theta) - R_\ell(\theta)\big|^q \le V_q. \tag{4}$$

This is a uniform finite-q analogue of Catoni's finite-variance scale condition. Equal means inside a color class are *not* required: Appendix A proves the class-wise deviation bound for independent, non-identically distributed variables around their average expectation, using only the averaged mean identity and equation 4.

Let $\Theta_\eta \subset \Theta$ be a finite approximation with $M_\eta = |\Theta_\eta|$, and define the approximation gap $\omega(\eta) := \inf_{\theta\in\Theta_\eta} R_{n,m}(\theta) - R^*_{n,m} \ge 0$. For example, if (Θ, d) is compact, $R_{n,m}$ is L_R-Lipschitz, and Θ_η is an η-net, then $\omega(\eta) \le L_R\eta$. Write $\Lambda = \log(2LM_\eta/\delta)$. For constants depending only on q, choose

$$\alpha_\ell = c_{\alpha,q}\left(\frac{\Lambda}{N_\ell V_q}\right)^{1/q}, \qquad \Lambda \le c_{\mathrm{conf},q} \min_\ell N_\ell. \tag{5}$$

Let $\widehat{R}_\ell(\theta)$ be any root of $\sum_{(t,s)\in D_\ell} \psi_q\{\alpha_\ell(X_{t,s}(\theta) - z)\} = 0$, and aggregate

$$\widehat{R}(\theta) = \sum_{\ell=1}^{L} \frac{N_\ell}{nm} \widehat{R}_\ell(\theta), \qquad \widehat{\theta}_\eta \in \arg\min_{\theta\in\Theta_\eta} \widehat{R}(\theta). \tag{6}$$

A highly unbalanced coloring can render equation 5 infeasible through its smallest active classes. For fixed (L, M_η, δ), feasibility requires $\min_\ell N_\ell \ge \Lambda / c_{\mathrm{conf},q}$. If this fails, one must use another *certified* graph/cover representation with adequate active-set sizes, fall back to a justified observation-level representation, or use a different robust estimator; arbitrary merging of color classes need not preserve independence.

Proposition 1 (Conservative moment-scale perturbation)**.** *Suppose the estimator is tuned with $\bar{V}_q$ satisfying $V_q \le \bar{V}_q \le \kappa V_q$ for some $\kappa \ge 1$. Every upper bound below remains valid with $\bar{V}_q$ in place of V_q, and its stochastic term is inflated by at most $\kappa^{1/q}$ relative to oracle tuning. Underestimation of V_q is not covered.*

This proposition separates conservative moment-scale calibration from dependence-count misspecification: the former widens a valid certificate predictably, whereas underestimating the dependence cost can destroy nominal coverage. Section 6 stress-tests the conservative regime.

4 EVIDENCE COUNTS AND RISK GUARANTEES

Theorem 1 (Profile-aware master bound). *Assume $G_{n,m}$ is valid, equation 4–equation 5 hold for a coloring $\mathcal{C}$, and $0 < \delta < 1/2$. Then, for a constant C_q depending only on q, with probability at least $1 - \delta$,*

$$R_{n,m}(\widehat{\theta}_\eta) - R^*_{n,m} \le \omega(\eta) + C_q V_q^{1/q} \left(\frac{\log(2LM_\eta/\delta)}{N_{\mathrm{eff},q}(\mathcal{C})} \right)^{a_q}. \tag{7}$$

In particular, $N_{\mathrm{eff},q}(\mathcal{C}) \ge nm/L$, recovering the balanced-coloring rate as a corollary.

The proof applies the finite-q Catoni deviation bound inside each independent color class, weights the deviations by $N_\ell/(nm)$, and retains the exact class-size profile before the optional concavity step; see Appendix B.

4.1 FRACTIONAL INDEPENDENT-SET REFINEMENT

Ordinary coloring need not be the sharpest graph description.

Definition 2 (Exact fractional independent-set cover). For a finite dependency graph $G = (V, E)$ with $|V| = nm$, an exact fractional cover is $\lambda = \{(I_j, \lambda_j)\}_{j=1}^S$ with nonempty independent sets I_j, positive weights $\lambda_j > 0$, and $\sum_{j: v \in I_j} \lambda_j = 1$ for every $v \in V$; zero-weight sets are omitted, so S is the active support size. Define

$$T_q(\lambda) = \sum_{j=1}^S \lambda_j |I_j|^{1/q}, \qquad N^{\mathrm{frac}}_{\mathrm{eff},q}(\lambda) = \left(\frac{nm}{T_q(\lambda)} \right)^{q/(q-1)}. \tag{8}$$

For each I_j, let $R_j(\theta) = |I_j|^{-1} \sum_{v \in I_j} \mathbb{E} X_v(\theta)$, assume the cover-wise analogue of equation 4 with the same bound V_q, compute a Catoni mean $\widehat{R}_j(\theta)$, and set

$$\widehat{R}_\lambda(\theta) = \frac{1}{nm} \sum_j \lambda_j |I_j| \widehat{R}_j(\theta), \qquad \widehat{\theta}_{\eta,\lambda} \in \arg\min_{\theta \in \Theta_\eta} \widehat{R}_\lambda(\theta).$$

Theorem 2 (Fractional refinement). *Let $\Lambda_\lambda = \log(2SM_\eta/\delta)$ and assume explicitly that $\Lambda_\lambda \le c_{\mathrm{conf},q} \min_{1 \le j \le S} |I_j|$. Then, with probability at least $1 - \delta$,*

$$R_{n,m}(\widehat{\theta}_{\eta,\lambda}) - R^*_{n,m} \le \omega(\eta) + C_q V_q^{1/q} \left(\frac{\Lambda_\lambda}{N^{\mathrm{frac}}_{\mathrm{eff},q}(\lambda)} \right)^{a_q}. \tag{9}$$

Every integral coloring is a feasible exact cover. The full bound, however, also pays through the support size S and the smallest active independent set, so one should not maximize equation 8 in isolation. Within a pre-specified certified family of covers, a natural choice minimizes the full certificate $\{\Lambda_\lambda / N^{\mathrm{frac}}_{\mathrm{eff},q}(\lambda)\}^{a_q}$ subject to the active-set confidence condition; optimizing globally over all covers can be combinatorial.

On a disjoint union of cliques of size r, the optimized fractional count is exactly nm/r, so the refinement does not alter the canonical block rate. On a C_5 coordinate graph with $q = 2$, every proper 3-coloring has profile $(2, 2, 1)$ and count $\left(5/(2\sqrt{2}+1)\right)^2 \approx 1.706$, while the five maximum independent pairs with weight $1/2$ give fractional count 2; Appendix C proves both statements.

4.2 TWO DEPENDENCE AXES AND FINITE MODEL SELECTION

If a sample graph on $[n]$ has color classes $A_1, \ldots, A_{J_n}$ and a coordinate graph on $[m]$ has color classes $C_1, \ldots, C_{K_m}$, their product coloring is valid whenever full residual vectors are independent across sample color classes and coordinates are independent within coordinate color classes. Its evidence count factorizes into the product of the corresponding one-axis evidence counts, and is at least $nm/(J_n K_m)$.

For a finite candidate set $\Theta_M = \{\theta_1, \ldots, \theta_M\}$, candidate-specific valid graphs are allowed. Let $\mathcal{C}_j$ be candidate j's valid coloring, with color-class sizes $N_{j\ell}$, and define the uniform evidence-count lower

bound $N_{\min} := \min_{1\le j\le M} N_{\mathrm{eff},q}(\mathcal{C}_j)$. If each candidate has at most $L_{\max}$ classes, the common moment bound V_q holds, and $\log(2L_{\max}M/\delta) \le c_{\mathrm{conf},q}\min_{j,\ell} N_{j\ell}$, let $\widehat{R}^{(j)}$ be its aggregated robust-risk estimate, choose $\widehat{j} \in \arg\min_{1\le j\le M} \widehat{R}^{(j)}$, and set $\widehat{\theta} = \theta_{\widehat{j}}$. Then simultaneous evaluation and selection obey

$$R_{n,m}(\widehat{\theta}) - \min_j R_{n,m}(\theta_j) \le C_q V_q^{1/q}\left(\frac{\log(2L_{\max}M/\delta)}{N_{\min}}\right)^{a_q} \tag{10}$$

with probability at least $1-\delta$. For a finite candidate set no parameter-space net is enumerated and $\omega(\eta) = 0$. The guarantee in equation 10 does *not* imply that dependence-aware and naive point rankings must differ; Section 6.2 gives an exact common-mode cancellation example.

5 When does resolution help? Lower bounds and phase laws

Upper bounds are meaningful only if the dependence penalty cannot be removed by a better estimator.

Theorem 3 (Coordinate-block lower bound)**.** *Fix $q \in (1,2]$ and an integer r dividing m. Let $\mathcal{P}_r$ be the family of laws with n independent observations whose m coordinates are partitioned into blocks of size r, all coordinates in one block being copies of one latent scalar and different blocks being independent. There exist $c_q > 0$, depending only on q, and a two-action residual-risk subfamily satisfying equation 4 such that*

$$\inf_{\widetilde{\theta}} \sup_{P\in\mathcal{P}_r} \mathbb{P}_P\Big(R_{n,m}(\widetilde{\theta}) - R^*_{n,m} \ge c_q V_q^{1/q}\left(\frac{r}{nm}\right)^{a_q}\Big) \ge \frac{3}{8}. \tag{11}$$

Here the infimum is over all estimators $\widetilde{\theta}$ measurable with respect to the observed residual field. Consequently, after possibly decreasing the q-dependent constant,

$$\inf_{\widetilde{\theta}} \sup_{P\in\mathcal{P}_r} \mathbb{E}_P\Big[R_{n,m}(\widetilde{\theta}) - R^*_{n,m}\Big] \ge c_q V_q^{1/q}\left(\frac{r}{nm}\right)^{a_q}.$$

The construction has $N = nm/r$ independent latent variables. Two alternatives differ only on a rare event of probability $p = 1/(4N)$ and induce opposite optimal actions. Le Cam's two-point inequality implies that, for every estimator, one of the two alternatives has testing error at least $3/8$. On that error event the excess risk equals the action gap, which is of order $V_q^{1/q}(r/(nm))^{a_q}$ after saturating the q-moment constraint. Thus Theorem 1's high-probability upper bound and Theorem 3's fixed-probability lower threshold have the same polynomial dependence on (n,m,r) up to logarithmic factors. We do not claim matching dependence on the confidence level δ.

Theorem 4 (Joint space–time lower bound)**.** *Fix $q \in (1,2]$ and integers $r \mid m$ and $g \mid n$. Let $\mathcal{P}_{r,g}$ be the corresponding family in which coordinates are partitioned into blocks of size r and observations into groups of size g, with one latent heavy-tailed scalar shared inside each space–time block and different blocks independent. There exists a constant $c_q > 0$, depending only on q, such that*

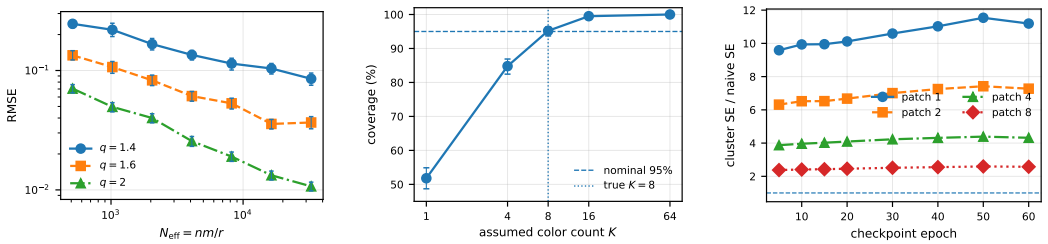
$$\inf_{\widetilde{\theta}} \sup_{P\in\mathcal{P}_{r,g}} \mathbb{P}_P\Big(R_{n,m}(\widetilde{\theta}) - R^*_{n,m} \ge c_q V_q^{1/q}\left(\frac{rg}{nm}\right)^{a_q}\Big) \ge \frac{3}{8}.$$

Here the infimum is over all estimators based on the observed residual field. Consequently, after possibly decreasing the q-dependent constant,

$$\inf_{\widetilde{\theta}} \sup_{P\in\mathcal{P}_{r,g}} \mathbb{E}_P\Big[R_{n,m}(\widetilde{\theta}) - R^*_{n,m}\Big] \ge c_q V_q^{1/q}\left(\frac{rg}{nm}\right)^{a_q}.$$

We deliberately claim matching only on these canonical families, not a complete minimax invariant for arbitrary graphs.

Proposition 2 (Canonical fixed-confidence matching)**.** *On the coordinate-block family, the optimized fractional count equals nm/r. For any fixed confidence level $\delta \in (0, 3/8)$ satisfying the confidence condition of Theorem 1, the theorem's stochastic upper term and the fixed-probability lower threshold in Theorem 3 have the same polynomial dependence on (n,m,r) up to constants and logarithmic factors. On the joint space–time block family, the same statement holds with effective block cost rg. We do not claim matching dependence on δ.*



(a) finite-q scaling (b) count misspecification (c) CIFAR uncertainty gap

Figure 1: Evidence counting from theorem to data. (a) RMSE decreases with $N_{\text{eff},q} = nm/r$ at slopes close to the finite-q prediction. (b) Undercounting the dependence/color factor causes severe undercoverage, while conservative factors over-cover. (c) On CIFAR-10 autoencoder residuals, image-cluster bootstrap uncertainty remains far above naive pixel-bootstrap uncertainty even after patch aggregation. Marker and line styles make all curves distinguishable without color.

The proposition turns the evidence count into an explicit resolution law rather than only an upper-bound device.

Corollary 1 (Coordinate-resolution phase law)**.** *On the coordinate-block family, along admissible integer block sizes $r_m \asymp m^\gamma$ with $\gamma \in [0, 1]$, and for any fixed $\delta \in (0, 3/8)$, if V_q and model-complexity/confidence terms introduce no additional polynomial dependence on m, the dependence-driven fixed-confidence minimax stochastic scale is*

$$V_q^{1/q} n^{-a_q} m^{-(1-\gamma)a_q} \tag{12}$$

up to constants and logarithms. Hence $\gamma = 0$ corresponds to essentially independent new coordinates, $0 < \gamma < 1$ to partial redundancy, and $\gamma = 1$ to no asymptotic gain from increasing spatial resolution.

Corollary 2 (Space–time phase law)**.** *On the joint space–time block family, if in addition $g_n \asymp n^\xi$, $\xi \in [0, 1]$, and for any fixed $\delta \in (0, 3/8)$, with the same control of moment and complexity terms, the dependence-driven fixed-confidence canonical minimax rate becomes*

$$V_q^{1/q} n^{-(1-\xi)a_q} m^{-(1-\gamma)a_q}$$

up to constants and logarithms. Spatial and temporal redundancy therefore enter multiplicatively in the canonical family.

As a scalar sanity check, set $m = 1$ and $q = 2$. If Θ is contained in a bounded d-dimensional parameter set of radius B and an η-net satisfies $\log M_\eta \lesssim d \log(B/\eta)$, balancing approximation and stochastic error recovers the familiar heavy-tailed scale $\widetilde{O}(\sqrt{d/n})$, consistent with classical robust regression results such as Zhang and Zhou (2018). Thus the new behavior appears only when the residual coordinate axis contributes additional, potentially dependent evidence.

6 Empirical evaluation: mechanism, real residuals, and decision boundaries

All primary settings, seeds, checkpoints, metrics, and stop rules were pre-specified and frozen; negative outcomes are retained. Appendix F gives full protocols, seeds, implementation details, and additional results.

6.1 From the rate to real residual fields

Figure 1 follows the theory-to-data chain.

Finite-q scaling. For block-correlated $|t_\nu|$ residuals with $(\nu, q) = (1.5, 1.4), (1.8, 1.6), (2.5, 2)$, log–log RMSE slopes versus $N_{\text{eff},q} = nm/r$ are -0.260 (95% CI $[-0.289, -0.232]$), -0.336 ($[-0.364, -0.307]$), and -0.466 ($[-0.489, -0.442]$), versus theoretical $-0.286, -0.375, -0.500$, with $R^2 = 0.986, 0.974, 0.995$. These support the predicted trend rather than exact finite-sample equality; for $q = 1.6, 2$ the theoretical exponents lie just outside the reported intervals.

Table 2: Theory-to-decision evidence chain; each frozen test supports a different link.

Question	Test	Supported conclusion
Heavy-tail rate?	Block-correlated $\lvert t_\nu \rvert$	Slopes track the predicted $-a_q$ trend as nm/r varies.
Dependence undercounted?	Held-out K stress test	Coverage rises from 51.8% at $K = 1$ to 95.2% at the true $K = 8$; undercounting is overconfident.
Real residual maps?	Eight CIFAR checkpoints	Cluster SE is 10.49× naive pixel SE on average; descriptive dependence, not graph certification.
Ranking changed?	Controlled selection	Not necessarily: common-mode cancellation leaves Top-1 almost unchanged.
Empirical deployment?	Target-aligned risk control	Not at $n \leq 128$: reference-feasible checkpoints exist, yet dependence-aware margins abstain under both the pre-specified and $3/8$-exponent-aligned decision rules.

Evidence-count misspecification. With $n = 128, m = 64, r = 8$, nominal 95% coverage is $51.8\%, 84.8\%, 95.2\%, 99.5\%, 100.0\%$ for prespecified $K = 1, 4, 8, 16, 64$. Thus $K < 8$ undercounts dependence, $K = 8$ matches it, and $K > 8$ is conservative. Here $q = 2$, so $a_q = 1/2$ and $N_{\text{eff},q} = nm/K$ in the balanced block representation; calibrating at $K = 8$ and scaling held-out radii by $\sqrt{K/8}$ therefore matches the evidence-count exponent. This tests count misspecification, not estimator superiority or graph discovery.

CIFAR-10 residual fields. Across eight fixed autoencoder checkpoints, bounded absolute residuals have mean horizontal/vertical correlations $0.465/0.469$ and kurtosis 11.76. Image-cluster bootstrap SE averages $10.49\times$ naive pixel-bootstrap SE at pixel resolution (range $[9.59, 11.54]\times$), remaining $6.87\times, 4.15\times, 2.49\times$ for patch sizes $2, 4, 8$. Squared residuals give $8.87\times$ at pixel level and $2.47\times$ at 8×8. Because both losses are bounded, these diagnose dependence, tail-heaviness, and uncertainty inflation—not infinite variance or a theorem-valid natural-image graph.

6.2 The inferential target matters

Dependence need not affect every downstream target. With candidate-shared additive perturbations, $X_v(\theta_j) = B_v(\theta_j) + A_v$ implies

$$\frac{1}{nm}\sum_v \{X_v(\theta_j) - X_v(\theta_k)\} = \frac{1}{nm}\sum_v \{B_v(\theta_j) - B_v(\theta_k)\},$$

so arithmetic-mean rankings are invariant although uncertainty changes. Naive pixel averaging and true-graph Catoni both achieve Top-1 oracle accuracy 74.7% at $n = 32$ (Wilson 95% CI $[67.2, 81.0]\%$) and 96.7% at $n = 256$ ($[92.4, 98.6]\%$), with nearly identical regrets. This is a boundary result, not estimator superiority.

A target-aligned risk-control audit selects the earliest checkpoint whose empirically calibrated UCB-style margin falls below a pseudo-risk threshold from the same process. The pre-specified square-root rule is empirical, not Theorem 1's unknown-constant 5% certificate. A 10k reference confirms epochs 50 and 60 are feasible, yet graph-aware and observation-level Catoni abstain on all 150 draws for every $n \leq 128$; naive $K = 1$ Catoni deploys 1/150 and 2/150 times at $n = 64, 128$, rising to 9/150 and 17/150 under the pre-specified $\alpha = 0.10$ sensitivity. Since $q = 1.6$ gives $a_q = 3/8$, a post-hoc reanalysis on the identical frozen estimates (no new randomness or retuning) replaces the square-root exponent by $3/8$: all rules abstain at $\alpha = 0.05$, while at $\alpha = 0.10$ naive counting deploys 1/150 and 2/150 times at $n = 64, 128$ and both dependence-aware rules still abstain. The qualitative boundary is unchanged; Appendix F.4 gives full results.

A whole-image audit gives a complementary target-shift boundary: noise and occlusion are nearly saturated, while blur yields clean-oracle Top-1 $= 0$ and matched-corruption Top-1 $= 1$ across tested sizes. Uncertainty cannot repair a validation distribution whose preferred checkpoint differs from deployment; Appendix Table 4 gives the $n = 128$ slice.

Figure 3 shows that conservative moment-scale multipliers $\kappa \in \{1, 1.5, 2, 4\}$ retain $R^2 \geq 0.968$ for all displayed finite-q slope fits. At $(n, m, K) = (2048, 3072, 64)$ on an RTX 3090, median times are 0.12 ms (mean), 1.94 ms (observation Catoni), and 63.46 ms (graph-aware Catoni), favoring offline validation/auditing over per-iteration training.

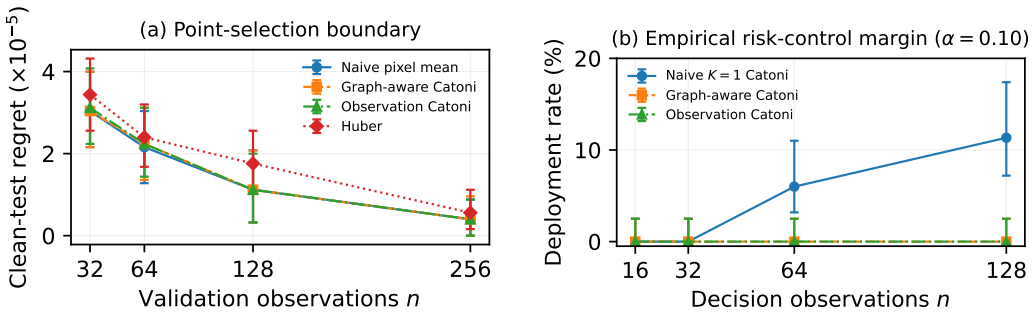


Figure 2: Decision-level boundaries. **Left:** under candidate-shared heavy-tailed perturbations, clean-test regret decreases with n and naive, graph-aware, and observation-level estimators nearly overlap. **Right:** with the pre-specified square-root target-aligned margins at $\alpha = 0.10$, graph-aware and observation-level Catoni abstain through $n = 128$, while naive $K = 1$ Catoni deploys 9/150 and 17/150 times at $n = 64, 128$. Appendix F.4 reports the post-hoc $3/8$ reanalysis. These are empirical decision margins, not Theorem 1's certificate.

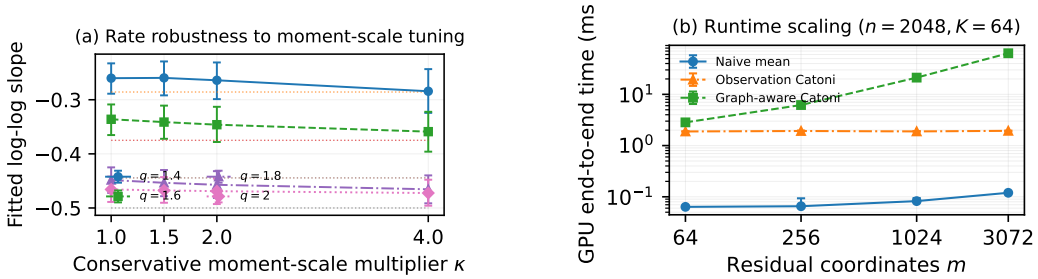


Figure 3: Robustness and computation. **Left:** finite-q slopes under conservative moment-scale multipliers $\kappa \geq 1$, with dotted theoretical exponents $-(q-1)/q$. **Right:** RTX 3090 runtime at $n = 2048, K = 64$; graph-aware Catoni scales approximately linearly in m, while observation-level Catoni is nearly flat. Medians with IQR error bars are shown.

7 Assumptions and Practical Scope

The guarantees assume residual dependence: input topology, proximity, or correlation need not certify Definition 1 for a globally coupled predictor. Table 1 distinguishes certified structure from diagnostic evidence. With independent observations, the within-observation graph gives $N_{\mathrm{eff},q} = n$; when no trusted structure is available, the complete space–time graph gives $N_{\mathrm{eff},q} = 1$. For nonstationary data, Eq. (2) represents a finite-design target unless sampling assumptions are imposed.

The theory uses a pre-specified q and an upper bound on V_q; data-dependent choices can be accommodated through uniform control or sample splitting. Small active classes may require care in confidence construction, while exact general-graph fractional covers can be computationally demanding. The real-data study uses one CIFAR-10 autoencoder family; squared-residual sensitivity tests robustness to the loss specification rather than cross-domain variation. Extensions to general-graph matching lower bounds, broader data settings, and adaptive certification are natural directions for future work.

8 Conclusion

High resolution helps only when it adds evidence for the target. The profile-aware evidence count $N_{\mathrm{eff},q}$ turns a valid dependence description into the finite-q scale, while canonical block lower bounds show the penalty is unavoidable up to logarithms. Experiments show undercoverage from dependence undercounting and large CIFAR uncertainty inflation; decision audits show that valid uncertainty need not change every ranking or deployment rule.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and manuscript drafting. All mathematical derivations, proof verification, experimental design, and code implementation were conducted entirely independently by the authors, who take full responsibility for the paper's scientific claims.

REPRODUCIBILITY STATEMENT

All assumptions used by the theoretical results are stated in the main text, with complete proofs and transfer arguments in the appendix. Appendix F records the frozen experimental protocols, candidate sets, seeds, primary metrics, uncertainty calculations, hardware, and negative outcomes. The final numerical claims are taken only from completed frozen runs; the target-aligned risk-control experiment additionally includes an independent replay of all root solves and decision reconstruction. Citation metadata used in the paper were cross-checked against publisher or proceedings records before submission.

REFERENCES

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

H. C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*. Mathematical Centre Tracts 112, Mathematisch Centrum, Amsterdam, 1979.

Sujay Bhatt, Guanhua Fang, Ping Li, and Gennady Samorodnitsky. Nearly optimal Catoni's M-estimator for infinite variance. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:1925–1944, 2022.

Richard C. Bradley. Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics*, 15(2):5523–5544, 2021.

Paul Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics 85, Springer, New York, 1994.

Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2):37–45, 2014.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

William Kengne and Modou Wade. Robust deep learning from weakly dependent data. *Neural Networks*, 185:107227, 2025.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-IID data. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, PMLR 5:416–423, 2009.

Emmanuel Rio. *Asymptotic Theory of Weakly Dependent Random Processes*. Probability Theory and Stochastic Modelling 80, Springer, 2017.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, pages 8913–8926, 2021.

Lihu Xu, Fang Yao, Qiuran Yao, and Huiming Zhang. Non-asymptotic guarantees for robust statistical learning under infinite variance assumption. *Journal of Machine Learning Research*, 24(92):1–46, 2023.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Rui-Ray Zhang and Massih-Reza Amini. Generalization bounds for learning under graph-dependence: A survey. *Machine Learning*, 113(7):3929–3959, 2024.

Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems*, volume 31, pages 1084–1094, 2018.

Rui (Ray) Zhang, Xingwu Liu, Yuyi Wang, and Liwei Wang. McDiarmid-type inequalities for graph-dependent variables and stability bounds. In *Advances in Neural Information Processing Systems*, volume 32, pages 10889–10899, 2019.

A Finite-q Catoni lemma for independent, non-identically distributed variables

Let $Y_1, \dots, Y_N$ be independent, not necessarily identically distributed, let $\mu = N^{-1}\sum_i \mathbb{E}Y_i$, and suppose $N^{-1}\sum_i \mathbb{E}|Y_i - \mu|^q \le V$ for $q \in (1, 2]$. Put $a_q = (q-1)/q$, $L_\delta = \log(2/\delta)$ and assume $L_\delta \le c_{\mathrm{conf},q} N$. Choose

$$u = A_q V^{1/q}(L_\delta/N)^{a_q}, \qquad \alpha = c_{\alpha,q}(L_\delta/(NV))^{1/q}.$$

Let $\widehat{\mu}$ be any root of $\sum_i \psi_q\{\alpha(Y_i - z)\} = 0$. Since the estimating score is nonincreasing in z,

$$\Pr(\widehat{\mu} \ge \mu + u) \le \Pr\left(\sum_i \psi_q\{\alpha(Y_i - \mu - u)\} \ge 0\right) \le \mathbb{E}\exp\left[\sum_i \psi_q\{\alpha(Y_i - \mu - u)\}\right].$$

By the upper envelope stated after equation 3 (see Chen et al. (2021)),

$$e^{\psi_q(\alpha(Y_i-\mu-u))} \le 1 + \alpha(Y_i - \mu - u) + \frac{|\alpha(Y_i - \mu - u)|^q}{q}.$$

Taking expectations, using the average identity $N^{-1}\sum_i \mathbb{E}(Y_i - \mu) = 0$, applying $1 + x \le e^x$, and multiplying independent factors yields

$$\mathbb{E}e^{S(u)} \le \exp\{-N\alpha u + C'_q N\alpha^q(V + u^q)\}.$$

Now $N\alpha u = c_{\alpha,q}A_q L_\delta$, $N\alpha^q V = c^q_{\alpha,q}L_\delta$, and $N\alpha^q u^q = c^q_{\alpha,q}A^q_q L_\delta(L_\delta/N)^{q-1}$. The confidence condition uniformly bounds the last factor. Choosing $c_{\alpha,q}$ sufficiently small and A_q sufficiently large gives upper-tail probability at most $\delta/2$; oddness gives the lower tail. Hence

$$|\widehat{\mu} - \mu| \le C_q V^{1/q}\left(\frac{\log(2/\delta)}{N}\right)^{a_q} \tag{13}$$

with probability at least $1 - \delta$. Common means are not required: the proof uses independence, the averaged centered-moment bound, and $\sum_i(\mathbb{E}Y_i - \mu) = 0$.

B PROOFS OF THE MASTER THEOREM AND GRAPH-TRANSFER STATEMENTS

Applying equation 13 to every color class and every $\theta \in \Theta_\eta$, with failure budget $\delta/(LM_\eta)$, gives simultaneously

$$|\widehat{R}_\ell(\theta) - R_\ell(\theta)| \le C_q V_q^{1/q} \left(\frac{\Lambda}{N_\ell}\right)^{a_q}.$$

Weighting by $N_\ell/(nm)$,

$$\begin{aligned}|\widehat{R}(\theta) - R_{n,m}(\theta)| &\le C_q V_q^{1/q} \Lambda^{a_q} \frac{1}{nm} \sum_{\ell=1}^{L} N_\ell^{1/q} \\ &= C_q V_q^{1/q} \left(\frac{\Lambda}{N_{\mathrm{eff},q}(\mathcal{C})}\right)^{a_q}.\end{aligned}$$

Let Δ_η denote the right-hand uniform deviation. Choose $\theta_\eta^\star \in \Theta_\eta$ with $R_{n,m}(\theta_\eta^\star) \le R_{n,m}^* + \omega(\eta)$. Then

$$R_{n,m}(\widehat{\theta}_\eta) - R_{n,m}^* \le \omega(\eta) + 2\Delta_\eta,$$

and the factor 2 is absorbed into C_q. Concavity gives $\sum_\ell N_\ell^{1/q} \le L^{a_q}(nm)^{1/q}$ and therefore $N_{\mathrm{eff},q} \ge nm/L$.

B.1 SAFE SUPERGRAPHS AND INDEPENDENT-PILOT TRANSFER

If $\bar{G} \supseteq G^\star$ and $G^\star$ is valid, any independent set of $\bar{G}$ is also an independent set of $G^\star$; hence every coloring of $\bar{G}$ is valid for the true law. If an independent pilot returns a random $\widehat{G}$ satisfying $\Pr(\widehat{G} \supseteq G^\star) \ge 1 - \delta_0$ for an underlying valid $G^\star$, condition on that event. The analysis sample is independent of the pilot, so $\widehat{G}$ is fixed conditionally and the master theorem applies; a union bound increases the final failure probability by at most δ_0.

B.2 PILOT MOMENT TRANSFER AND CONSERVATIVE MOMENT SCALE

Suppose an independent pilot returns $\widehat{V}_q$ satisfying $\Pr(\widehat{V}_q \ge V_q) \ge 1 - \delta_V$. Conditional on that event, all class-wise moment assumptions remain valid with $\widehat{V}_q$ in place of V_q; independence of pilot and analysis samples makes the tuning fixed conditionally, and the final failure probability increases by at most δ_V. This is a transfer interface for an independently validated upper bound, not a claim of fully adaptive estimation under only a finite q-th moment.

If the analysis estimator uses $\bar{V}_q$ with $V_q \le \bar{V}_q \le \kappa V_q$, the same exponential-moment proof remains valid and the stochastic radius is inflated by at most $\kappa^{1/q}$. Underestimation is not covered.

B.3 APPROXIMATE OPTIMIZATION

If $\widetilde{\theta}_\eta \in \Theta_\eta$ satisfies $\widehat{R}(\widetilde{\theta}_\eta) \le \inf_{\theta \in \Theta_\eta} \widehat{R}(\theta) + \epsilon_{\mathrm{opt}}$, then the master bound remains valid after adding ϵ_{opt} to its right-hand side. Indeed, on the uniform-deviation event, the usual empirical-minimizer comparison gives $R_{n,m}(\widetilde{\theta}_\eta) - R_{n,m}^* \le \omega(\eta) + 2\Delta_\eta + \epsilon_{\mathrm{opt}}$.

B.4 FINITE-CANDIDATE SELECTION AND PROFILE-ADAPTIVE CONFIDENCE

For each candidate θ_j, apply the class-wise bound with its own valid coloring and failure budget δ/M. A union bound over at most $ML_{\max}$ candidate/class pairs yields simultaneous deviations with logarithmic term $\log(2L_{\max}M/\delta)$. Comparing the selected robust-risk minimizer with the best candidate gives equation 10 after absorbing the usual factor two into C_q.

More generally, assign candidate/class-specific budgets $\delta_{j\ell} > 0$ with $\sum_{j,\ell} \delta_{j\ell} \le \delta$ and require $\log(2/\delta_{j\ell}) \le c_{\mathrm{conf},q} N_{j\ell}$ for every active pair. Then, with probability at least $1 - \delta$,

$$R_{n,m}(\widehat{\theta}) - \min_j R_{n,m}(\theta_j) \le C_q V_q^{1/q} \max_j \frac{1}{nm} \sum_\ell N_{j\ell}^{1/q} \{\log(2/\delta_{j\ell})\}^{a_q}.$$

Feasibility itself imposes $\delta_{j\ell} \geq 2e^{-c_{\mathrm{conf},q} N_{j\ell}}$. If these lower bounds already exceed the total confidence budget, no allocation rule can repair the supplied graph description. In particular, arbitrary unions of color classes are not a valid fallback because the union need not be independent.

B.5 Parameter error under curvature and separation

Suppose $\theta^\star$ is the unique minimizer, $R_{n,m}(\theta) - R^*_{n,m} \geq \kappa d(\theta, \theta^\star)^\rho$ whenever $d(\theta, \theta^\star) \leq r_0$, and $\Delta_0 := \inf_{d(\theta,\theta^\star)\geq r_0}\{R_{n,m}(\theta) - R^*_{n,m}\} > 0$. On any event where an excess-risk bound $B_\eta < \Delta_0$ holds, the estimator lies inside the local neighborhood and

$$d(\widehat{\theta}_\eta, \theta^\star) \leq (B_\eta/\kappa)^{1/\rho}.$$

Thus the evidence-counting risk rate transfers directly to parameter error whenever the target risk has the stated local geometry.

C Fractional-cover proofs and examples

For an exact cover, the target risk decomposes exactly:

$$R_{n,m}(\theta) = \frac{1}{nm}\sum_j \lambda_j \sum_{v\in I_j} \mathbb{E}X_v(\theta) = \frac{1}{nm}\sum_j \lambda_j |I_j| R_j(\theta),$$

because each vertex receives total cover weight one. Applying equation 13 to each I_j and each $\theta \in \Theta_\eta$ gives

$$|\widehat{R}_j(\theta) - R_j(\theta)| \leq C_q V_q^{1/q} \left(\frac{\Lambda_\lambda}{|I_j|}\right)^{a_q}.$$

Therefore

$$|\widehat{R}_\lambda(\theta) - R_{n,m}(\theta)| \leq C_q V_q^{1/q} \Lambda_\lambda^{a_q} \frac{1}{nm}\sum_j \lambda_j |I_j|^{1/q} = C_q V_q^{1/q} \left(\frac{\Lambda_\lambda}{N^{\mathrm{frac}}_{\mathrm{eff},q}(\lambda)}\right)^{a_q},$$

and the empirical-minimizer comparison is identical to Theorem 1.

For a disjoint union of $B = nm/r$ equal cliques of size r, every independent set has size at most B. Exact coverage gives $\sum_j \lambda_j |I_j| = nm$, and since $1/q - 1 < 0$,

$$T_q(\lambda) = \sum_j \lambda_j |I_j|\, |I_j|^{1/q-1} \geq B^{1/q-1} \sum_j \lambda_j |I_j| = r B^{1/q}.$$

The r balanced color classes attain equality, so $N^{\mathrm{frac}}_{\mathrm{eff},q} = B = nm/r$.

For C_5 and $q = 2$, every proper 3-coloring has profile $(2, 2, 1)$ and integral count $\left(5/(2\sqrt{2}+1)\right)^2 \approx 1.706$. The five maximum independent pairs, each weighted by $1/2$, form an exact cover with $T_2 = (5/2)\sqrt{2} = 5/\sqrt{2}$ and hence fractional count 2. Conversely, $|I_j| \leq 2$ and exactness imply $T_2 \geq 5/\sqrt{2}$, so the displayed fractional value is optimal.

D Approximate independence and dependent observations

D.1 Class-local total-variation transfer

Consider a finite candidate set. For candidate j and active class $D_{j\ell}$, let $P_{j\ell}$ be the joint law of the class residual vector and let $Q_{j\ell}$ be the product of the same one-dimensional marginals. If $d_{\mathrm{TV}}(P_{j\ell}, Q_{j\ell}) \leq \epsilon_{j\ell}$, then every Catoni deviation event has probability under $P_{j\ell}$ at most its probability under $Q_{j\ell}$ plus $\epsilon_{j\ell}$. Allocating the original confidence budget and summing the transfer errors yields the finite-candidate excess-risk bound with failure probability at most $\delta + \sum_{j,\ell} \epsilon_{j\ell}$. No risk-bias term appears because the marginals, hence class means and marginal moments, are preserved.

A coarser whole-law fallback follows from maximal coupling. If the complete residual-data law P is within ϵ in total variation of a reference law Q under which the graph is valid and both laws satisfy $\sup_{\theta,t,s} \mathbb{E}|X_{t,s}(\theta)|^q \le B_q$, write R_P and R_Q for the risk in equation 2 computed under P and Q, respectively. Hölder's inequality then gives $\sup_\theta |R_P(\theta) - R_Q(\theta)| \le 2B_q^{1/q}\epsilon^{a_q}$. Coupling failure adds ϵ to the failure probability and the two risk-comparison terms add at most $4B_q^{1/q}\epsilon^{a_q}$ to the bound. This is a robustness sanity check, not a claim that pairwise correlation controls total variation.

D.2 β-MIXING OBSERVATIONS

Suppose the observation process is stationary and absolutely regular with coefficient $\beta(k)$. Retain a coordinate coloring with K_m classes as in Section 4.2. For a spacing parameter $b \ge 0$, set $L_b = 2b+1$ and partition time indices into residue classes modulo L_b. Sequential Berbee coupling (Berbee, 1979; Doukhan, 1994; Bradley, 2005; Yu, 1994; Rio, 2017) produces independent copies within each residue class with total mismatch probability at most $n\beta(L_b)$. On the successful coupling event the product-color theorem applies with temporal factor L_b; adding coupling failure yields probability at least $1 - \delta - n\beta(L_b)$.

Writing stoch for the stochastic term in the corresponding excess-risk bound, for exponential mixing $\beta(b) \le B_0 e^{-B_1 b}$, choosing $b \asymp B_1^{-1}\log(n/\delta)$ gives, up to constants depending on (B_0, B_1, q),

$$\mathrm{stoch} \lesssim V_q^{1/q}\left(\frac{K_m \log(n/\delta)\log(K_m M_\eta n/\delta)}{nm}\right)^{a_q}.$$

For polynomial mixing $\beta(b) \le B_0(1+b)^{-\zeta}$ with $\zeta > 1$, taking $b \asymp (n/\delta)^{1/\zeta}$ yields

$$\mathrm{stoch} \lesssim V_q^{1/q}\left(\frac{K_m \delta^{-1/\zeta}\log(K_m M_\eta n/\delta)}{mn^{1-1/\zeta}}\right)^{a_q},$$

again up to constants depending only on the displayed mixing parameters and q. These are theoretical extensions; the main empirical claims do not rely on estimating a mixing coefficient from CIFAR-10.

E LOWER-BOUND PROOFS

Proof of Theorem 3. There are $N = nm/r$ independent latent variables. Let $0 < p < 1/4$. Under P_+, a latent variable W equals 1 with probability p and 0 otherwise; under P_- it equals -1 with probability p and 0 otherwise. For actions $\theta \in \{+,-\}$ define

$$X(+;W) = \frac{a}{2}(1-W), \qquad X(-;W) = \frac{a}{2}(1+W).$$

Under P_+ action $+$ is optimal and choosing $-$ incurs excess risk ap; the roles reverse under P_-. For either $\theta \in \{+,-\}$ and either alternative, writing $X = X(\theta;W)$,

$$\mathbb{E}|X - \mathbb{E}X|^q = \left(\frac{a}{2}\right)^q [p(1-p)^q + (1-p)p^q] \le 2^{1-q} p a^q \le p a^q.$$

Thus every residual in a valid color class satisfies the centered-moment bound in equation 4 when $a = (V_q/p)^{1/q}$, and the action gap is $ap = V_q^{1/q} p^{a_q}$. The single-latent total-variation distance is p, so the N-fold product distance is at most Np. Taking $p = 1/(4N)$ makes this distance at most $1/4$, while the action gap becomes

$$ap = 4^{-a_q} V_q^{1/q}\left(\frac{r}{nm}\right)^{a_q}.$$

Le Cam's two-point inequality therefore gives, for every estimator, a testing error of at least $3/8$ under one of the two alternatives, proving the constant-probability statement in equation 11 for a q-dependent constant $c_q > 0$. Since excess risk is nonnegative, the inequality $\mathbb{E}Y \ge t\,\mathbb{P}(Y \ge t)$ for $Y = R_{n,m}(\widetilde{\theta}) - R^*_{n,m}$ yields the stated expectation lower bound after possibly decreasing c_q.

Proof of Theorem 4. There are $N = (n/g)(m/r) = nm/(rg)$ independent space–time latent variables. Copy one latent scalar to every rg entries in a block and apply the same two-law construction. With $p = 1/(4N)$, the product total-variation distance is again at most $1/4$, so one testing error is at least $3/8$. On that error event the excess risk is $4^{-a_q} V_q^{1/q}(rg/(nm))^{a_q}$, proving the constant-probability lower bound after choosing a q-dependent $c_q > 0$. The expectation lower bound follows from the same nonnegativity argument as above.

F EXPERIMENTAL PROTOCOLS AND FULL DECISION AUDITS

F.1 FROZEN MECHANISM EXPERIMENTS

For finite-q scaling we fix $n = 512, m = 64$, block sizes $r \in \{1, 2, 4, 8, 16, 32, 64\}$, and 250 repetitions per point. The three settings $(\nu, q) = (1.5, 1.4), (1.8, 1.6), (2.5, 2)$ use fixed seeds 2026091610–2026091612. A separate two-million-sample pilot supplies an oracle tuning proxy for the centered moment scale; it is not a theorem-certified moment upper confidence bound. Slope intervals use 3,000 bootstrap resamples with frozen bootstrap seed 2026091699.

For the count-misspecification diagnostic we fix $n = 128, m = 64$, generator block size $r = 8$, and $q = 2$, so $a_q = 1/2$. We calibrate the interval radius at the matching color factor $K = 8$ on 1,000 repetitions, then evaluate a disjoint 1,000-run set after multiplying that radius by $\sqrt{K/8}$ for $K \in \{1, 4, 8, 16, 64\}$. In this balanced block diagnostic $N_{\mathrm{eff},q} = nm/K$, so the square-root adjustment is exactly the corresponding finite-q evidence-count scaling. Thus $K < 8$ deliberately undercounts dependence, $K = 8$ matches the generator, and $K > 8$ is conservative; this experiment isolates evidence-count calibration rather than comparing graph estimators. Wilson intervals are reported.

F.2 CIFAR-10 RESIDUAL DIAGNOSTICS

A fixed convolutional autoencoder is trained on CIFAR-10 using split seed 20260921: Conv3×3 stride-2 layers $3 \to 32 \to 64 \to 128$ with ReLU, followed by ConvTranspose4×4 stride-2 layers $128 \to 64 \to 32 \to 3$, ReLU, and a final sigmoid. Training uses Adam, learning rate 10^{-3}, batch size 256, no augmentation, and a 45k/5k train/validation split; checkpoints are saved at epochs $\{5, 10, 15, 20, 30, 40, 50, 60\}$. Absolute reconstruction residual is primary and squared residual is a frozen sensitivity. For each checkpoint we report horizontal/vertical residual correlation, kurtosis, and image-cluster versus naive pixel-bootstrap standard errors over 20 fixed validation subsets of 256 images with 1,000 bootstrap replicates each. Patch sizes $1, 2, 4, 8$ are all retained. Because the image residuals are bounded, these are descriptive dependence/tail-shape measurements rather than evidence of infinite variance, and no natural-image graph is declared valid from correlation. Across the eight checkpoints, squared-residual sensitivity gives an average cluster-to-pixel SE ratio of $8.87\times$ at pixel resolution and $2.47\times$ at 8×8, compared with $10.49\times$ and $2.49\times$ for the primary absolute residual.

F.3 CONTROLLED CHECKPOINT SELECTION

Eight fixed checkpoints form the primary candidate set. The clean 10k test set defines oracle reconstruction risk and regret only. Each pseudo-observation has $m = 64$ residual coordinates partitioned into blocks with primary $r = 8$. Within each block, a real validation residual patch receives a block-shared absolute Student-t perturbation plus coordinate jitter; the same additive perturbation realization is shared across candidates, preserving the population ordering by construction. We use formal seeds 20260921–20260925, 30 paired draws per seed, and $n \in \{32, 64, 128, 256\}$ (150 draws per n). Table 3 reports Wilson intervals for the two representative point-ranking rules.

Table 3: Controlled checkpoint-selection boundary. Top-1 clean-oracle selection accuracy; 95% Wilson intervals in brackets.

n	Naive pixel mean	True-graph Catoni
32	74.7 [67.2, 81.0]	74.7 [67.2, 81.0]
64	82.0 [75.1, 87.3]	81.3 [74.3, 86.8]
128	90.7 [84.9, 94.4]	90.7 [84.9, 94.4]
256	96.7 [92.4, 98.6]	96.7 [92.4, 98.6]

The whole-image corruption audit uses a frozen fallback pipeline: Gaussian blur (kernel 5, $\sigma = 1$), additive Gaussian noise ($\sigma = 0.1$ with clipping to $[0, 1]$), and an 8×8 zero-filled occlusion at a seeded location (corruption seed 2026092601). Noise and occlusion are nearly saturated for all strong methods. Under blur, every strong method has clean-oracle Top-1 equal to 0 and matched-corruption-

oracle Top-1 equal to 1 across tested validation sizes, so we treat this as a target-sensitivity audit rather than evidence for graph-aware dominance.

Table 4: Whole-image corruption selection at $n = 128$ (150 draws): Top-1 ranges across eight frozen aggregators.

Validation condition	Clean-oracle Top-1	Matched-oracle Top-1	Interpretation
Clean	$[0.713, 0.953]$	$[0.713, 0.953]$	finite-sample selection
Noise	1.000	1.000	saturated
Occlusion	1.000	1.000	saturated
Blur	0.000	1.000	target ordering changes

F.4 TARGET-ALIGNED RISK-CONTROLLED CHECKPOINT SELECTION

The final pre-specified audit fixes the target mismatch present in an earlier stress test. Calibration and decision pools are disjoint but use the same pseudo-field

$$X_{t,s}^{(j)} = W_{t,b} B_{t,s}^{(j)} + \sigma s_0 |\epsilon_{t,s}|, \qquad W_{t,b} = \frac{|T_{\nu,t,b}|}{\mathbb{E}|T_\nu|}, \qquad T_{\nu,t,b} \overset{\text{i.i.d.}}{\sim} t_\nu.$$

Here $b = b(s)$ indexes the block containing coordinate s, $B_{t,s}^{(j)}$ is checkpoint j's baseline residual sampled from the fixed CIFAR validation residual pool, $\epsilon_{t,s} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is independent coordinate-level jitter, and s_0 is the fixed global residual scale computed once on the calibration pool and shared across all checkpoints. The block multipliers are independent across (t, b) and satisfy $\mathbb{E}W_{t,b} = 1$. We use $m = 64, r = 8, \nu = 1.8, \sigma = 0.1, q = 1.6$, eight fixed checkpoints, and $n \in \{16, 32, 64, 128\}$. Let $R_{\text{cal}}^{\text{pseudo}}(j)$ denote checkpoint j's empirical pseudo-risk on the fixed calibration pool, and set $\tau(\alpha) = (1 + \alpha) \min_j R_{\text{cal}}^{\text{pseudo}}(j)$ with primary $\alpha = 0.05$ and a pre-specified sensitivity $\alpha = 0.10$. Formal seeds are 20260921–20260925 with 30 paired draws per seed; calibration and bootstrap seeds are 2026091902 and 2026091904. A large independent sanity run verifies that calibration and decision pseudo-risk means agree up to the finite-pool difference and Monte Carlo error. The pre-specified decision rule reuses the frozen empirical reference radius 0.0135138880347, scaled by $\sqrt{1024/N_{\text{eff},q}}$ for the chosen evidence count. This square-root scaling is inherited from the original empirical calibration; it is not the finite-q exponent of Theorem 1. Because $q = 1.6$, the theorem has $a_q = (q - 1)/q = 3/8$. We therefore additionally report a post-hoc *theory-exponent-aligned* reanalysis on exactly the same frozen candidate estimates, with no new randomness or retuning, using

$$r_{\text{align}}(N_{\text{eff},q}) = 0.0135138880347 \left(\frac{1024}{N_{\text{eff},q}} \right)^{3/8}.$$

Neither margin is the unknown-constant bound of Theorem 1; the aligned analysis matches only its $N_{\text{eff},q}^{-a_q}$ exponent and is not a theorem-certified 5% risk guarantee.

Under the pre-specified square-root calibration, at $\alpha = 0.05$ naive counting deploys 0/150, 0/150, 1/150, and 2/150 times for $n = 16, 32, 64, 128$; graph-aware Catoni and observation-level Catoni deploy 0/150 times at every n. The independent 10k test-pool reference places epochs 50 and 60 below this threshold, so universal graph-aware abstention cannot be attributed to an empty reference-feasible set. No unsafe deployment is observed among the three naive deployments, but the conditional sample size is far too small to claim calibrated safety. At $\alpha = 0.10$, naive deployments rise to 9/150 at $n = 64$ and 17/150 at $n = 128$, while both dependence-aware rules still abstain on all draws. Figure 4 shows the primary deployment rates.

Theory-exponent-aligned consistency check. Reapplying only the decision layer to the same frozen candidate estimates with the $N_{\text{eff},q}^{-3/8}$ scaling does not reverse the qualitative conclusion. At the primary $\alpha = 0.05$, naive, graph-aware Catoni, and observation-level Catoni all abstain on all 150 draws for every $n \in \{16, 32, 64, 128\}$. At the pre-specified $\alpha = 0.10$ threshold, naive counting deploys 0/150, 0/150, 1/150, and 2/150 times as n increases, whereas graph-aware and observation-level Catoni still abstain on every draw. No unsafe event is observed among these three naive deployments, which is far too few to support a conditional safety claim. Thus the finite-sample conservatism of the dependence-aware decision rules persists when the evidence-count exponent is aligned with the finite-q theory.

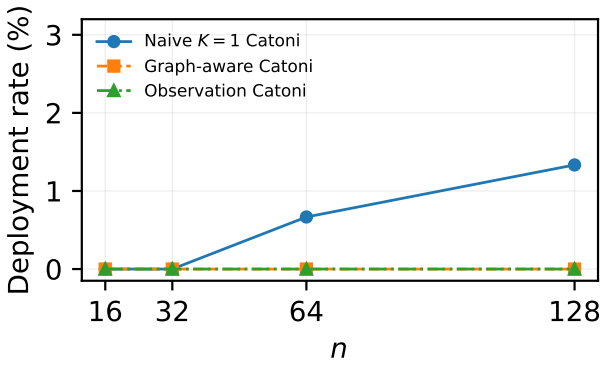


Figure 4: Target-aligned risk-control audit under the pre-specified square-root calibration at the primary $\alpha = 0.05$. The graph-aware and observation-level rules remain too conservative to deploy for $n \leq 128$, although the independent test-pool reference identifies checkpoints below the frozen threshold.

F.5 Fractional, profile, tuning, and runtime stress tests

The profile-aware evidence count is a high-probability certificate, not a claim that realized finite-sample RMSE must collapse exactly as $N_{\mathrm{eff},q}^{-a_q}$. In the frozen class-imbalance stress test, the crude N/L axis has higher RMSE-collapse R^2 than the profile-aware evidence count for the tested profiles, and extreme one-large-many-small configurations can contain singleton classes outside the useful finite-RMSE/moderate-confidence regime. We therefore do not use this experiment as positive evidence for exact finite-sample RMSE prediction.

On the C_5 graph with $q = 2$, the fractional certificate is strictly larger ($2n$ versus $1.706n$ for the integral profile), yet realized RMSEs are nearly identical: at $n = 128$, integral and fractional Catoni give 0.12345 and 0.12363, respectively, while the fractional implementation revisits more sets and is slower. This supports the narrower theorem-level statement that fractional covers can tighten the upper certificate, not that they must dominate integral estimators empirically.

Conservative moment-scale perturbation is also tested directly. Replacing V_q by κV_q changes the theoretical radius by exactly $\kappa^{1/q}$; the frozen sensitivity suite verifies this implementation across $q \in \{1.4, 1.6, 1.8, 2\}$ and $\kappa \in \{1, 1.5, 2, 4\}$. The deliberately theorem-violating multiplier 0.5 is reported separately.

Runtime is measured for $n \in \{128, 512, 2048\}$, $m \in \{64, 256, 1024, 3072\}$, and $K \in \{1, 4, 16, 64\}$ with 20 repeats on CPU and an NVIDIA GeForce RTX 3090. Graph/coloring and fractional-cover construction are timed separately from robust-mean evaluation. At $(2048, 3072, 64)$ the GPU median end-to-end times are 0.12 ms (mean), 1.94 ms (observation Catoni), 63.46 ms (integral graph-aware Catoni), and 63.83 ms (fractional Catoni).