Sharp collision criteria for frozen color histograms

Scott Schubr
Luminary Oddities

October 26, 2026
Archive edition 2

Abstract

A probability law selects fixed-size subsets of a finite ground set, whose elements are independently colored once. We compare the histogram law conditional on that frozen coloring with the prescribed multinomial law. Under upper negative-cylinder inequalities, the normalized marginal collision sum gives a sharp criterion for expected chi-square convergence and its leading coefficient, uniformly over positive varying color probabilities and color counts. A covariance statistic gives a linear collision lower bound for expected total variation without any dependence hypothesis on the subset law. Under the upper-cylinder hypothesis, we obtain an exact convergence criterion for binary colors with varying probabilities that combines subset collisions and color rarity. For a fixed positive color vector, the same hypothesis gives total-variation convergence exactly when the collision sum vanishes. A non-real-stable example and applications to spanning trees and matroid-base rounding illustrate the scope.

2020 Mathematics Subject Classification: 60F05; 60C05; 60E15.
Keywords: multinomial approximation; frozen coloring; negative dependence; chi-square divergence; total variation.

1 The frozen experiment and its collision criterion

Let E be finite, $n \geq 1$, and let μ be a probability law on $\binom{E}{n}$, the n-element subsets of E. Write

$$p_e = \mathbb{P}_\mu(e \in S), \qquad \tau = \frac{1}{n}\sum_{e\in E} p_e^2, \qquad \ell = \max_{e\in E} p_e. \tag{1}$$

Thus $\sum_e p_e = n$ and $0 < \tau \leq \ell \leq 1$. Our dependence hypothesis is

$$\mathbb{P}_\mu(I \subseteq S) \leq \prod_{e\in I} p_e \quad \text{for every } I \subseteq E. \tag{2}$$

We call (2) the upper negative-cylinder property.

Independently of S, assign independent colors $H_e \in \{1, \dots, m\}$ with $\mathbb{P}(H_e = i) = q_i > 0$ and $\sum_i q_i = 1$. This assignment is made once. For a fixed coloring h, let P_h be the law, over $S \sim \mu$ only, of

$$C_h(S) = \left(|S \cap h^{-1}(i)|\right)_{i=1}^m.$$

Preprint available at https://research.luminaryoddities.com/sharp-collision-criteria-frozen-color-histograms/v/2.

The reference is the prescribed law $M = \mathrm{Mult}(n, q)$ on $\mathcal{A}_{n,m} = \{a \in \mathbb{Z}_{\geq 0}^m : \sum_i a_i = n\}$, rather than a multinomial with a coloring-dependent mean. We use

$$\chi^2(P\|M) = \sum_{a \in \mathcal{A}_{n,m}} \frac{(P(a) - M(a))^2}{M(a)}, \qquad \mathrm{TV}(P, M) = \frac{1}{2} \sum_a |P(a) - M(a)|.$$

The source law μ is fixed independently of H. Averaging the histogram itself over H always gives M, for any fixed-size source law. The question here concerns the distance of its *conditional* law P_H from M.

The experiment in three steps

Draw once	Independently assign each ground element a color using the prescribed vector q.
Freeze and sample	Hold the coloring h fixed; redraw subsets from the same independent source law μ to obtain P_h.
Compare	Compare P_h with $\mathrm{Mult}(n, q)$, then average the distance over colorings. The reference does not use empirical color proportions.

Theorem 1 (Collision criterion)**.** *Assume* (2) *and* $m \geq 2$*, and put* $r = m - 1$*. Let* S, T *be independent with law* μ*, and* $J = |S \cap T|$*. Then*

$$\mathbb{E}_H \chi^2(P_H\|M) = \sum_{k=1}^{n} \binom{r+k-1}{k} \frac{\mathbb{E}(J)_k}{(n)_k}. \tag{3}$$

Here $(x)_k = x(x-1)\cdots(x-k+1)$ *is a falling factorial. If* $e\tau < 1$*,*

$$r\tau \;\leq\; \mathbb{E}_H \chi^2(P_H\|M) \leq r\tau + (1 - e\tau)^{-r} - 1 - er\tau, \tag{4}$$

$$\mathbb{E}_H \chi^2(P_H\|M) \leq (1 - e\tau)^{-r} - 1. \tag{5}$$

For arbitrary sequences of such experiments, allowing n, E, m, q, μ *to vary,*

$$\mathbb{E}_H \chi^2(P_H\|\,\mathrm{Mult}(n, q)) \longrightarrow 0 \quad \Longleftrightarrow \quad (m-1)\tau \longrightarrow 0. \tag{6}$$

In this regime,

$$\mathbb{E}_H \chi^2(P_H\|M) = (m-1)\tau(1 + o(1)), \tag{7}$$

uniformly over the stated class and all positive q*.*

The identity (3) requires no dependence hypothesis. Its kernel and orthogonal decomposition are classical multinomial reproducing-kernel and Hoeffding–Krawtchouk machinery: in particular, the shared-observation joint law is explicit in Diaconis and Griffiths [1, Remark 7, equation (30)]. We give a finite trace proof to make its normalization and use in this experiment explicit. The additional input is the upper-cylinder control of replica-overlap factorial moments, which yields (6)–(7).

Theorem 2 (Total variation and color rarity)**.** *For every fixed-size subset law, without assuming* (2)*, and every color* i *with* $0 < q_i < 1$*, put* $s_i = q_i(1 - q_i)$ *and* $\kappa_i = \max\{1 - 6s_i, 0\}$*. Then*

$$\mathbb{E}_H\, \mathrm{TV}(P_H, \mathrm{Mult}(n, q)) \geq \frac{\tau}{2\{2 + 2\ell + \kappa_i \ell/(ns_i)\}} \geq \frac{\tau}{10} \min\{1, ns_i\}. \tag{8}$$

Under (2)*, the following convergence criteria hold:*

1. *For binary colors $q = (1-u, u)$, allowing n, μ and $u \in (0,1)$ to vary arbitrarily,*

$$\mathbb{E}_H \operatorname{TV}(P_H, \operatorname{Bin}(n,u)) \longrightarrow 0 \quad \Longleftrightarrow \quad \tau \min\{1, nu(1-u)\} \longrightarrow 0. \tag{9}$$

Here a binary histogram is identified with its second coordinate.

2. *If $m \geq 2$ and the positive vector q are fixed along a sequence,*

$$\mathbb{E}_H \operatorname{TV}(P_H, \operatorname{Mult}(n,q)) \longrightarrow 0 \quad \Longleftrightarrow \quad \tau \longrightarrow 0. \tag{10}$$

The chi-square criterion allows varying positive q and growing m. For binary total variation, (9) also resolves the rare-color regime. The fixed-vector conclusion (10) does not assert an equivalence for arbitrary growing color counts.

Convergence criteria and their parameter regimes

Expected chi-square	Under upper negative cylinders, positive q and the color count may vary; the criterion is $(m-1)\tau \to 0$, with leading coefficient one.
Expected total variation	For fixed positive q and fixed $m \geq 2$, the criterion is $\tau \to 0$. Its necessary direction holds for every fixed-size source law.
Binary total variation	For varying $u \in (0,1)$, the criterion is $\tau \min\{1, nu(1-u)\} \to 0$. Upper negative cylinders are required only for sufficiency.

Overlap and applications For two independent source samples S, T, the parameter is $\tau = \mathbb{E}|S \cap T|/n$. In a spanning-tree law, substitute the edge marginals $p_e = w_e R_{\text{eff}}(e)$; in a fixed matroid-base rounding law, substitute the prescribed marginals $p_e = x_e$. Corollary 5 turns these data into finite envelopes and probability bounds over the frozen coloring.

1.1 Relation to preceding approximation results

Multinomial approximation of dependent variables is classical; Loh [2] provides a general Stein-method treatment. Approximate de Finetti and permutation-mixture results compare exchangeable marginals or permutation mixtures with empirical or mean-matched product references [3, 4, 5, 6]. Their cancellation of first-degree terms is a different normalization from the prescribed-q experiment here. For example, a uniform n-subset of an L-element population has $\tau = n/L$. The exact first term of our expected chi-square is $(m-1)n/L$. Comparing instead to a multinomial with the frozen population's empirical color frequencies removes that first-marginal fluctuation and changes the approximation problem.

For real-stable fixed-total count laws, the author's earlier covariance certificate [7] gives a mean-matched multinomial TV bound. Applied after grouping coordinates, together with control of the coloring-dependent mean, it already supplies fixed-color TV sufficiency for real-stable subset laws. The present results instead need only (2), determine the expected chi-square criterion and its first coefficient without reciprocal-q_i losses, and give fixed-color TV necessity for arbitrary fixed-size laws. The nonstable example in Section 4 demonstrates that the broader source class is substantive. The classical kernel, negative-dependence inputs, and existing rounding algorithms are credited ingredients.

2 The classical overlap trace and factorial control

Lemma 3 (Overlap trace)**.** *Equation (3) holds for every probability law on* $\binom{E}{n}$.

Proof. All sums are finite and M is strictly positive. Expanding the square and using independent $S, T \sim \mu$ gives

$$1 + \mathbb{E}_H \chi^2(P_H \| M) = \mathbb{E}_{S,T} \sum_a \frac{\mathbb{P}_H(C_H(S) = a, C_H(T) = a)}{M(a)}. \tag{11}$$

For fixed S, T with overlap j, their common elements contribute j shared independent categorical observations; the other $n - j$ observations in each histogram are mutually independent. This joint histogram law is also obtained by starting with n independent q-colored coordinates, retaining a uniformly chosen j-subset of their positions, independently refreshing the rest with colors of law q, and forgetting the order. Indeed, conditional on the first histogram, any fixed j positions have the same color-count law as a uniform j-subset of positions. Let K_j be this histogram transition matrix. It is stationary for M, and its contribution to the sum in (11) is $\operatorname{Tr} K_j$.

Choose an orthonormal basis $1, \phi_1, \ldots, \phi_r$ of $L^2(q)$, with the last $r = m - 1$ functions centered. Tensor products form an orthonormal basis on the ordered n coordinates. Uniform keep-j/refresh resampling acts on a product with k nonconstant coordinates by the scalar $(j)_k/(n)_k$: every one of those coordinates must be retained, or its refreshed centered factor integrates to zero. This scalar is zero when $k > j$.

Permutation-invariant functions identify isometrically with histogram functions in $L^2(M)$. In degree k, symmetrized tensor products are indexed by multisets of k labels from $\{1, \ldots, r\}$. Distinct multisets have disjoint tensor-basis supports, so these products are orthogonal and can be normalized. There are $d_{r,k} = \binom{r+k-1}{k}$ such multisets. Together over $0 \le k \le n$ (with $d_{r,0} = 1$), their dimension is $\sum_{k=0}^{n} d_{r,k} = \binom{n+m-1}{n} = |\mathcal{A}_{n,m}|$; they exhaust the histogram space. Therefore

$$\operatorname{Tr} K_j = \sum_{k=0}^{n} d_{r,k} \frac{(j)_k}{(n)_k}.$$

Average over S, T in (11) and remove the constant term. □

For nonnegative numbers $x = (x_e)_{e \in E}$, let $e_k(x) = \sum_{|I|=k} \prod_{e \in I} x_e$ be the elementary symmetric polynomial, with an empty sum interpreted as zero.

Lemma 4 (Replica factorial moments)**.** *Under (2), for* $1 \le k \le n$,

$$a_k := \frac{\mathbb{E}(J)_k}{(n)_k} \le \min\left\{ \ell^k, \frac{k! e_k((p_e^2)_e)}{(n)_k}, \frac{(n\tau)^k}{(n)_k} \right\}. \tag{12}$$

Moreover $a_1 = \tau$ *exactly, and* $a_k \le (e\tau)^k$.

Proof. Writing $u_I = \mathbb{P}(I \subseteq S)$, independence of the two replicas gives

$$\mathbb{E}(J)_k = k! \sum_{|I|=k} u_I^2 \le k! e_k((p_e^2)_e) \le \left(\sum_e p_e^2 \right)^k.$$

The last inequality retains just the ordered distinct-index terms in the nonnegative expansion of the power. Conditional on a realized S, the same factorial moment sums over ordered distinct k-tuples in S. The inclusion probability of each such tuple in T is at most ℓ^k, and there are $(n)_k$ tuples. This proves (12). For $k = 1$, $\mathbb{E}J = \sum_e p_e^2$ without any dependence assumption.

To pay the falling-factorial denominator uniformly, use the increasing function $f(x) = -\log(1 - x/n)$ on $[0, n)$. Its left Riemann sum gives

$$\log \frac{n^k}{(n)_k} = \sum_{j=0}^{k-1} f(j) \le \int_0^k f(x)\,dx$$
$$= k + (n-k)\log(1-k/n) \le k.$$

At $k = n$ the integral is improper but finite, with $0 \log 0 = 0$. Hence $n^k/(n)_k \le e^k$, which proves the final bound. □

Proof of Theorem 1. Lemma 3 gives the identity. Each coefficient and each overlap moment is nonnegative, so retaining $k = 1$ gives the lower bound $r\tau$. With $x = e\tau < 1$, use $a_k \le x^k$ for $k \ge 2$ and

$$\sum_{k=0}^{\infty} \binom{r+k-1}{k} x^k = (1-x)^{-r}$$

to obtain (4). Applying the same bound also at $k = 1$ gives (5).

Suppose $r\tau \to 0$. Since $r \ge 1$, $\tau \to 0$ and $x \to 0$. The expansion $-\log(1-x) = x + O(x^2)$ gives, uniformly when $rx \to 0$,

$$(1-x)^{-r} - 1 - rx = O(r^2x^2 + rx^2).$$

Thus the upper remainder in (4), divided by $r\tau$, is $O(r\tau + \tau) = o(1)$. This proves (7), and its sufficient convergence direction. The lower bound $r\tau$ proves the necessary direction. □

Corollary 5 (Finite envelopes and typical colorings)**.** *The finite quantity*

$$U = \sum_{k=1}^{n} \binom{r+k-1}{k} \min\left\{\ell^k, \frac{k! e_k((p_e^2)_e)}{(n)_k}, \frac{(n\tau)^k}{(n)_k}\right\} \tag{13}$$

is a valid upper bound on $\mathbb{E}_H \chi^2(P_H \| M)$ for all τ. When $\ell < 1$, $(1-\ell)^{-r} - 1$ is another valid bound; when $e\tau < 1$, either upper bound in Theorem 1 is valid. For any such bound U,

$$\mathbb{E}_H \operatorname{TV}(P_H, M) \le \min\{1, \tfrac{1}{2}\sqrt{U}\}, \tag{14}$$
$$\mathbb{P}_H\{\operatorname{TV}(P_H, M) > \tfrac{1}{2}\sqrt{U/\delta}\} \le \delta, \quad 0 < \delta < 1. \tag{15}$$

In particular, if $r\tau \to 0$, expected TV is at most $(1/2 + o(1))\sqrt{r\tau}$ and tends to zero in probability over H.

Proof. The finite bound follows termwise from Lemma 4. The maximum-inclusion relaxation follows from the same binomial series. Cauchy–Schwarz gives $\operatorname{TV}(P, M) \le \frac{1}{2}\sqrt{\chi^2(P\|M)}$. Jensen gives (14); applying Markov to chi-square gives (15). If $U = 0$, the distance vanishes almost surely and the same conclusions hold directly. □

3 Total-variation necessity through a covariance statistic

The necessary direction of Theorem 2 cannot follow from a bare chi-square lower bound. We instead use a statistic whose moments are controlled under both the joint experiment and its independent reference.

Proof of Theorem 2. First take binary colors $H_e \sim \text{Bernoulli}(q)$, $0 < q < 1$, and set $s = q(1-q)$ and $Y_e = H_e - q$. Put

$$A(H) = \frac{\sum_e p_e Y_e}{\sqrt{ns}}, \qquad Z(H,S) = \frac{\sum_{e\in S} Y_e}{\sqrt{ns}}.$$

Let $\mathcal{J}$ be the joint law of (H, C) with $C = \sum_{e\in S} H_e$, where S is independent of H, and let $\mathcal{R} = \mathcal{L}(H) \otimes \text{Bin}(n, q)$. Finite conditioning gives

$$\text{TV}(\mathcal{J}, \mathcal{R}) = \mathbb{E}_H \,\text{TV}(P_H, \text{Bin}(n, q)) =: D. \tag{16}$$

The statistic $F(H, C) = A(H)(C - nq)/\sqrt{ns}$ is a function of this observed pair. Independence and centering of the Y_e give

$$\mathbb{E}_{\mathcal{J}} F = \frac{1}{n}\mathbb{E}_S \sum_{e\in S} p_e = \tau, \qquad \mathbb{E}_{\mathcal{R}} F = 0, \qquad \mathbb{E}_{\mathcal{R}} F^2 = \tau.$$

The last identity uses $\mathbb{E}A^2 = \tau$ and the unit variance of the standardized binomial under the product law.

For fixed S, put $X_S = n^{-1}\sum_{e\in S} p_e$. Expanding four independent centered Bernoulli factors, whose fourth cumulant is $s(1-6s)$, gives exactly

$$\mathbb{E}_H[F(H,C)^2 \mid S] = \tau + 2X_S^2 + \frac{1-6s}{n^2 s}\sum_{e\in S} p_e^2.$$

Indeed the pairings give $\text{Var}(A)\,\text{Var}(Z) = \tau$ and twice the squared covariance $2X_S^2$; the remaining same-index terms give the cumulant. Averaging over S yields

$$\mathbb{E}_{\mathcal{J}} F^2 = \tau + 2\mathbb{E}_S X_S^2 + \frac{1-6s}{n^2 s}\sum_e p_e^3. \tag{17}$$

Now $0 \le X_S \le \ell$ and $\mathbb{E}_S X_S = \tau$, so $\mathbb{E}_S X_S^2 \le \ell\tau$. Also $\sum_e p_e^3 \le \ell n\tau$. Writing $\kappa = \max\{1-6s, 0\}$ and retaining the sign of the cumulant therefore gives

$$\mathbb{E}_{\mathcal{J}} F^2 + \mathbb{E}_{\mathcal{R}} F^2 \le \tau\{2 + 2\ell + \kappa\ell/(ns)\}. \tag{18}$$

This bound retains a factor of τ that a uniform fourth-moment bound would discard.

For two finite laws P, Q and real F, Cauchy–Schwarz against $P + Q$ implies

$$\begin{aligned}|\mathbb{E}_P F - \mathbb{E}_Q F|^2 &\le (\mathbb{E}_P F^2 + \mathbb{E}_Q F^2)\sum_x \frac{(P(x) - Q(x))^2}{P(x) + Q(x)} \\ &\le 2\,\text{TV}(P, Q)(\mathbb{E}_P F^2 + \mathbb{E}_Q F^2).\end{aligned}$$

Terms with zero denominator are zero. Apply this to $\mathcal{J}, \mathcal{R}$ and use (16)–(18). It gives $D \ge \tau/[2\{2 + 2\ell + \kappa\ell/(ns)\}]$. Since $\ell, \kappa \le 1$, this is at least $\tau/(8 + 2/(ns))$, and hence at least $(\tau/10)\min\{1, ns\}$.

For m colors, project the histogram onto coordinate i. The indicator $\mathbf{1}_{\{H_e = i\}}$ is an independent Bernoulli coloring of parameter q_i. TV contracts under projection, proving (8). For binary colors, write $D = \mathbb{E}_H \,\text{TV}(P_H, \text{Bin}(n, u))$ and $s = u(1-u)$. Comparing each law with the point mass at the all-majority histogram, then averaging, gives

$$D \le 2n\min\{u, 1-u\} \le 4ns.$$

Under (2), Theorem 1 and Corollary 5 also imply $D \to 0$ whenever $\tau \to 0$, uniformly in u. The lower bound proves necessity in (9). For sufficiency, suppose its right-hand side tends to zero but $D \geq \varepsilon > 0$ on a subsequence. The rare-color upper bound forces $ns \geq \varepsilon/4$ there. Thus $\min\{1, ns\}$ is bounded below, forcing $\tau \to 0$, a contradiction. This argument does not require n or u to have a limit.

When m, q are fixed and positive, $ns_i \geq s_i > 0$ because $n \geq 1$. The lower bound forces $\tau \to 0$ if expected full-histogram TV tends to zero. Conversely, under (2), $\tau \to 0$ implies $r\tau \to 0$ for fixed r; Corollary 5 gives sufficiency. □

4 Source classes and applications

4.1 A diffuse non-real-stable upper-cylinder family

Real stability is sufficient for the usual fixed-size strongly Rayleigh negative-dependence inputs [8], but is not necessary for our theorem. Label five elements by $\mathbb{Z}/5\mathbb{Z}$. Assign probability $3/20$ to each cycle pair $\{i, i+1\}$ and $1/20$ to each of the five other pairs. All marginals are $p_i = 2/5$. Pair inclusion is at most $3/20 < 4/25$, singleton inclusion is equality, and inclusion of three or more elements is impossible. Thus (2) holds.

To verify nonstability directly, let $g(z) = \sum_{i<j} \mu(\{i,j\}) z_i z_j$ be its homogeneous quadratic generating polynomial and B its Hessian. The constant vector has eigenvalue $2/5$. Real Fourier vectors of frequency one have eigenvalue

$$\lambda = \frac{3}{10}\cos(2\pi/5) + \frac{1}{10}\cos(4\pi/5) = \frac{\sqrt{5}-2}{20} > 0.$$

Take a nonzero such real vector v with $\sum_i v_i = 0$. Then $\nabla g(\mathbf{1}) = (2/5)\mathbf{1}$, $g(\mathbf{1}) = 1$, and $g(v) = \lambda\|v\|^2/2 > 0$. Homogeneity gives

$$g(v + t\mathbf{1}) = g(v) + t\nabla g(\mathbf{1}) \cdot v + t^2 g(\mathbf{1}) = g(v) + t^2.$$

This has a root with $\Im t > 0$, at which every coordinate of $v + t\mathbf{1}$ has positive imaginary part. Hence g is not real stable.

For an arbitrarily diffuse sequence, replace each parent element by L clones. After selecting the parent pair, independently select one uniform clone from each chosen parent. Each clone has marginal $2/(5L)$, so $\tau = 2/(5L)$. Two clones of the same parent cannot both appear; for different parents, pair inclusion is the former pair weight divided by L^2. This proves all upper-cylinder inequalities. Identifying clones within each parent recovers g; such identification preserves real stability directly from its upper-half-plane definition. Thus every cloned law remains nonstable, while $\tau \to 0$.

4.2 Strongly Rayleigh samples and spanning trees

A multiaffine probability generating polynomial is real stable when it has no zero with all coordinates in the upper half-plane. Fixed-size strongly Rayleigh measures have negative association and therefore satisfy (2) [8]. In particular, a finite connected graph on at least two vertices with positive conductances has a strongly Rayleigh weighted spanning-tree law. Its edge inclusion probabilities are $p_e = w_e R_{\mathrm{eff}}(e)$, with the conductance convention for effective resistance. Our parameter is

$$\tau = \frac{\sum_e (w_e R_{\mathrm{eff}}(e))^2}{|V| - 1}.$$

On K_v, a uniform spanning tree has $n = v - 1$ and $p_e = 2/v$ by symmetry and fixed total, hence $\tau = 2/v$. For $m = o(v)$, Theorem 1 gives the sharp averaged chi-square asymptotic

$$\mathbb{E}_H \chi^2(P_H \| \operatorname{Mult}(v-1, q)) \sim \frac{2(m-1)}{v}.$$

Its TV upper consequence is $O(\sqrt{m/v})$. The balanced stable-law TV range was already accessible from the earlier covariance certificate; the sharp chi-square characterization is the assertion emphasized here. These are frozen independent *edge* colors, not the vertex-degree identity arising from Prüfer codes or an adversarial edge-color theorem.

4.3 Fixed matroid-base rounding

Let x be a fixed point in the base polytope of a matroid of rank $n \geq 1$. Fix, independently of the coloring, a randomized base-rounding law with marginal preservation and upper negative cylinders. Randomized swap and pipage rounding supply such laws under their usual hypotheses [9, Theorem 1.1 and Section 4]. Their output S has exactly n elements, $p_e = x_e$, and our parameter is

$$\tau = \frac{1}{n} \sum_e x_e^2.$$

All conclusions apply to its conditional color histogram. This is a consequence for the established rounding law. A point only in an independence polytope may produce a random output size; silently replacing that size by n would change the experiment.

4.4 Why a collision parameter can beat maximum inclusion

Select one deterministic element, and one uniform element from each of $n - 1$ disjoint groups of size L. The generating polynomial is a product of a monomial and positive linear forms, so the law is real stable and satisfies (2). Here

$$\ell = 1, \qquad \tau = \frac{1 + (n-1)/L}{n}.$$

For $L = n^2$ and $m = \lfloor \sqrt{n} \rfloor$, the collision criterion holds although an infinite-series relaxation based on $\ell < 1$ cannot be used. This illustrates the difference between the two finite envelopes. The earlier covariance/good-mean argument also covers its TV convergence.

5 Sharpness and boundaries of the experiment

5.1 A classical square-root example

For $n = 1$, let μ be uniform on L singleton elements. Then $\tau = 1/L$, and P_H is an empirical categorical law. For uniform $q_i = 1/m$,

$$\mathbb{E}_H \operatorname{TV}(P_H, q) = \frac{m}{2L} \mathbb{E} \left| \operatorname{Bin}(L, 1/m) - L/m \right|.$$

For fixed $m \geq 2$, or for growing m with $L/m \to \infty$, the classical binomial central limit theorem gives

$$\mathbb{E}_H \operatorname{TV}(P_H, q) \sim \sqrt{\frac{m-1}{2\pi L}}.$$

Indeed the standardized binomial converges to a standard normal and its absolute value is uniformly integrable by its unit second moment. This old empirical-distribution example shows why a general $O(\sqrt{(m-1)\tau})$ TV order cannot be replaced by $O((m-1)\tau)$. The new criterion's chi-square coefficient one is exact already in degree one of the overlap expansion.

5.2 Rare colors and adaptive laws

For binary $q = (1-\varepsilon, \varepsilon)$, compare both histogram laws with the point mass on the all-first-color histogram. The union bound and the triangle inequality give, for every fixed-size source law,

$$\mathbb{E}_H \operatorname{TV}(P_H, \operatorname{Mult}(n, q)) \le 2n\varepsilon.$$

For example, a deterministic selected subset has $\tau = 1$, but its expected TV tends to zero if $n\varepsilon \to 0$. This example explains the rarity factor in the binary criterion (9); the chi-square criterion retains no such factor.

Independence of the source law from the coloring also matters. For example, color a $2n$-element ground set with binary fair colors, and then select deterministically n elements of a most numerous color. For each realized coloring this deterministic law has upper cylinders, but its histogram is a point mass at an endpoint and is far from $\operatorname{Bin}(n, 1/2)$. The iid coloring/independent-source joint calculation in Lemma 3 is absent. Our conclusions are for each fixed law, not simultaneous control over laws chosen after seeing H. They concern absolute distributional error.

6 Reproducibility and assistance

The accompanying standard-library Python programs perform finite exact rational controls. The first enumerates all colorings for twelve small fixtures, including spanning trees and a source without negative cylinders, and checks the overlap identity against direct histogram calculation. A separate control enumerates binary colorings and checks the exact fourth-moment identity and rarity-sensitive lower bound, including sources without negative cylinders. The collision control checks six negative-cylinder collision fixtures and four forced-element families against exact factorial envelopes. Their acceptance checks remain active under Python's optimization flag. These computations illustrate the finite identities and examples; the general proofs above do not depend on a finite test or a numerical limit. Build and execution instructions are in the edition-specific source package: Edition 2 source package.

A note on the use of AI

OpenAI ChatGPT and Codex assisted mathematical exploration, literature comparison, drafting, calculation, and proof and code checks. The author assumes responsibility for all content.

References

[1] P. Diaconis and R. C. Griffiths, Reproducing kernel orthogonal polynomials on the multinomial distribution (2019). arXiv:1802.07122, version 3.

[2] W.-L. Loh, Stein's method and multinomial approximation, *Annals of Applied Probability* **2**(3) (1992), 536–554. doi:10.1214/aoap/1177005648.

[3] S. G. Bobkov, Generalized symmetric polynomials and an approximate de Finetti representation, *Journal of Theoretical Probability* **18** (2005), 399–412. doi:10.1007/s10959-005-3509-6.

[4] B. Roos, On Bobkov's approximate de Finetti representation via approximation of permanents of complex rectangular matrices (2014). arXiv:1203.2871, version 2.

[5] Y. Han and J. Niles-Weed, Approximate independence of permutation mixtures, *Annals of Statistics* **54**(3) (2026), 1594–1621. doi:10.1214/26-AOS2619; arXiv:2408.09341, version 3.

[6] Y. Han and J. Niles-Weed, Elementary symmetric polynomial inequalities for centered vectors and matrices (2026). arXiv:2607.23836.

[7] S. Schubr, Moment certificates for homogeneous stable count laws, preprint, edition 6 (2026). https://research.luminaryoddities.com/moment-certificates-homogeneous-stable-count-laws/v/6.

[8] J. Borcea, P. Brändén and T. M. Liggett, Negative dependence and the geometry of polynomials, *Journal of the American Mathematical Society* **22** (2009), 521–567. arXiv:0707.2340.

[9] C. Chekuri, J. Vondrák and R. Zenklusen, Dependent randomized rounding for matroid polytopes and applications. arXiv:0909.4348, version 2 (2009).