THE GEOMETRY OF RANDOMIZED SMOOTHING ON FEASIBLE SETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized smoothing certifies the probability of a fixed output event as the center of Gaussian noise moves. Feasibility or confidence filtering reports label probabilities only among retained proposals, producing a ratio. Its numerator is a fixed Gaussian event mass, while its denominator is the probability of retention and can change with the center. Substituting this ratio into the ordinary smoothing formula can therefore certify a ball that contains a decision boundary. We separate the problem into a geometric question and a certification question. Geometry determines when conditioning preserves Gaussian comparisons. Convex retained sets preserve the full comparison, while general sets require geometric control of the retained law as the center moves. Without such control, conditional probabilities imply no positive universal radius. Joint retention-and-label probabilities always yield a valid certificate for the same filtered predictor. A uniform covariance bound transfers divergence certificates to the retained law and can yield larger radii even when the Gaussian event comparison fails. Both methods admit finite-sample bounds. For a learned image classifier with a training-selected nonconvex filter, conditional Rényi bounds certify more images than joint-mass bounds without additional model evaluations. A released confidence filter exhibits verified label changes inside radii obtained by conditional substitution. An application of adaptive Gaussian composition covers causal finite-horizon executions with history-dependent center shifts under a pathwise energy bound.

1 INTRODUCTION

Randomized smoothing turns predictions under Gaussian noise into a robustness certificate (Cohen et al., 2019). Now consider a system that discards nine out of ten noisy inputs. If one label receives nine out of ten surviving votes, the reported probability is 0.9. Yet, the event that an input is both retained and assigned that label may have a Gaussian mass of only 0.09. The standard smoothing proof controls the event mass, not the fraction among survivors.

Let $s_y(a)$ be the Gaussian probability that a proposal centered at a is retained and assigned label y. Let $\zeta_\sigma(a)$ be the probability of retention. The filtered vote is $p_y(a) = s_y(a)/\zeta_\sigma(a)$. Dividing all label masses by the same value preserves their order at a given center. It does not preserve the comparison between two centers because $\zeta_\sigma(a)$ changes with a. A large majority among retained proposals can therefore carry little Gaussian mass and can change label inside the radius obtained by substituting $p_y(a)$ into the ordinary formula.

This observation leads to two questions. When does conditioning preserve the Gaussian comparison, and how can the filtered predictor be certified when it does not? The retained distribution answers the first question. Its one-dimensional projections determine whether event probabilities obey the Gaussian comparison. The curvature of log retention probability controls KL and Rényi divergence, while conditional covariance describes the same behavior locally. Convex retained sets satisfy the strongest condition and preserve the full event comparison.

The second question has two practical answers. Joint retained-label masses are probabilities of fixed Gaussian events, so they always produce a sound certificate without changing the filtered predictor or adding model evaluations. A uniform covariance bound instead permits direct certification from conditional probabilities through the Rényi bound of Li et al. (2019). This can yield a larger radius

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.

without requiring the full Gaussian event comparison. Without either geometric control or joint masses, the conditional label probabilities alone cannot determine a positive universal radius.

We test both conclusions on a learned image model. A training-selected nonconvex filter on CIFAR-10.2 admits a global covariance proof, which permits a direct comparison between covariance and joint-mass certificates. A separate AuditVotes case study finds independently verified label changes inside radii computed from conditional probabilities. The first study shows how geometric control recovers a direct certificate. The second exhibits the failure when that control is absent. Our derivations combine truncated exponential-family identities, strong log-concavity, and rejection-aware smoothing (Nielsen, 2022; Gopi et al., 2022; Sheikholeslami et al., 2022).

Contributions.

- We characterize when conditioning retains Gaussian event and divergence comparisons. The criteria cover arbitrary fixed measurable retained sets and identify conditional covariance as the exact local quantity.
- We prove that conditional probabilities alone provide no positive universal radius. Algorithm 1 gives a finite-sample certificate for the unchanged filtered predictor from joint masses. A uniform covariance bound transfers divergence certificates to the retained law, with an explicit spatial requirement when the bound is local.
- Exact counterexamples and two learned-model studies demonstrate both failure and recovery. A finite-horizon extension covers history-dependent center shifts chosen before each Gaussian proposal block under one common causal execution.

The single-step results compare centers under the same measurable retention and task rules. Appendix F treats fixed post-processing alternatives. Appendix I develops the causal finite-horizon result and Algorithm 2 gives its finite-sample procedure.

2 Filtered Gaussian smoothing

Fix a noise scale $\sigma > 0$ and a finite-dimensional affine space H with direction space $L = H - H$. Let I_L be the identity on L and let $G_a = \mathcal{N}_H(a, \sigma^2 I_L)$ be the Gaussian proposal centered at $a \in H$. The retention rule $R : H \to \{0, 1\}$ and task rule $f : H \to \mathcal{Y}$ are measurable and fixed, where $\mathcal{Y}$ is finite and $|\mathcal{Y}| \geq 2$. Define

$$K = \{z : R(z) = 1\}, \qquad E_y = K \cap \{z : f(z) = y\}.$$

Assume $G_a(K) > 0$. This holds at every finite center once it holds at one center. Gaussian measures with equal nonsingular covariance are mutually absolutely continuous.

The joint retained-label mass, acceptance probability, and conditional label probability are

$$s_y(a) = G_a(E_y), \qquad \zeta_\sigma(a) = G_a(K) = \sum_y s_y(a), \qquad p_y(a) = \frac{s_y(a)}{\zeta_\sigma(a)}.$$

The filtered classifier uses one fixed tie rule and satisfies

$$g(a) = \arg\max_y p_y(a) = \arg\max_y s_y(a). \tag{1}$$

Thus the denominator affects the numerical probabilities but not their order.

Write Φ for the standard normal CDF. The Gaussian event comparison states that every measurable $E \subseteq H$ and every $a, b \in H$ satisfy

$$\Phi\left(\Phi^{-1}(G_a(E)) - \frac{\|a-b\|}{\sigma}\right) \leq G_b(E) \leq \Phi\left(\Phi^{-1}(G_a(E)) + \frac{\|a-b\|}{\sigma}\right),$$

using the extended Gaussian quantiles at zero and one. A certificate of radius r at a is invalid if the smoothed decision changes at some b with $\|b - a\| < r$. In particular, if A and B are the top and runner-up labels at a, the *substituted conditional radius* is

$$r_{\mathrm{cond}}(a) = \frac{\sigma}{2}\left[\Phi^{-1}(p_A(a)) - \Phi^{-1}(p_B(a))\right]_+.$$

Here $[t]_+ = \max\{t, 0\}$.

Whole-set conditioning gives $Q_a = G_a(\cdot \mid K)$, whose label probabilities are $p_y(a)$. The Gaussian event comparison applies to the joint mass $s_y(a)$, not directly to this conditional ratio. By contrast, a projection, classifier, or randomized repair applied after the Gaussian draw acts identically at both centers. The comparison therefore carries through by data processing. Appendix F formalizes this distinction.

The perturbation changes only the center a. For a learned center map $a_\theta(x)$, an observation-space certificate also needs a verified bound on $\|a_\theta(x') - a_\theta(x)\|$. A support or filter that changes with x requires a separate analysis.

3 Occupancy controls conditional Gaussian geometry

Define the Gaussian occupancy

$$\zeta_\sigma(a) = G_a(K), \qquad \ell(a) = \log \zeta_\sigma(a).$$

For centers $a, b \in \mathrm{aff}(K)$, conditioned KL differs from ordinary Gaussian KL by the exact first-order Taylor remainder of log occupancy.

$$D_{\mathrm{KL}}(Q_a \| Q_b) = \frac{\|a-b\|^2}{2\sigma^2} + \ell(b) - \ell(a) - \langle \nabla \ell(a), b - a \rangle. \tag{2}$$

The same exponential-family calculation gives a three-point correction for finite-order Rényi divergence. Write D_α for order-α Rényi divergence, with $D_1 = D_{\mathrm{KL}}$. Appendix A gives the calculation.

Theorem 1 (Global divergence-rate criterion)**.** *Assume K is measurable with positive intrinsic measure in $H = \mathrm{aff}(K)$. The following are equivalent. First, ℓ is concave on H. Second, $D_{\mathrm{KL}}(Q_a \| Q_b) \le \|a-b\|^2/(2\sigma^2)$ for every ordered pair. Third, $D_\alpha(Q_a \| Q_b) \le \alpha \|a-b\|^2/(2\sigma^2)$ for every ordered pair and every finite $\alpha > 0$, with $D_1 = D_{\mathrm{KL}}$.*

Concavity of ℓ characterizes the divergence bound. Preserving every Gaussian event comparison requires the stronger one-dimensional condition below.

Corollary 2 (Convex conditioning preserves the Gaussian comparison)**.** *If K is convex, then every measurable $E \subseteq K$ and every $a, b \in H$ satisfy*

$$\Phi\left(\Phi^{-1}(Q_a(E)) - \frac{\|a-b\|}{\sigma}\right) \le Q_b(E) \le \Phi\left(\Phi^{-1}(Q_a(E)) + \frac{\|a-b\|}{\sigma}\right).$$

Consequently r_{cond} is sound for every fixed convex retained set. The same conclusion holds when nominal conditional label probabilities are replaced by familywise-valid lower and upper bounds.

This is the strongly log-concave tradeoff comparison of Gopi et al. (2022, Theorem 13) specialized in the intrinsic affine hull. It is sharp because $K = H$ recovers the ordinary Gaussian comparison. The proof and scale calculation are in Appendix A.

For a general retained set, the full comparison still has an exact one-dimensional test.

Proposition 3 (Ordered-pair event criterion)**.** *Let $a \ne b$, $u = (b-a)/\|b-a\|$, $\rho = \|b-a\|/\sigma$, and*

$$X(z) = \langle u, z - a \rangle / \sigma, \qquad F_c(t) = Q_c\{X \le t\}.$$

Every measurable $E \subseteq K$ obeys the two-sided Gaussian event comparison from Q_a to Q_b if and only if, for every $t \in \mathbb{R}$,

$$\Phi^{-1}(F_b(t)) \ge \Phi^{-1}(F_a(t)) - \rho. \tag{3}$$

Extended quantiles cover zero and one.

The likelihood ratio dQ_b/dQ_a is increasing in X. Neyman–Pearson ordering therefore reduces all measurable events to the nested halfspaces in Proposition 3. The proof is in Appendix A.

The occupancy identities also yield a local diagnostic.

Proposition 4 (Integrated and local covariance control)**.** *Let $X_c \sim Q_c$ and write $C_c = \mathrm{Cov}_L(X_c)$. For $d = b - a$,*

$$D_{\mathrm{KL}}(Q_a \| Q_b) = \frac{1}{\sigma^4} \int_0^1 (1-t) d^\top C_{a+td} d \, dt. \tag{4}$$

If $\dim L \geq 1$, then for every unit $u \in L$ and every finite $\alpha > 0$, with $D_1 = D_{\mathrm{KL}}$,

$$\lim_{t \to 0} \frac{D_\alpha(Q_a \| Q_{a+tu})}{\alpha t^2/(2\sigma^2)} = \frac{\mathrm{Var}\langle u, X_a \rangle}{\sigma^2}.$$

Consequently the worst local divergence ratio is

$$\Lambda(a) = \lambda_{\max}(\mathrm{Cov}_L(X_a)) / \sigma^2.$$

If $\Lambda(a) > 1$, the ordinary Gaussian KL bound fails along a corresponding direction for all sufficiently small nonzero displacements. If $\lambda_{\max}(C_c) \leq \Lambda\sigma^2$ along the segment from a to b, then $D_{\mathrm{KL}}(Q_a \| Q_b) \leq \Lambda \|a - b\|^2/(2\sigma^2)$.

Appendix K.8 compares this local quantity with KL at finite displacement. A local ratio above one proves failure of the ordinary Gaussian KL bound. Certifying a decision radius additionally requires either locating a label change or bounding covariance throughout the candidate ball.

The conditional substitution. The Gaussian comparison controls the joint event probability $G_b(E_y)$ from the nominal mass $G_a(E_y)$. For the filtered classifier, the latter equals $\zeta_\sigma(a) p_y(a)$ rather than $p_y(a)$. The lower comparison would require $G_a(E_A) \geq \underline{p}_A$, while a conditional premise gives only $G_a(E_A) \geq \zeta_\sigma(a) \underline{p}_A$. Dividing perturbed joint masses by $\zeta_\sigma(b)$ preserves their order, but it cannot restore the missing nominal factor.

Proposition 5 (Confidence-filter counterexample)**.** *Let $H = \mathbb{R}$, $\sigma = 1$, and let the two logits at proposal u be $(0, (\log 9)u)$. Retain a proposal when its largest softmax probability is strictly above 0.9. Thus $K = (-\infty, -1) \cup (1, \infty)$, with label zero on the left tail and label one on the right tail. At center $a = -0.1$,*

$$p_0(a) = \frac{\Phi(-0.9)}{\Phi(-0.9) + \Phi(-1.1)} = 0.575680\ldots, \qquad r_{\mathrm{cond}}(a) = 0.190855\ldots.$$

The conditional decision boundary is at zero. Center $b = 0.01$ therefore has the opposite label although $|b - a| = 0.11 < r_{\mathrm{cond}}(a)$. The joint-mass radius in Theorem 7 equals 0.1, the exact distance from a to the boundary.

Proposition 6 (Compact finite-displacement witness)**.** *Let*

$$K = ([0,1] \times [0, \tfrac{1}{2}]) \cup ([0, \tfrac{1}{2}] \times [\tfrac{1}{2}, 1]), \qquad \sigma = \tfrac{3}{20},$$

and take $a = (13/20, 49/100)$ and $b = (66/100, 49/100)$, so $\|a - b\| = 1/100$. For $E = K \cap \{z_1 \leq 633/1000\}$, exact-rational outward interval propagation certifies

$$Q_a(E) > 0.52747658, \qquad Q_b(E) < 0.49888514, \qquad Q_a(E) - \Phi(1/15) > 9/10000.$$

It also certifies

$$D_{\mathrm{KL}}(Q_a \| Q_b) - \frac{\|a-b\|^2}{2\sigma^2} > \frac{1}{5000}, \qquad d_{\mathrm{TV}}(Q_a, Q_b) - \left(2\Phi\left(\frac{\|a-b\|}{2\sigma}\right) - 1\right) > \frac{1}{500}.$$

Consequently the binary rule $f = \mathbf{1}_E$ changes its conditioned label at b, although b lies strictly inside $r_{\mathrm{cond}}(a)$.

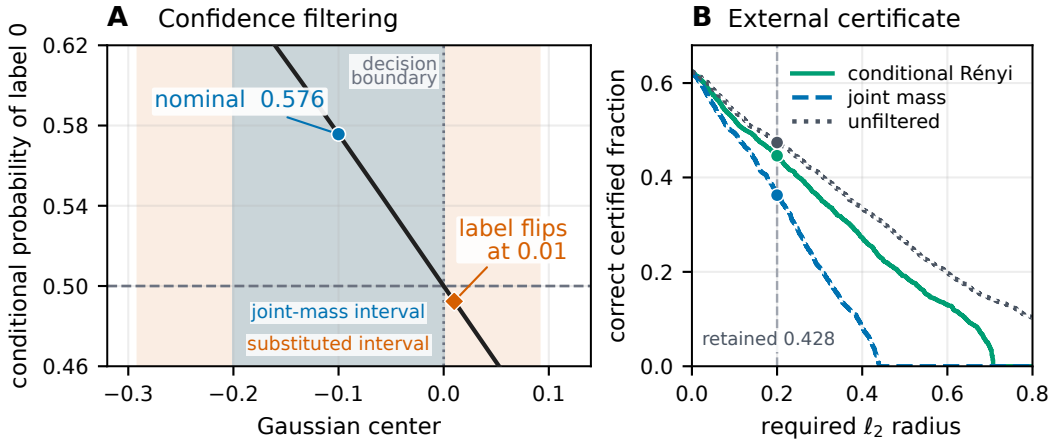

Figure 1: Normalization failure and geometry-controlled recovery. Panel A shows Proposition 5. The conditional label changes at zero. The substituted interval crosses this boundary, while the joint-mass interval ends at it. Panel B uses the projected-band filter selected before CIFAR-10.2 access. Curves give the fraction of all test images with the correct selected label and a positive valid radius at least as large as the abscissa. Conditional Rényi and joint mass certify the same filtered predictor. Unfiltered smoothing is a separate baseline. The vertical line marks radius 0.2.

Prékopa log-concavity also makes a convex support satisfy Theorem 1 (Prékopa, 1971). Convexity is sufficient but not necessary for log occupancy to be concave. The confidence filter and the connected L-shape show that simple fixed nonconvex supports need not satisfy either the substituted decision radius or the Gaussian divergence rate. The outward interval algorithm and certified enclosures are in Appendix D. Figure 3 shows the retained supports.

4 Certifying the unchanged filtered classifier

Equation (1) gives a direct correction. The conditional predictor can be left unchanged while its certificate is computed from the probabilities of the fixed Gaussian events E_y.

Theorem 7 (Joint retained-label certificate)**.** *Fix the retention rule, task rule, label set, and noise scale from Section 2. Let A be the selected label at center a. Suppose numbers L_A and U_y satisfy*

$$L_A \le s_A(a), \qquad s_y(a) \le U_y \quad \text{for every } y \ne A.$$

Set $U_B = \max_{y \ne A} U_y$ and set $r_{\mathrm{mass}}(a) = 0$ when $L_A \le U_B$. Otherwise, the conditional classifier is constant throughout the open ball of radius

$$r_{\mathrm{mass}}(a) = \frac{\sigma}{2}\left[\Phi^{-1}(L_A) - \Phi^{-1}(U_B)\right]. \tag{5}$$

At the population level one may take $L_A = s_A(a)$ and $U_B = \max_{y \ne A} s_y(a)$.

The proof applies the lower Gaussian comparison to E_A and the upper comparison to each E_y. Their bounds remain strictly ordered inside (5). Equation (1) then transfers the result to the conditional classifier. The full proof is in Appendix C.

Sheikholeslami et al. (2022) previously retained rejection as an output and certified joint selected-label and selected-or-rejection masses. If $L_{A\cup\perp}$ lower bounds the combined mass of the selected and rejected outputs, then $1 - L_{A\cup\perp}$ upper bounds every task competitor. Explicit competitor bounds can be smaller in multiclass problems. Neither finite-sample construction uniformly dominates because they divide the confidence budget differently. Algorithm 1 combines both bounds within one confidence event.

Proposition 8 (Sharpness and conditional-only impossibility)**.** *For every $0 < s_B < s_A$ with $s_A + s_B \le 1$, there is a one-dimensional fixed filter and binary task whose joint masses at the nominal center are s_A and s_B and whose nearest decision boundary is exactly $\frac{\sigma}{2}[\Phi^{-1}(s_A) - \Phi^{-1}(s_B)]$. Hence,*

no uniformly larger binary radius can be determined from these two masses alone. In contrast, for every $p \in (1/2, 1)$ and every $\varepsilon > 0$, there is a fixed filter with conditional top probability p whose decision boundary is less than ε away. Conditional probabilities and σ alone therefore give no positive universal radius over arbitrary fixed filters.

The sharp construction uses two oppositely oriented half-lines and rejects the interval between them. The impossibility construction moves the two retained tails outward while moving the nominal center toward their symmetry point. Appendix C gives both calculations.

Transferring divergence certificates through conditioning. A covariance bound controls divergences of the conditional law even when the Gaussian event comparison fails. It therefore permits direct use of the Rényi classification bound of Li et al. (2019). For $p > q > 0$ and $p + q \le 1$, define

$$C_\alpha(p, q) = -\log(1 - p - q + 2M_{1-\alpha}(p, q)), \qquad M_t(p, q) = \left(\frac{p^t + q^t}{2}\right)^{1/t},$$

with $M_0(p, q) = \sqrt{pq}$. Zero probabilities use continuous limits. This is the least reverse categorical Rényi divergence permitting a tie.

Proposition 9 (Conditional divergence certificates)**.** *Fix a and $R \in (0, \infty]$. Let $X_c \sim Q_c$ and write $C_c = \mathrm{Cov}_L(X_c)$. Suppose $\lambda_{\max}(C_c) \le \Lambda\sigma^2$ for every c with $\|c - a\| < R$, where $\Lambda > 0$. If $p_A(a) > p_B(a)$, the conditional decision is fixed throughout the open ball of radius*

$$r_{\mathrm{R}} = \sup_{\alpha \ge 1} \min\left\{\frac{R}{\alpha},\ \sigma\sqrt{\frac{2C_\alpha(p_A, p_B)}{\alpha\Lambda}}\right\}. \tag{6}$$

For any finite set of orders containing one, its maximum is also valid. For conditional confidence bounds, replace p_A, p_B by $L_A, \min\{U_B, 1 - L_A\}$, returning zero when these are not strictly ordered.

The covariance premise bounds $D_\alpha(Q_b \| Q_a)$ by $\alpha\Lambda\|b - a\|^2/(2\sigma^2)$. Orders above one require control as far as the tilted center $a + \alpha(b - a)$, which explains R/α. Covariance at a alone is insufficient. The order-one case uses reverse KL. It improves on the forward-KL radius $r_{\mathrm{cov}} = \min\{R, \sigma\sqrt{2\mathcal{J}(p_A, p_B)/\Lambda}\}$, where $\mathcal{J}(p, q) = p\log(2p/(p + q)) + q\log(2q/(p + q))$. Order selection reuses one confidence event and adds no model evaluations. Appendix C proves the transfer and the improvement.

Certified nonconvex recovery. Let $\sigma = 1$ and retain the two horizontal bands $K = \mathbb{R} \times ([-1, -0.9] \cup [0.9, 1])$, labeled by band. At $a = (0, -0.1)$ the nearest decision boundary is exactly 0.1 away. The law factorizes and Popoviciu gives $\mathrm{Cov}(Q_c) \preceq I$ at every center. Thus $\Lambda = 1$ is certified globally although K is nonconvex and has infinite diameter. Outward rational intervals certify

$$r_{\mathrm{mass}} = 0.04067\ldots < r_{\mathrm{cov}} = 0.09470\ldots < 0.1 < r_{\mathrm{cond}} = 0.11888\ldots.$$

This example shows that the ordinary Gaussian KL bound can hold even when the full Gaussian event comparison fails. Even the forward-KL certificate is larger than the joint-mass certificate while remaining below the exact decision boundary. The construction extends to a unit vector u learned on separate data and then fixed in the filter $R(x) = \mathbf{1}\{\alpha \le |u^\top x + b| \le \beta\}$, where $0 < \alpha < \beta \le \sigma$. The same global covariance bound holds. Appendix E gives the calculation and this extension, and Figure 4 gives a continuous sweep.

Write $\mathrm{CP}_{\mathrm{lower}}(c, n, \eta)$ and $\mathrm{CP}_{\mathrm{upper}}(c, n, \eta)$ for one-sided Clopper–Pearson endpoints with tail failure probability η (Clopper & Pearson, 1934).

Algorithm 1 Filtered Gaussian certification with familywise competitor bounds

Fixed inputs center a, scale σ, filter R, task rule f, labels $\mathcal{Y}$, tie rule
Statistical inputs selection size N_0, estimation size N, error δ

Draw $Z_i^{(0)} \overset{\text{iid}}{\sim} G_a$ for $i = 1, \dots, N_0$
$C_y^{(0)} \leftarrow \sum_i \mathbf{1}\{R(Z_i^{(0)}) = 1,\ f(Z_i^{(0)}) = y\}$ for every $y \in \mathcal{Y}$
Select $\widehat{A}$ from $(C_y^{(0)})_y$ using the fixed tie rule
Draw a fresh batch $Z_i \overset{\text{iid}}{\sim} G_a$ for $i = 1, \dots, N$
$C_y \leftarrow \sum_i \mathbf{1}\{R(Z_i) = 1,\ f(Z_i) = y\}$ for every $y \in \mathcal{Y}$
$C_\perp \leftarrow N - \sum_{y \in \mathcal{Y}} C_y$
$L_{\widehat{A}} \leftarrow \mathrm{CP}_{\mathrm{lower}}(C_{\widehat{A}}, N, \delta/3)$
$L_{\widehat{A}\cup\perp} \leftarrow \mathrm{CP}_{\mathrm{lower}}(C_{\widehat{A}} + C_\perp, N, \delta/3)$
for $y \in \mathcal{Y} \setminus \{\widehat{A}\}$ **do**
    $U_y \leftarrow \mathrm{CP}_{\mathrm{upper}}(C_y, N, \delta/[3(|\mathcal{Y}| - 1)])$
$U_B \leftarrow \min\{\max_{y \neq \widehat{A}} U_y,\ 1 - L_{\widehat{A}\cup\perp}\}$
if $L_{\widehat{A}} \leq U_B$ **then**
    return abstain
$r \leftarrow \frac{\sigma}{2}[\Phi^{-1}(L_{\widehat{A}}) - \Phi^{-1}(U_B)]$
return $(\widehat{A}, r)$

With probability at least $1 - \delta$, the bounds hold jointly for each input. The algorithm computes two valid upper bounds on the strongest competitor and uses the smaller one. One comes from explicit competitor counts and the other from the selected-or-rejected mass (Sheikholeslami et al., 2022). Rejected proposals count toward the common denominator N, so both bounds use the same model evaluations. Normalizing by the retained count instead estimates conditional probabilities and does not support the ordinary Gaussian radius. A single-batch variant bounds every label before selection.

Which comparison applies. A fixed classifier, projection, or randomized repair is common post-processing and uses the ordinary event certificate. Convex conditioning uses the same certificate by Corollary 2. Log-concave occupancy gives the KL and Rényi rate in Theorem 1, but not by itself the event comparison. A certified uniform covariance bound gives r_{R}. Arbitrary fixed conditioning uses joint retained-label masses. Appendix I gives the corresponding finite-horizon comparison for predictable Gaussian blocks under one common causal program.

5 Experiments

We evaluate a geometry-controlled certificate on a learned predictor, then measure normalization error and locate individual conditional-label changes for a different released filter.

Global covariance certificate on a learned model. Using CIFAR-10 training images, we selected a unit direction and offset for the nonconvex filter $\alpha \leq |\langle u, z\rangle + b| \leq \beta$ with $0 < \alpha < \beta = \sigma = 0.25$. This support satisfies the analytic bound $\mathrm{Cov}(Q_c) \preceq \sigma^2 I$ at every center. We evaluated the resulting filter on all $2{,}000$ CIFAR-10.2 images. All images, including abstentions, remain in each reported denominator. The joint confidence bounds have familywise error at most 0.001.

At radius 0.2, conditional Rényi, joint-mass, and unfiltered smoothing correctly certify 892, 725, and 948 images. The original forward-KL analysis certifies 829. At radius 0.5, Rényi certifies 381 images, where forward KL and joint mass certify none and unfiltered smoothing certifies 529. The Rényi calculation is a reanalysis of the same saved counts and confidence bounds, with no additional model evaluations. On 877 correctly classified images, its lower radius bound exceeds the joint-mass upper radius bound by at least 0.01. Figure 1B gives the curves. Appendix K.9 reports the original experiment, reanalysis, and a matched-model-call comparison. Unfiltered smoothing remains stronger at these radii, including with matched model calls.

Released image model. We use the AuditVotes Gaussian image code, released CIFAR-10 (Krizhevsky, 2009) ResNet-110 checkpoint, and $\sigma = 0.25$ (Lai et al., 2026). The filter retains

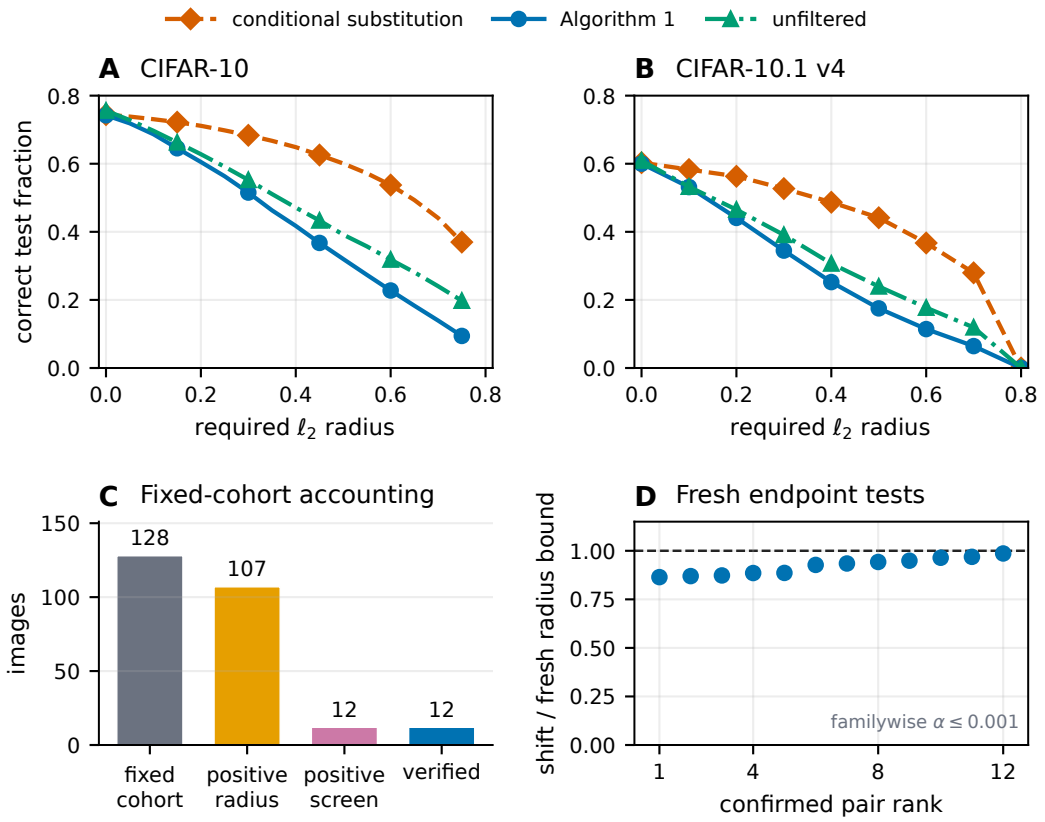

Figure 2: Released-model evaluation and fixed-cohort attack. Panels A and B use all CIFAR-10 and CIFAR-10.1 v4 images with $N_0 = 100$, $N = 10{,}000$, and $\delta = 0.001$. The diagnostic conditional-substitution curve shares proposal batches with the certified calculations. Panel C includes all 128 deterministically selected images. Panel D divides twelve independently confirmed shifts by familywise lower bounds on their substituted radii. Every ratio is below one, its endpoint labels differ, and familywise error is at most 0.001.

a proposal when its largest softmax probability exceeds 0.9. We evaluate all 10,000 CIFAR-10 test images and all 2,021 CIFAR-10.1 v4 images (Recht et al., 2019). A deterministic protocol specified the latter dataset, full evaluation, and attack cohort before model inference. Each image uses independent batches of $N_0 = 100$ proposals for label selection and $N = 10{,}000$ for estimation at $\delta = 0.001$. Appendix K.2 gives provenance and full results.

The released certificate uses conditional label probabilities as the nominal inputs to the Gaussian halfspace formula. Its implementation estimates those probabilities using the retained count as the binomial denominator (Lai et al., 2026).

For each image, every method uses the same selection and estimation batches. The released one-sided calculation uses $\sigma[\Phi^{-1}(\underline{p}_A)]_+$, where $\underline{p}_A$ is a retained-count binomial lower bound. The explicit-competitor calculation keeps N as the denominator and applies Theorem 7. The rejection-complement calculation is due to Sheikholeslami et al. (2022). Algorithm 1 combines both runner bounds in one familywise confidence event. Each curve reports the fraction of all images with the correct selected label and radius greater than ε. The joint-mass and unfiltered curves report certified accuracy. The conditional-substitution curve is diagnostic.

At radius 0.5, Algorithm 1 certifies 32.02% of CIFAR-10 and 17.52% of CIFAR-10.1 images. Unfiltered smoothing certifies 39.17% and 23.95%. The conditional substitution reports 60.07% and 44.14%. Mean retention is 49.34% and 35.11%. A nine-threshold study selects no confidence

filtering on development data and confirms the choice on untouched indices. Appendices K.1– K.5 give the complete calculations.

The endpoint study reports all 128 images in a cohort selected without model output. Fresh independent samples verify opposite population conditional labels in twelve endpoint pairs. Every displacement is smaller than its familywise lower radius bound, with displacement-to-bound ratios from 0.864 to 0.986. Nine cases occur among the 80 correctly classified images with a positive conditional-substitution radius. Appendix K.3 gives the search, joint inference, and every confirmed case.

Interpreting the comparisons. Figure 1B compares two sound certificates for one projected-band predictor under a global covariance proof. Figure 2 studies a distinct confidence filter, for which no covariance premise is asserted.

Algorithm 1 reuses the conditional calculation's model evaluations. Rejected proposals remain in the fixed denominator. Sampling to a fixed number of acceptances requires stopping-rule-aware inference.

Sequential extension. We also test the finite-horizon result on SafetyPointGoal2 (Ji et al., 2023). Across three PPO-Lagrangian controllers and two proposal settings per controller, the lower confidence bound on cost avoidance remains above 0.5 for every center-shift sequence with cumulative Euclidean shift budget $B \leq 0.02$. This certificate concerns avoidance of positive native cost until goal or timeout. It does not establish task completion, observation robustness, or forward invariance.

6 Related work

Gaussian smoothing certifies fixed output events under isotropic noise (Lécuyer et al., 2019; Li et al., 2019; Cohen et al., 2019). Likelihood-ratio methods cover broader noise laws (Dvijotham et al., 2020; Yang et al., 2020). Rejection-aware smoothing treats rejection as an output (Sheikholeslami et al., 2022; Däubener et al., 2024). We retain that joint-event foundation and study probabilities conditioned on acceptance.

Projection is fixed post-processing (Pfrommer et al., 2023), while known-domain methods certify bounded inputs (Kou et al., 2022; Voracek & Hein, 2023). Neither divides by a center-dependent acceptance probability. Input-dependent smoothing changes the noise law (Súkeník et al., 2022), whereas Double Sampling keeps its primary law fixed (Li et al., 2022).

Truncated exponential-family identities yield our occupancy calculation (Nielsen & Nock, 2011; Nielsen, 2022). Work on bounded Gaussian mechanisms studies related normalizers (Chen & Hale, 2022; Fu et al., 2023; Hu et al., 2024). Action-constrained learning maps, masks, projects, or truncates actions (Brahmanage et al., 2023; Hung et al., 2025; Stolz et al., 2024; Lee et al., 2025; Stolz et al., 2026; Chzhen & Donti, 2026). Policy Smoothing certifies adaptive trajectories (Kumar et al., 2022), and Adaptive Randomized Smoothing certifies multi-step computation (Lyu et al., 2024). Our trajectory result specializes fully adaptive Gaussian composition (Smith & Thakurta, 2022; Koskela et al., 2023) to correlated proposal blocks and a common causal feasibility program.

7 Limitations and conclusion

The single-step guarantees require unchanged retention and task rules. Observation perturbations also require a verified map to Gaussian centers. The learned-model studies use one checkpoint and noise scale. The CIFAR-10.2 experiment evaluates one training-selected filter. The AuditVotes $12/128$ result concerns a selected cohort, not dataset prevalence.Outward arithmetic validates the CIFAR-10.2 count-to-radius calculations, not noise generation or model execution (Voráček & Hein, 2023). The trajectory theorem certifies cost avoidance under proposal-center shifts, not task completion, observation robustness, or forward invariance.

Filtering changes a Gaussian event probability into a ratio whose denominator depends on the center. Convex retention controls this denominator strongly enough to preserve the ordinary event comparison. General nonconvex retention requires a verified geometric condition or joint retained-label events. The resulting certificates leave the predictor unchanged.

Reproducibility. The appendix and supplement provide proofs, protocols, statistical definitions, code, tests, configurations, and figure sources.

AI DISCLOSURE

Our use of generative AI was limited to language editing and manuscript consistency checks. The mathematical arguments and experimental implementation were developed by the authors independently of these tools. The authors take responsibility for the final text, claims, results, and code.

REFERENCES

Janaka Brahmanage, Jiajing Ling, and Akshat Kumar. FlowPG: Action-constrained policy gradient with normalizing flows. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20118–20132, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3fd9fe8ec6d7238bf71784797399bb61-Abstract-Conference.html.

Bo Chen and Matthew Hale. The bounded gaussian mechanism for differential privacy, 2022. URL https://arxiv.org/abs/2211.17230.

Maria Chzhen and Priya L. Donti. Improving feasibility via fast autoencoder-based projections. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7991a1ef0ff5f56d2e0ac8134dfe68d0-Abstract-Conference.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Sina Däubener, Kira Maag, David Krueger, and Asja Fischer. Integrating uncertainty quantification into randomized smoothing based robustness guarantees, 2024. URL https://arxiv.org/abs/2410.20432.

Krishnamurthy Dvijotham, Jamie Hayes, Borja Balle, J. Zico Kolter, Chongli Qin, András György, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlKrkSFPH.

Jie Fu, Zhiyu Sun, Haitao Liu, and Zhili Chen. Truncated laplace and gaussian mechanisms of RDP, 2023. URL https://arxiv.org/abs/2309.12647.

Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1948–1989, 2022. URL https://proceedings.mlr.press/v178/gopi22a.html.

Shengyuan Hu, Saeed Mahloujifar, Virginia Smith, Kamalika Chaudhuri, and Chuan Guo. Privacy amplification for the gaussian mechanism via bounded support, 2024. URL https://arxiv.org/abs/2403.05598.

Wei Hung, Shao-Hua Sun, and Ping-Chun Hsieh. Efficient action-constrained reinforcement learning via acceptance-rejection method and augmented MDPs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/34b70ece5f8d273fd670a17e2248d034-Abstract-Conference.html.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Antti Koskela, Marlon Tobaben, and Antti Honkela. Individual privacy accounting with gaussian differential privacy. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JmC_Tld3v-f.

Yiwen Kou, Qinyuan Zheng, and Yisen Wang. Certified adversarial robustness under the bounded support set. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11559–11597, 2022. URL https://proceedings.mlr.press/v162/kou22a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=mwdfai8NBrJ.

Yuni Lai, Yulin Zhu, Yixuan Sun, Yulun Wu, Bin Xiao, Gaolei Li, Jianhua Li, Qi Xie, and Kai Zhou. AuditVotes: Elevating provable defense for GNNs with efficient augmentation and conditional smoothing. In *ACM SIGSAC Conference on Computer and Communications Security*, 2026. doi: 10.1145/3830454.3832651. URL https://arxiv.org/abs/2503.22998.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy*, pp. 656–672, 2019. doi: 10.1109/SP.2019.00044.

Ganghun Lee, Minji Kim, Minsu Lee, and Byoung-Tak Zhang. Truncated gaussian policy for debiased continuous control. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (17):18071–18081, 2025. doi: 10.1609/aaai.v39i17.33988.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/335cd1b90bfa4ee70b39d08a4ae0cf2d-Paper.pdf.

Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13163–13208, 2022. URL https://proceedings.mlr.press/v162/li22aa.html.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Puya Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. URL https://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4260. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f1e7c90552850afcc2558d78950c519d-Abstract-Conference.html.

Frank Nielsen. Statistical divergences between densities of truncated exponential families with nested supports: Duo Bregman and duo Jensen divergences. *Entropy*, 24(3):421, 2022. doi: 10.3390/e24030421.

Frank Nielsen and Richard Nock. On Rényi and Tsallis entropies and divergences for exponential families, 2011. URL https://arxiv.org/abs/1105.3259.

Samuel Pfrommer, Brendon G. Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=FObkvLwNSo.

András Prékopa. Logarithmic concave measures with application to stochastic programming. *Acta Scientiarum Mathematicarum*, 32(3–4):301–316, 1971. URL https://acta.bibl.u-szeged.hu/14319/.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Fatemeh Sheikholeslami, Wan-Yi Lin, Jan Hendrik Metzen, Huan Zhang, and J. Zico Kolter. Denoised smoothing with sample rejection for robustifying pretrained classifiers. In *Workshop on Trustworthy and Socially Responsible Machine Learning at NeurIPS*, 2022. URL https://openreview.net/forum?id=i1lF1WqMw3j.

Adam Smith and Abhradeep Thakurta. Fully adaptive composition for gaussian differential privacy, 2022. URL https://arxiv.org/abs/2210.17520.

Roland Stolz, Hanna Krasowski, Jakob Thumm, Michael Eichelbeck, Philipp Gassert, and Matthias Althoff. Excluding the irrelevant: Focusing reinforcement learning through continuous action masking. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Roland Stolz, Michael Eichelbeck, and Matthias Althoff. Improving stochastic action-constrained reinforcement learning via truncated distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25617–25626, 2026. doi: 10.1609/aaai.v40i30.39758.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20697–20743, 2022. URL https://proceedings.mlr.press/v162/sukeni-k22a.html.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback–leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Vaclav Voracek and Matthias Hein. Improving ℓ_1-certified robustness via randomized smoothing by leveraging box constraints. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35198–35222, 2023. URL https://proceedings.mlr.press/v202/voracek23a.html.

Václav Voráček and Matthias Hein. Sound randomized smoothing in floating-point arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HaHCoGcpV9.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10693–10705, 2020. URL https://proceedings.mlr.press/v119/yang20c.html.

A OCCUPANCY IDENTITIES AND GLOBAL CHARACTERIZATION

This appendix gives proofs and supporting arguments for the analytic statements and the provenance of the finite nonconvex witness. All measures and derivatives are intrinsic to the affine hull introduced in Section 2. No ambient density is used when the feasible set is lower dimensional.

Write λ_H for intrinsic Lebesgue measure on H and

$$\varphi_{\sigma,a}(z) = (2\pi\sigma^2)^{-\dim L/2} \exp\left(-\frac{\|z-a\|^2}{2\sigma^2}\right)$$

for the intrinsic Gaussian density. Thus $\zeta_\sigma(a) = \int_K \varphi_{\sigma,a}\, d\lambda_H$ and Q_a has density $q_a(z) = \mathbf{1}_K(z)\varphi_{\sigma,a}(z)/\zeta_\sigma(a)$. Positive intrinsic measure and strict positivity of the Gaussian density imply $\zeta_\sigma(a) > 0$ for every $a \in H$. Let $m_a = \mathbb{E}_{Q_a} X_a$ and $C_a = \mathrm{Cov}_L(X_a)$.

Lemma 10 (Differentiation envelope)**.** *Fix a compact set of centers $U \subset H$. For each derivative order k, every kth center derivative of $\varphi_{\sigma,a}(z)$, uniformly over $a \in U$, is bounded in absolute value by an integrable polynomial times a Gaussian density with a variance strictly larger than σ^2. Consequently ζ_σ and $\ell = \log \zeta_\sigma$ are smooth on H, and their derivatives may be taken under the integral sign.*

Proof. Each center derivative is a degree-k polynomial in $(z-a)/\sigma^2$ times $\varphi_{\sigma,a}(z)$. Fix $a_\star \in H$ and r with $U \subseteq \{a : \|a - a_\star\| \le r\}$. If $\sigma' > \sigma$, completing the square shows that, uniformly on this ball,

$$\varphi_{\sigma,a}(z) \le c\, \varphi_{\sigma',a_\star}(z)$$

for a finite $c = c(\sigma, \sigma', r)$. The polynomial factor is bounded by a polynomial in $\|z - a_\star\| + r$, and every polynomial is integrable against the wider Gaussian. Restricting the integral to measurable K preserves the bound. Differentiation under the integral follows at every order. Continuity and positivity of ζ_σ then give smoothness of its logarithm. □

Differentiating once and twice gives the identities

$$\nabla \ell(a) = \frac{m_a - a}{\sigma^2}, \qquad \nabla^2 \ell(a) = \frac{C_a}{\sigma^4} - \frac{I_L}{\sigma^2}. \tag{7}$$

Indeed,

$$\nabla \zeta_\sigma(a) = \frac{1}{\sigma^2} \int_K (z-a)\varphi_{\sigma,a}(z)\, d\lambda_H(z),$$

and division by $\zeta_\sigma(a)$ yields the gradient. A second derivative gives

$$\frac{\nabla^2 \zeta_\sigma(a)}{\zeta_\sigma(a)} = \frac{\mathbb{E}_{Q_a}[(X_a - a)(X_a - a)^\top]}{\sigma^4} - \frac{I_L}{\sigma^2}.$$

Subtracting $\nabla\ell(a)\nabla\ell(a)^\top$ yields the covariance form.

For completeness, the common-support likelihood ratio is integrable and equals, Q_a-almost surely,

$$\log \frac{q_a(z)}{q_b(z)} = \frac{\|z-b\|^2 - \|z-a\|^2}{2\sigma^2} + \ell(b) - \ell(a).$$

Taking its Q_a expectation, expanding the two squares, and using the first identity in (7) proves (2).

For $\alpha \in (0,\infty) \setminus \{1\}$, the same common support gives

$$\int_K q_a(z)^\alpha q_b(z)^{1-\alpha}\, d\lambda_H(z) = \frac{1}{\zeta_\sigma(a)^\alpha \zeta_\sigma(b)^{1-\alpha}} \int_K \varphi_{\sigma,a}(z)^\alpha \varphi_{\sigma,b}(z)^{1-\alpha}\, d\lambda_H(z).$$

The square-completion identity

$$\alpha\|z-a\|^2 + (1-\alpha)\|z-b\|^2 = \|z - (\alpha a + (1-\alpha)b)\|^2 + \alpha(1-\alpha)\|a-b\|^2$$

holds for every real α. Hence the integral is finite for every finite positive order, including $\alpha > 1$, and

$$D_\alpha(Q_a \| Q_b) = \frac{\alpha\|a-b\|^2}{2\sigma^2} + \frac{\ell(\alpha a + (1-\alpha)b) - \alpha\ell(a) - (1-\alpha)\ell(b)}{\alpha - 1}. \tag{8}$$

Proof of Theorem 1. Suppose first that ℓ is concave. Its supporting-hyperplane inequality makes the correction in (2) nonpositive, proving the KL bound. If $0 < \alpha < 1$, concavity makes the numerator in (8) nonnegative while $\alpha - 1 < 0$. If $\alpha > 1$, put $c = \alpha a + (1-\alpha)b$ and observe that

$$a = \frac{1}{\alpha} c + \left(1 - \frac{1}{\alpha}\right) b.$$

Concavity now makes the same numerator nonpositive while $\alpha - 1 > 0$. Thus the Rényi correction is nonpositive at every finite positive order.

Conversely, the KL bound and (2) imply

$$\ell(b) \le \ell(a) + \langle \nabla \ell(a), b - a \rangle \qquad \text{for every } a, b \in H.$$

This is the first-order characterization of concavity for the smooth function ℓ. Finally, the all-order statement contains its declared order-one KL case, so it implies the KL statement. This proves all three equivalences. □

Proof of Corollary 2. Use orthonormal coordinates on H. On the convex set K, the density of Q_c is proportional to $\exp(-F_c)$, where $F_c(z) = \|z - c\|^2/(2\sigma^2)$. Both F_a and F_b are σ^{-2}-strongly convex. Their difference is affine with Lipschitz constant $\|a - b\|/\sigma^2$. The strongly log-concave tradeoff comparison of Gopi et al. (2022, Theorem 13) therefore dominates the tradeoff curve of two unit Gaussians separated by $(\|a - b\|/\sigma^2)/\sqrt{\sigma^{-2}} = \|a - b\|/\sigma$. Applying its lower bound to E gives the left inequality. Applying the same bound to E^c gives the right inequality. The usual top-versus-runner argument then gives r_{cond} and its simultaneous-bound version. The zero-dimensional case is immediate. □

Proof of Proposition 3. On the common support K,

$$\frac{dQ_b}{dQ_a}(z) = C \exp(\rho X(z))$$

for a positive constant C. The likelihood ratio is therefore increasing in X. The projection X has a continuous law under Q_a because K has positive intrinsic measure. For a prescribed value $p \in (0, 1)$, the Neyman–Pearson lemma shows that the event with Q_a mass p and smallest Q_b mass is $\{X \le F_a^{-1}(p)\}$. Thus the lower Gaussian event comparison holds for every event exactly when

$$F_b(F_a^{-1}(p)) \ge \Phi\big(\Phi^{-1}(p) - \rho\big) \qquad \text{for every } p \in (0, 1).$$

This is equivalent to (3). Applying the lower comparison to complements gives the upper comparison. Mutual absolute continuity supplies the endpoint cases. □

B Integrated and local covariance identities

Proof of Proposition 4. First write $d = b - a$. The second-order integral remainder gives

$$\ell(b) - \ell(a) - \langle \nabla \ell(a), d \rangle = \int_0^1 (1 - t) d^\top \nabla^2 \ell(a + td) d \, dt.$$

Combining this identity with (2) and (7) yields

$$D_{\mathrm{KL}}(Q_a \| Q_b) = \int_0^1 (1 - t) d^\top \left(\frac{I_L}{\sigma^2} + \nabla^2 \ell(a + td) \right) d \, dt = \frac{1}{\sigma^4} \int_0^1 (1 - t) d^\top C_{a+td} d \, dt.$$

The segmentwise eigenvalue bound follows immediately because $\int_0^1 (1 - t) \, dt = 1/2$.

Fix $a \in H$, a unit $u \in L$, and set $b = a + tu$. Lemma 10 permits a second-order Taylor expansion of ℓ around a. For $\alpha \ne 1$, the third center in (8) is $\alpha a + (1-\alpha)b = a + (1-\alpha)tu$. The constant and linear Taylor terms cancel, while the quadratic occupancy numerator is

$$\frac{t^2}{2}\big((1-\alpha)^2 - (1-\alpha)\big)\langle u, \nabla^2 \ell(a) u \rangle + o(t^2) = \frac{\alpha(\alpha - 1)t^2}{2} \langle u, \nabla^2 \ell(a) u \rangle + o(t^2).$$

Substitution in (8) gives

$$D_\alpha(Q_a \| Q_{a+tu}) = \frac{\alpha t^2}{2\sigma^2}\left(1 + \sigma^2\langle u, \nabla^2 \ell(a) u\rangle\right) + o(t^2).$$

For $\alpha = 1$, the same expansion follows directly from (2). Equation (7) identifies the factor in parentheses as

$$1 + \sigma^2\langle u, \nabla^2 \ell(a) u\rangle = \frac{\langle u, C_a u\rangle}{\sigma^2} = \frac{\mathrm{Var}\langle u, X_a\rangle}{\sigma^2}.$$

Dividing proves the claimed limit. Maximizing this Rayleigh quotient over unit vectors gives $\lambda_{\max}(C_a)/\sigma^2$. If that value is strictly larger than one, the expansion and the definition of $o(t^2)$ make the corresponding divergence strictly larger than its Gaussian comparator for every sufficiently small nonzero t. □

C JOINT-MASS PROOFS AND GEOMETRY-CONTROLLED RADII

Proof of Theorem 7. If $L_A \le U_B$, the stated open ball is empty. Assume $L_A > U_B$. Let $d = \|a - b\|$. The Gaussian event comparison applied to E_A gives

$$s_A(b) \ge \Phi\big(\Phi^{-1}(s_A(a)) - d/\sigma\big) \ge \Phi\big(\Phi^{-1}(L_A) - d/\sigma\big).$$

For every $y \ne A$, the other side of the comparison gives

$$s_y(b) \le \Phi\big(\Phi^{-1}(s_y(a)) + d/\sigma\big) \le \Phi\big(\Phi^{-1}(U_B) + d/\sigma\big).$$

The first displayed lower bound is strictly larger than the second displayed upper bound whenever

$$d < \frac{\sigma}{2}\left[\Phi^{-1}(L_A) - \Phi^{-1}(U_B)\right].$$

Thus $s_A(b) > s_y(b)$ for every competitor. Division by the common positive acceptance probability $\zeta_\sigma(b)$ preserves these inequalities, so the conditional classifier also selects A. □

Proof of Proposition 8. First fix $0 < s_B < s_A$ with $s_A + s_B \le 1$ and write $q_A = \Phi^{-1}(s_A)$ and $q_B = \Phi^{-1}(s_B)$. At nominal center zero, let

$$E_A = (-\infty, \sigma q_A], \qquad E_B = [-\sigma q_B, \infty).$$

The condition on the masses makes these half-lines disjoint. The interval between them is rejected. At center t their joint masses are

$$s_A(t) = \Phi(q_A - t/\sigma), \qquad s_B(t) = \Phi(q_B + t/\sigma).$$

They are equal first at $t = \frac{\sigma}{2}(q_A - q_B)$, exactly the population radius in Theorem 7. This proves binary sharpness from the two masses.

For the second assertion, fix $p \in (1/2, 1)$. For $T > 0$, retain the symmetric tails

$$K_T = (-\infty, -T] \cup [T, \infty)$$

and assign one label to each tail. At center $-d$, the conditional probability of the left label is

$$F_T(d) = \frac{\Phi((d - T)/\sigma)}{\Phi((d - T)/\sigma) + \Phi(-(T + d)/\sigma)}.$$

The function is continuous and strictly increasing from $F_T(0) = 1/2$ toward one, so there is a unique $d_T > 0$ with $F_T(d_T) = p$. Put

$$c_* = \frac{\sigma^2}{2}\log\frac{p}{1 - p}.$$

Gaussian tail asymptotics, applied at $d = c/T$, give uniformly for c in any fixed bounded interval

$$\log\frac{\Phi((c/T - T)/\sigma)}{\Phi(-(T + c/T)/\sigma)} = \frac{2c}{\sigma^2} + o(1).$$

For every $\eta \in (0, c_*)$, the last display is eventually smaller than $\log\{p/(1 - p)\}$ at $c = c_* - \eta$ and larger at $c = c_* + \eta$. Monotonicity therefore gives

$$c_* - \eta < T d_T < c_* + \eta$$

for all sufficiently large T. Hence $T d_T \to c_*$ and $d_T \to 0$. Symmetry places the decision boundary at zero. Taking T large enough makes its distance from the nominal center smaller than any prescribed ε while keeping the same conditional probability p. □

Proof of Proposition 9. Let $d = b - a$ and let $\psi(c) = \|c\|^2/(2\sigma^2) + \log \zeta_\sigma(c)$, using coordinates on the affine hull. Log-normalizer differentiation gives $\nabla^2\psi(c) = \mathrm{Cov}_L(X_c)/\sigma^4$. Write $g(t) = \psi(a + td)$. Whenever the complete segment from a to $a + \alpha d$ lies in the stated ball,

$$g''(t) \le v, \qquad v = \Lambda\|d\|^2/\sigma^2.$$

The exponential-family identity gives, for $\alpha > 1$,

$$D_\alpha(Q_b\|Q_a) = \frac{g(\alpha) - \alpha g(1) + (\alpha - 1)g(0)}{\alpha - 1} \le \frac{\alpha v}{2}.$$

For the inequality, $g(t) - vt^2/2$ is concave, so its value at one is at least the interpolation of its values at zero and α. The limit at one is $D_{\mathrm{KL}}(Q_b\|Q_a) = g(0) - g(1) + g'(1) \le v/2$. Thus the order-one bound requires only the segment to b, while order $\alpha > 1$ requires $\alpha\|d\| < R$.

For completeness, consider nominal category probabilities (p, q, s), where $s = 1 - p - q$ and $p > q$. Among distributions whose second category ties or exceeds the first, the reverse Rényi divergence is minimized on the tie. This follows from convexity of $\sum_j x_j^\alpha p_j^{1-\alpha}$ and the location of its unconstrained minimum. Setting the tied probabilities to t leaves the objective

$$t^\alpha(p^{1-\alpha} + q^{1-\alpha}) + (1 - 2t)^\alpha s^{1-\alpha}.$$

Put $m = M_{1-\alpha}(p, q)$ and $S = 2m + s$. The minimizer is $(m/S, m/S, s/S)$, and the objective is $S^{1-\alpha}$. The minimum divergence is therefore $-\log S = C_\alpha(p, q)$, as in Li et al. (2019, Lemma 1). At order one, the same calculation uses $m = \sqrt{pq}$ and gives the reverse-KL minimum. Boundary probabilities follow by continuity. Data processing to the selected label, any competitor, and their complement now rules out a tie whenever $\alpha\Lambda\|d\|^2/(2\sigma^2) < C_\alpha(p_A, p_B)$. This proves each radius in (6). Their supremum certifies the union of these concentric open balls.

For $\alpha \ge 1$ and $p > q > 0$, differentiating $S = 1 - p - q + 2m$ gives

$$\partial_p S = -1 + (m/p)^\alpha < 0, \qquad \partial_q S = -1 + (m/q)^\alpha > 0.$$

Thus C_α increases with p and decreases with q. The same upper bound on the strongest competitor covers every smaller competitor. Replacing the probabilities by their stated confidence bounds is conservative. Every order is a deterministic function of the same bounds, so maximizing over orders requires no additional probability statement. □

Why reverse KL improves the forward bound. Let $s = p + q$ and $t = (p - q)/s$. Then

$$C_1(p, q) = -\log(1 - s + s\sqrt{1 - t^2}) \ge s(1 - \sqrt{1 - t^2}) \ge \mathcal{J}(p, q).$$

The first inequality uses $-\log(1-x) \ge x$. For the second, the derivative of $1-\sqrt{1-t^2}$ is $t/\sqrt{1-t^2}$, which is at least $\mathrm{arctanh}(t)$, the derivative of $\{(1+t)\log(1+t) + (1-t)\log(1-t)\}/2$. Both vanish at zero. Consequently the order-one radius is never smaller than r_{cov} under the same covariance premise. The forward cost is at most $\log 2$, which imposes the ceiling $\sigma\sqrt{2\log 2/\Lambda}$. The reverse cost has no such ceiling as (p, q) approaches $(1, 0)$.

The categorical bound is an existing Rényi smoothing result. The conditional application here follows from the covariance identity, with the additional spatial requirement R/α for a local covariance certificate. It does not require the Gaussian event comparison for the retained law.

Forward-KL and bounded-support alternatives. The uniform covariance premise and (4) give

$$D_{\mathrm{KL}}(Q_a\|Q_b) \le \frac{\Lambda}{2\sigma^2}\|a - b\|^2$$

whenever the segment from a to b stays in the stated ball. Suppose a competitor y ties or overtakes A at b. Data processing under the three-cell partition formed by E_A, E_y, and their complement in K gives

$$D_{\mathrm{KL}}(Q_a\|Q_b) \ge \mathcal{J}(p_A(a), p_y(a)).$$

Indeed, among categorical laws in which the second cell ties or overtakes the first, forward KL from $(p, q, 1 - p - q)$ is minimized at $((p + q)/2, (p + q)/2, 1 - p - q)$. The attained value is $\mathcal{J}(p, q)$. For $p > q > 0$,

$$\partial_p\mathcal{J} = \log\frac{2p}{p + q} > 0, \qquad \partial_q\mathcal{J} = \log\frac{2q}{p + q} < 0.$$

Thus every competitor requires KL at least $\mathcal{J}(p_A(a), p_B(a))$. The covariance KL upper bound stays strictly below this value inside r_{cov}, which rules out a tie. The same monotonicity permits simultaneous bounds L_A and $\widetilde{U}_B = \min\{U_B, 1 - L_A\}$ in the displayed radius. The second term is also an upper bound because the conditional label probabilities sum to one. If $p_B = 0$, every competitor is null and Gaussian mutual absolute continuity makes the decision constant over all centers.

If K has diameter D, every scalar projection of a random variable supported on K has range at most D. Popoviciu's inequality therefore gives

$$\lambda_{\max}(C_c) \le D^2/4.$$

Taking $\Lambda = D^2/(4\sigma^2)$ gives the first of the two radii

$$r_{\mathrm{KL}} = \frac{2\sigma^2}{D}\sqrt{2\mathcal{J}(p_A, p_B)}, \qquad r_{\mathrm{odds}} = \frac{\sigma^2}{D}\log\frac{p_A}{p_B}. \tag{9}$$

For the remaining radius, the likelihood ratio dQ_b/dQ_a has logarithmic oscillation at most $D\|b - a\|/\sigma^2$ over K. Hence

$$\frac{Q_b(E_A)}{Q_b(E_B)} \ge \exp\left(-\frac{D\|b-a\|}{\sigma^2}\right)\frac{Q_a(E_A)}{Q_a(E_B)}.$$

The right side exceeds one inside r_{odds}. If $p_B = 0$, Gaussian mutual absolute continuity makes the competitor null at every center, giving the stated infinite value.

Finite-sample coverage. In Algorithm 1, the selection batch is independent of the estimation batch. Conditional on the selected label, its lower tail and the selected-or-rejected lower tail each fail with probability at most $\delta/3$. The union of the $|\mathcal{Y}| - 1$ explicit competitor upper-tail failures has probability at most $\delta/3$. When none fail, both runner upper bounds in Algorithm 1 are valid, so their minimum is valid and Theorem 7 applies. If one batch is used for selection and estimation, one may form all label lower bounds, all label-or-rejection lower bounds, and all label upper bounds before selection. Assigning total error $\delta/3$ to each family gives simultaneous coverage.

For the covariance-controlled radius, condition on a positive retained count. The retained label counts are multinomial with probabilities $(p_y(a))_y$. An independent selection batch or simultaneous conditional bounds therefore provides L_A and U_B with the stated coverage. The covariance upper bound needs its own simultaneous guarantee over the complete center ball. A sample covariance at the anchor does not supply that guarantee.

This statement treats binomial endpoints and Gaussian quantiles as real numbers. The empirical implementation uses SciPy to evaluate them. It therefore does not by itself address adversarial floating-point execution. A deployment requiring a machine-arithmetic guarantee must use validated probability and quantile routines, as in sound floating-point smoothing (Voráček & Hein, 2023). The compact witness in Appendix D uses separate outward rational intervals.

D Outward-rational certificate for the compact witness

We prove Proposition 6 without treating floating-point quadrature as exact. Write ϕ and Φ for the standard-normal density and CDF. For a rectangle $R = \prod_{k=1}^{d}[r_k^-, r_k^+]$, center c, and scale σ, define

$$\alpha_k = (r_k^- - c_k)/\sigma, \qquad \beta_k = (r_k^+ - c_k)/\sigma, \qquad p_k = \Phi(\beta_k) - \Phi(\alpha_k).$$

Gaussian factorization gives $G_c(R) = \prod_k p_k$. Because the two rectangles in the witness have disjoint interiors, their masses add exactly. The same factorization gives the unnormalized first-coordinate moment

$$[c_1 p_1 + \sigma\{\phi(\alpha_1) - \phi(\beta_1)\}]\prod_{k\ne 1} p_k.$$

These formulas yield both occupancies and the conditional mean used in the KL identity (2).

The corresponding unnormalized second moment of the first coordinate is

$$\left[c_1^2 p_1 + 2c_1\sigma\{\phi(\alpha_1) - \phi(\beta_1)\} + \sigma^2\{p_1 + \alpha_1\phi(\alpha_1) - \beta_1\phi(\beta_1)\}\right]\prod_{k\ne 1} p_k.$$

Adding this quantity over the two rectangles, dividing by occupancy, and subtracting the squared conditional mean gives the local variance ratio reported in the certified enclosure table below.

For $b = a + de_1$ with $d > 0$, the common-support likelihood ratio obeys

$$\log \frac{q_a(z)}{q_b(z)} = \frac{d}{\sigma^2}(x_* - z_1), \qquad x_* = \frac{a_1 + b_1}{2} + \frac{\sigma^2}{d} \log \frac{\zeta_\sigma(b)}{\zeta_\sigma(a)}.$$

Thus $q_a > q_b$ precisely to the left of x_*, up to a null hyperplane, and

$$d_{\mathrm{TV}}(Q_a, Q_b) = Q_a\{z_1 \le x_*\} - Q_b\{z_1 \le x_*\}.$$

All primitive evaluations are enclosed by rational intervals. We use the alternating series for e^{-y} and

$$\int_0^t e^{-x^2/2}\, dx = \sum_{n \ge 0} \frac{(-1)^n t^{2n+1}}{2^n n!(2n+1)},$$

after their terms decrease. Adjacent partial sums enclose the value. For $x > 0$, we enclose

$$\log x = 2 \sum_{k \ge 0} \frac{z^{2k+1}}{2k+1}, \qquad z = \frac{x-1}{x+1},$$

with tail at most $2|z|^{2N+3}/((2N+3)(1-z^2))$. Machin's identity and alternating arctangent series enclose π. Closed rational interval arithmetic then propagates these bounds through rectangle masses, normalization, logarithms, and event differences. Every divisor interval has a strictly positive lower endpoint.

At the rational data in Proposition 6, the resulting outward enclosures are

Quantity	certified enclosure
$D_{\mathrm{KL}}(Q_a \| Q_b)$	$[0.002423524,\ 0.002423525]$
Gaussian KL comparator	$1/450 = 0.002222\ldots$
$\mathrm{Var}_{Q_a}(Z_1)/\sigma^2$	$[1.09153866764843002220836 2,\ 1.09153866764843002220836 3]$
$d_{\mathrm{TV}}(Q_a, Q_b)$	$[0.028641422,\ 0.028641423]$
Gaussian TV comparator	$[0.026591227,\ 0.026591228]$
$Q_a(E)$	$[0.527476587,\ 0.527476588]$
$Q_b(E)$	$[0.498885133,\ 0.498885134]$
$Q_a(E) - \Phi(1/15)$	$[0.000900122,\ 0.000900123]$
x_*	$[0.623175082,\ 0.623175083]$

The KL excess is therefore greater than $0.0002013 > 1/5000$, and the TV excess is greater than $0.0020501 > 1/500$. The event bounds give opposite strict binary labels at a and b. Finally, $Q_a(E) > \Phi(1/15)$ implies

$$\sigma \Phi^{-1}(Q_a(E)) > \frac{3}{20}\frac{1}{15} = \frac{1}{100} = \|a - b\|,$$

so the changed label is strictly inside the imported ambient-Gaussian radius. This proves every assertion in Proposition 6.

The retained verifier evaluates the same rational series and records every remainder and propagated interval. A focused test checks the printed strict inequalities and fails if an endpoint is interchanged. Figure 3 shows the connected support and the certified displacement.

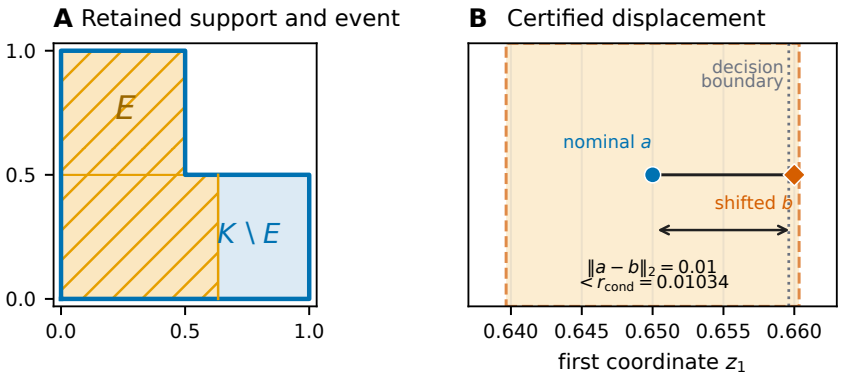

Figure 3: Compact nonconvex witness. Panel A shows the connected L-shaped support, the event E, and its complement $K \setminus E$. Panel B shows the tested displacement, the conditional decision boundary, and the substituted radius. The outward-certified bounds above establish the label change.

E A COVARIANCE-CONTROLLED NONCONVEX FAMILY

Let $L = L_0 \oplus L_1$ be an orthogonal decomposition and let $K = L_0 \times S$, where $S \subseteq L_1$ has positive measure and diameter at most 2σ. The conditioned Gaussian factors into an unrestricted Gaussian on L_0 and a Gaussian conditioned on S in L_1. For every center c, the covariance has blocks $\sigma^2 I_{L_0}$ and $\mathrm{Cov}(Y_c \mid Y_c \in S)$. Every unit projection of the second block has range at most $\mathrm{diam}(S)$, so Popoviciu's inequality gives $\mathrm{Cov}(Q_c) \preceq \sigma^2 I_L$ for every c. Thus the covariance radius in Proposition 9 applies globally with $\Lambda = 1$ even for a disconnected S. The result does not assert the Gaussian event comparison.

This construction gives a proof-carrying learned filter. Fix a unit vector u, an offset b, and $0 < \alpha < \beta \le \sigma$ after training, and set

$$R(x) = \mathbf{1}\{\alpha \le |u^\top x + b| \le \beta\}.$$

For $X_c \sim \mathcal{N}(c, \sigma^2 I)$, the component orthogonal to u is independent of $u^\top X_c + b$ and is unchanged by retention. The retained scalar lies in $[-\beta, \beta]$. Popoviciu's inequality therefore gives

$$\mathrm{Cov}(X_c \mid R(X_c) = 1) = \sigma^2(I - uu^\top) + \mathrm{Var}(u^\top X_c \mid R(X_c) = 1)uu^\top \preceq \sigma^2 I$$

for every center c. The direction and offset may be learned on separate data because the proof uses only their fixed values and $\|u\| = 1$.

The same statement holds for an orthonormal matrix $U = [u_1, \ldots, u_k]$ and the product rule $\prod_j \mathbf{1}\{\alpha_j \le |u_j^\top x + b_j| \le \beta_j\}$ with $\beta_j \le \sigma$. The Gaussian coordinates in the columns of U are independent, the retained event factorizes, and every retained coordinate has variance at most β_j^2. The orthogonal complement remains unchanged.

Let $H = \mathbb{R}^2$, $\sigma = 1$, and

$$K = \mathbb{R} \times \big([-1, -0.9] \cup [0.9, 1]\big).$$

The two horizontal bands receive different labels. For $a_\delta = (0, -\delta)$, reflection gives the lower-band and upper-band joint masses

$$s_-(\delta) = \Phi(1 - \delta) - \Phi(0.9 - \delta), \qquad s_+(\delta) = \Phi(1 + \delta) - \Phi(0.9 + \delta).$$

The lower band is selected for $\delta > 0$. Reflection makes the masses equal when the second center coordinate is zero. Pointwise Gaussian likelihood ordering makes their order strict on either side. The nearest decision boundary to a_δ is therefore exactly δ away.

Conditioning does not affect the first coordinate, so

$$\mathrm{Cov}(Q_c) = \mathrm{diag}(1, \mathrm{Var}(Y \mid Y \in [-1, -0.9] \cup [0.9, 1])) .$$

The retained second coordinate lies in $[-1, 1]$. Popoviciu's inequality gives variance at most one for every center. Hence $\lambda_{\max}(\mathrm{Cov}(Q_c)) \le 1$ globally. Proposition 9 applies with $R = \infty$ and $\Lambda = 1$. Notice that K has infinite diameter, so the two diameter-based formulas in (9) are unavailable.

This numerical witness uses the forward-KL alternative proved in Appendix C. Writing $p_- = s_-/(s_- + s_+)$ and $p_+ = 1 - p_-$, the compared radii are

$$r_{\mathrm{cond}} = \Phi^{-1}(p_-), \qquad r_{\mathrm{mass}} = \frac{1}{2}\left[\Phi^{-1}(s_-) - \Phi^{-1}(s_+)\right], \qquad r_{\mathrm{cov}} = \sqrt{2\mathcal{J}(p_-, p_+)}.$$

At $\delta = 0.1$, exact rational interval propagation gives

Quantity	Certified enclosure
$G_{a_\delta}(K)$	$[0.05078446622171157326,\ 0.05078446622171157327]$
$p_-(a_\delta)$	$[0.54731840865059743984,\ 0.54731840865059743985]$
r_{mass}	$[0.04067994841241465132,\ 0.04067994841241465133]$
r_{cov}	$[0.09470767664273030523,\ 0.09470767664273030524]$
r_{cond}	$[0.11888914386462333130,\ 0.11888914386462333131]$

The Rényi certificate is at least as large as r_{cov}, but it is not the quantity enclosed in this table. The strict ordering printed in the main text follows. Figure 4 traces the four quantities. The same covariance bound also gives the global KL and finite-order Rényi rates through Theorem 1. The substituted event radius nevertheless crosses the exact boundary. Thus Gaussian divergence contraction does not imply the Gaussian event comparison.

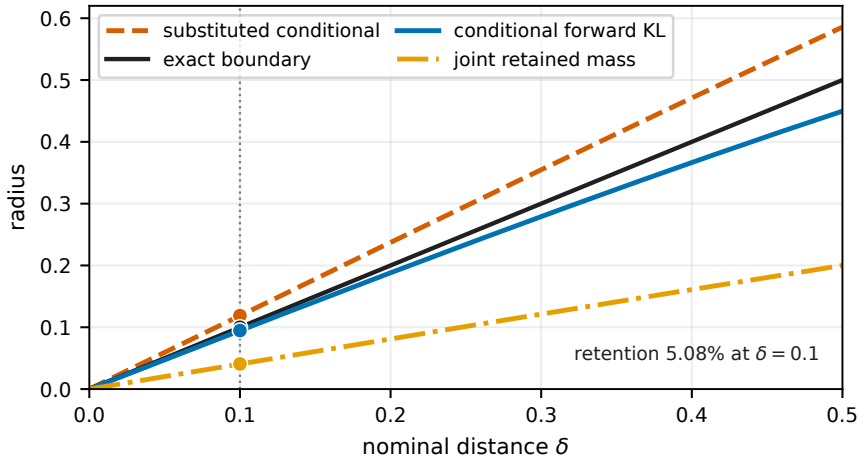

Figure 4: End-to-end radii for the nonconvex two-band family. The curves use the analytic Gaussian CDF formulas above. The points at $\delta = 0.1$ use the outward-certified calculation. The forward-KL radius remains close to the exact boundary and is more than twice the joint-mass radius at the certified point. The substituted conditional radius crosses the boundary.

The verifier uses exact fractions, alternating Gaussian-integral brackets, an atanh logarithm bound, and outward square roots. It records all interval endpoints and strict checks in the supplement. The displayed sweep uses standard floating-point Gaussian CDF evaluations and is illustrative. The certified point does not depend on those evaluations.

F Alternative feasible kernels

A common repair channel and a fixed-weight mixture change the sampling kernel rather than recertifying whole-set conditioning. They remain useful when the application permits that change.

Proposition 11 (Fixed randomized repair)**.** *Let $U \sim \rho$ be independent of the center and let $T : H \times \mathcal{U} \to K$ be one measurable map used unchanged at both centers. Then $P_a = T_{\#}(G_a \otimes \rho)$ satisfies*

the Gaussian event and total-variation comparisons and every finite-order Gaussian divergence bound. This includes a fixed categorical component followed by a component-specific projection.

Suppose $K = \bigcup_{j=1}^{m} K_j$, where the K_j are convex subsets of one affine space with positive intrinsic measure and every distinct pair has zero-intrinsic-measure intersection. All component-conditioned laws $Q_{a,j} = G_a(\cdot \mid K_j)$ use scale σ. For a center-independent probability vector w, define $P_a^w = \sum_j w_j Q_{a,j}$.

Theorem 12 (Fixed-weight comparison)**.** *For every finite $\alpha > 0$,*

$$D_\alpha(P_a^w \| P_b^w) \le \frac{\alpha \|a-b\|^2}{2\sigma^2}, \qquad D_1 = D_{\mathrm{KL}}.$$

The family also satisfies the Gaussian event comparison and

$$d_{\mathrm{TV}}(P_a^w, P_b^w) \le 2\Phi\left(\frac{\|a-b\|}{2\sigma}\right) - 1.$$

For a bounded common-affine partition and $\tau \in [0,1]$, let $\pi_j(a) = G_a(K_j)/G_a(K)$, fix anchor a_0, and define

$$w_j^{(\tau,a_0)}(a) = \frac{\pi_j(a_0)^{1-\tau}\pi_j(a)^\tau}{\sum_k \pi_k(a_0)^{1-\tau}\pi_k(a)^\tau}, \qquad \nu_a^{(\tau,a_0)} = \sum_j w_j^{(\tau,a_0)}(a) Q_{a,j}.$$

Proposition 13 (Anchored comparison)**.** *If $D_K = \sup_{x,y\in K} \|x-y\|$, then*

$$D_{\mathrm{KL}}\left(\nu_b^{(\tau,a_0)} \middle\| \nu_c^{(\tau,a_0)}\right) \le \left(\frac{1}{2\sigma^2} + \frac{\tau^2 D_K^2}{8\sigma^4}\right) \|b-c\|^2.$$

For $p_y^\nu(a) = \nu_a^{(\tau,a_0)}(f^{-1}(y))$, write the coefficient as C_τ. A top label A and runner-up B at a have the valid radius

$$r_{\mathrm{Pin}}^\nu(a,\tau,a_0) = \frac{p_A^\nu(a) - p_B^\nu(a)}{\sqrt{2C_\tau}}.$$

The endpoint $\tau = 0$ has fixed weights and the Gaussian event certificate. Positive temperatures use only the corrected bound above. The proofs are in Appendices G and H. Exact evaluation requires certified component occupancies, and positive temperatures also require a certified diameter. These quantities can be intractable for neural confidence regions. We therefore record the construction as an alternative for applications with an explicit convex partition rather than use it in the learned-model experiment.

G Fixed-weight and anchored comparisons

We first record the two component facts used below. If K_j is convex, Theorem 1 and Corollary 2 give, for every finite $\alpha > 0$,

$$D_\alpha(Q_{a,j} \| Q_{b,j}) \le \frac{\alpha \|a-b\|^2}{2\sigma^2}. \tag{10}$$

and, for every measurable E,

$$\Phi\big(\Phi^{-1}(Q_{a,j}(E)) - r\big) \le Q_{b,j}(E) \le \Phi\big(\Phi^{-1}(Q_{a,j}(E)) + r\big), \qquad r = \frac{\|a-b\|}{\sigma}. \tag{11}$$

The intrinsic scale in both statements is $r = \|a-b\|/\sigma$.

Proof of Theorem 12. Attach the component index to a sample and define the tagged laws

$$\widetilde{P}_a(j, dz) = w_j Q_{a,j}(dz), \qquad \widetilde{P}_b(j, dz) = w_j Q_{b,j}(dz).$$

For $\alpha \neq 1$, let $d_{\alpha,j} = D_\alpha(Q_{a,j} \| Q_{b,j})$. Directly from the definition,

$$\exp\Big((\alpha-1) D_\alpha(\widetilde{P}_a \| \widetilde{P}_b)\Big) = \sum_j w_j \exp((\alpha-1) d_{\alpha,j}).$$

Let $C_\alpha = \alpha\|a-b\|^2/(2\sigma^2)$. Equation (10) gives $d_{\alpha,j} \le C_\alpha$. For $\alpha > 1$, the exponential is increasing and division by $\alpha - 1$ keeps the resulting upper bound. For $0 < \alpha < 1$, the exponential is decreasing, so the sum is bounded below by $\exp((\alpha-1)C_\alpha)$. Division by the negative $\alpha - 1$ reverses the inequality and gives the same upper bound. At $\alpha = 1$, the tagged KL chain rule gives

$$D_{\mathrm{KL}}(\widetilde{P}_a\|\widetilde{P}_b) = \sum_j w_j D_{\mathrm{KL}}(Q_{a,j}\|Q_{b,j}) \le \frac{\|a-b\|^2}{2\sigma^2}.$$

Forgetting the tag is a measurable channel. Data processing for every positive Rényi order (van Erven & Harremoës, 2014) transfers the tagged bounds to P_a^w and P_b^w.

For the event statement, set $h_r(p) = \Phi(\Phi^{-1}(p) - r)$. For $r > 0$, writing $p = \Phi(x)$ gives

$$h_r'(p) = \exp(rx - r^2/2),$$

which is increasing in p. Hence h_r is convex. It is the identity when $r = 0$. Applying the lower half of (11), averaging, and using Jensen's inequality gives

$$P_b^w(E) \ge \sum_j w_j h_r(Q_{a,j}(E)) \ge h_r\left(\sum_j w_j Q_{a,j}(E)\right) = h_r(P_a^w(E)).$$

Applying this lower bound to E^c gives the upper event comparison. Finally, for $p = \Phi(x)$, both possible endpoint differences are maximized at the midpoint of the two shifted Gaussian means and equal $2\Phi(r/2) - 1$. Taking the supremum over events proves the total-variation bound. □

Proof of Proposition 13. Let $Z_j(x) = G_x(K_j) > 0$ and let $m_{x,j}$ be the mean of $Q_{x,j}$. From (7), applied to component K_j,

$$\nabla_x \log Z_j(x) = \frac{m_{x,j} - x}{\sigma^2}. \tag{12}$$

Because the common-affine union is bounded, conditional means lie in the closures of their convex components, and therefore $\|m_{x,j} - m_{x,k}\| \le D_K$.

Write $v_x(j) = w_j^{(\tau,a_0)}(x)$ for the anchored categorical law. For two centers b, c, define

$$\delta_j = \tau\big(\log Z_j(c) - \log Z_j(b)\big).$$

Terms common to every component cancel on normalization, so $v_c(j) = v_b(j)e^{\delta_j}/\mathbb{E}_{v_b}e^{\delta_J}$. Consequently

$$D_{\mathrm{KL}}(v_b\|v_c) = \log \mathbb{E}_{v_b} e^{\delta_J} - \mathbb{E}_{v_b}\delta_J. \tag{13}$$

For every j, k, integration of (12) along the segment from b to c gives

$$|(\delta_j - \delta_k)| \le \frac{\tau D_K}{\sigma^2}\|b - c\|.$$

Thus the range of the random variable δ_J has width at most the right-hand side. Hoeffding's lemma applied to (13) yields

$$D_{\mathrm{KL}}(v_b\|v_c) \le \frac{\tau^2 D_K^2}{8\sigma^4}\|b-c\|^2. \tag{14}$$

Attach the component tag to $\nu_b^{(\tau,a_0)}$ and $\nu_c^{(\tau,a_0)}$. The tagged KL chain rule, the convex-component order-one case of (10), and (14) give

$$D_{\mathrm{KL}}(\widetilde{\nu}_b\|\widetilde{\nu}_c) \le \left(\frac{1}{2\sigma^2} + \frac{\tau^2 D_K^2}{8\sigma^4}\right)\|b-c\|^2.$$

Forgetting the tag and applying KL data processing proves the asserted bound with coefficient C_τ.

Pinsker's inequality then gives, for any $a, b \in H$,

$$d_{\mathrm{TV}}(\nu_a^{(\tau,a_0)}, \nu_b^{(\tau,a_0)}) \le \sqrt{\frac{C_\tau}{2}}\,\|a-b\|.$$

Every fixed label event changes by at most this quantity. If A is top at a and $y \neq A$, then

$$p_A^\nu(b) - p_y^\nu(b) \geq p_A^\nu(a) - p_y^\nu(a) - \sqrt{2C_\tau}\,\|a - b\|.$$

Since $p_B^\nu(a)$ is the largest competing probability, every such difference is strictly positive whenever

$$\|a - b\| < \frac{p_A^\nu(a) - p_B^\nu(a)}{\sqrt{2C_\tau}}.$$

This proves the radius and independently checks its factor of two. □

H FIXED RANDOMIZED REPAIR IS A COMMON CHANNEL

Proof of Proposition 11. Let $U \sim \rho$ be auxiliary randomness whose law does not depend on the center, and define $M_a = G_a \otimes \rho$. The likelihood ratio between M_a and M_b depends only on the Gaussian coordinate. Hence

$$D_\alpha(M_a \| M_b) = D_\alpha(G_a \| G_b), \qquad d_{\mathrm{TV}}(M_a, M_b) = d_{\mathrm{TV}}(G_a, G_b)$$

for every finite positive order, with the order-one convention. Applying the same measurable map $T : H \times \mathcal{U} \to K$ at both centers and using data processing gives

$$D_\alpha(P_a \| P_b) \leq D_\alpha(G_a \| G_b) = \frac{\alpha \|a - b\|^2}{2\sigma^2}, \qquad 0 < \alpha < \infty,$$

with $D_1 = D_{\mathrm{KL}}$, and

$$d_{\mathrm{TV}}(P_a, P_b) \leq 2\Phi\left(\frac{\|a - b\|}{2\sigma}\right) - 1.$$

For completeness, the event statement also survives the auxiliary mixture. For $E \subseteq K$ and fixed u, put $A_u = \{z : T(z, u) \in E\}$ and $h_r(p) = \Phi(\Phi^{-1}(p) - r)$ with $r = \|a - b\|/\sigma$. The ordinary Gaussian event comparison gives $G_b(A_u) \geq h_r(G_a(A_u))$. The function h_r is convex, as shown in Appendix G. Averaging over u and applying Jensen yields

$$P_b(E) \geq \mathbb{E}_\rho h_r(G_a(A_U)) \geq h_r(\mathbb{E}_\rho G_a(A_U)) = h_r(P_a(E)).$$

Applying the lower bound to E^c gives the upper event comparison.

A deterministic repair is recovered when ρ is a point mass. A component-indexed repair takes U to be a categorical component index with a fixed center-independent law and lets $T(\cdot, U)$ be a component map fixed in advance, such as clipping or projection. This argument permits a state-specific channel T_s when the compared centers belong to the same fixed state fibre. It does not compare different states, different feasible sets, center-dependent auxiliary laws, or channels that are recomputed after the center perturbation. □

I ADAPTIVE FEASIBLE-TRAJECTORY COMPARISON

At step t, a fixed controller maps history h_t to $m_t(h_t)$. A causal shift rule chooses Δ_t before the current proposal block. Fix $N_t \geq 1$, $\sigma_t > 0$, and $0 \leq \rho_t < 1$. Let $\mathcal{F}_t$ contain the public pre-block history and any private seed used by the shift rule. Fresh variables $\xi_{t,0}, \ldots, \xi_{t,N_t}$ are conditionally independent standard Gaussians given $\mathcal{F}_t$ and are independent of $\mathcal{F}_t$. The proposals are

$$Z_{t,k} = m_t(h_t) + \Delta_t + \sigma_t\big(\sqrt{\rho_t}\xi_{t,0} + \sqrt{1 - \rho_t}\xi_{t,k}\big), \qquad \kappa_t = \frac{N_t}{1 + (N_t - 1)\rho_t}.$$

One fixed causal program receives the complete block and returns its first verified action, uses a finite fallback library, or abstains without a system transition.

Policy Smoothing gives Gaussian event comparisons for adaptive trajectory perturbations under a cumulative Euclidean budget (Kumar et al., 2022, Theorem 1). Adaptive Randomized Smoothing applies Gaussian differential privacy to adaptive multi-step computation (Lyu et al., 2024, Theorem 2.3). Below, the fully adaptive Gaussian composition result of Smith & Thakurta (2022) and Koskela et al. (2023) gives a comparison under a pathwise budget. Whitening computes the exact Mahalanobis cost of each correlated proposal block, and the common causal feasibility program preserves the comparison.

Theorem 14 (Adaptive feasible-trajectory comparison)**.** *Fix a finite horizon, a standard-Borel history space, and a Borel action subset of $\mathbb{R}^d$. Let P_0 and P_Δ be the public-trace laws obtained from the same measurable controller, verifier, fallback, dynamics, initial law, and stopping rule, with the requested shifts respectively suppressed and applied. History-dependent feasible sets are allowed. If the fresh Gaussian innovations underlying the current block are independent of the pre-block information and every reachable history and private shift seed obeys*

$$\sum_t \kappa_t \|\Delta_t\|_2^2/\sigma_t^2 \le \mu^2,$$

then every measurable public-trace event E obeys

$$\Phi\big(\Phi^{-1}(P_0(E)) - \mu\big) \le P_\Delta(E) \le \Phi\big(\Phi^{-1}(P_0(E)) + \mu\big).$$

Proof of Theorem 14. Let the action dimension be d. For step t, write

$$C_t = (1-\rho_t) I_{N_t} + \rho_t \mathbf{1}\mathbf{1}^\top.$$

The proposal block conditioned on the current history is Gaussian with covariance $\sigma_t^2 C_t \otimes I_d$. Since $0 \le \rho_t < 1$, this covariance is nonsingular. The nominal and perturbed conditional means differ by $\mathbf{1} \otimes \Delta_t$. Whitening gives squared displacement

$$\begin{aligned} r_t^2 &= \frac{1}{\sigma_t^2}(\mathbf{1}\otimes\Delta_t)^\top (C_t^{-1}\otimes I_d)(\mathbf{1}\otimes\Delta_t) \\ &= \frac{\mathbf{1}^\top C_t^{-1}\mathbf{1}}{\sigma_t^2}\|\Delta_t\|_2^2 = \frac{\kappa_t\|\Delta_t\|_2^2}{\sigma_t^2}. \end{aligned}$$

The last equality follows because $\mathbf{1}$ is an eigenvector of C_t with eigenvalue $1 + (N_t - 1)\rho_t$.

The shift rule chooses Δ_t from the previously visible history and any private randomness before the current Gaussian block is generated. Couple its private seed under both laws and include it in the latent filtration. Write u_t for the whitened mean displacement and X_t for the corresponding standard-normal block under the nominal law. Then u_t is predictable and $\|u_t\|^2 = r_t^2$. After the last block, append one unobserved scalar Gaussian with predictable displacement

$$u_* = \left(\mu^2 - \sum_t r_t^2\right)^{1/2}.$$

This quantity is real by the pathwise energy condition. The enlarged experiment has total squared displacement exactly μ^2.

Under its nominal law, define

$$M = \sum_t \langle u_t, X_t\rangle + u_* X_*.$$

Iterated conditional expectation of the Gaussian moment generating function gives, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}_0 \exp\left(\lambda M - \frac{\lambda^2\mu^2}{2}\right) = 1.$$

Thus $M \sim \mathcal{N}(0, \mu^2)$. The sequential Gaussian density ratio of the enlarged shifted experiment with respect to the enlarged nominal experiment is

$$\exp\left(M - \frac{\mu^2}{2}\right).$$

For any event of nominal probability p, Neyman–Pearson ordering places the smallest shifted probability on the lower tail of M and the largest on its upper tail. Direct evaluation of these two Gaussian tails gives

$$\Phi\big(\Phi^{-1}(p) - \mu\big) \quad \text{and} \quad \Phi\big(\Phi^{-1}(p) + \mu\big).$$

An event that ignores the appended scalar is an event of the enlarged experiment, so the same bounds hold before the scalar is appended. Forgetting the private seed is also common post-processing.

At a fixed generated history, proposal ordering, clipping, feasibility tests, fallback construction, dynamics, process randomness, and stopping use the same conditional kernel under both laws. Their causal composition is common post-processing of the Gaussian blocks. This remains true when the feasible action set changes with the generated state. It fails if the perturbation changes the verifier, fallback, dynamics, or feasible set through another input. After stopping, append zero-shift dummy blocks to the fixed horizon. Standard-Borel history and action spaces ensure the required regular conditional kernels exist.

For a measurable public-trace event E, the Gaussian tradeoff gives

$$\Phi\big(\Phi^{-1}(P_0(E)) - \mu\big) \le P_\Delta(E) \le \Phi\big(\Phi^{-1}(P_0(E)) + \mu\big)\,.$$

Applying the same comparison to all measurable events yields $d_{\mathrm{TV}}(P_0, P_\Delta) \le 2\Phi(\mu/2) - 1$. The enlarged likelihood-ratio experiment also has Rényi divergence $\alpha\mu^2/2$ in both directions at every order $\alpha > 1$. Therefore data processing gives

$$D_\alpha(P_\Delta \| P_0) \le \frac{\alpha\mu^2}{2}, \qquad D_\alpha(P_0 \| P_\Delta) \le \frac{\alpha\mu^2}{2}.$$

Letting $\alpha \downarrow 1$ gives both directed KL bounds $\mu^2/2$.

Suppose $\sigma_t = \sigma$ and $\kappa_t = \kappa$. Every perturbation with $\sum_t \|\Delta_t\|_2^2 \le B^2$ has $\mu \le \sqrt{\kappa}B/\sigma$. Replacing $P_0(E)$ by a simultaneous lower confidence bound p_L obtained from independent nominal trajectories and solving

$$\Phi\left(\Phi^{-1}(p_L) - \frac{\sqrt{\kappa}B}{\sigma}\right) \ge q$$

for a target probability $q \in (0,1)$ gives

$$B_{\mathrm{cert}} = \frac{\sigma}{\sqrt{\kappa}}\left[\Phi^{-1}(p_L) - \Phi^{-1}(q)\right]_+ .$$

When $p_L \ge q$, this radius inherits the coverage of p_L. When $p_L < q$, the target probability is not certified even at $B = 0$.

The public trace includes every externally visible selection index, fallback result, and abstention symbol. If abstention occurs, the stopping rule does not execute a physical action. An event requiring no abstention therefore charges finite-library exhaustion directly. The result bounds that event's probability. It does not prove that the feasible-action fibre is nonempty or that the physical dynamics are forward invariant. □

Executable certification procedure. Fix before sampling the horizon, event E, schedules (N_t, σ_t, ρ_t), controller, candidate map, verifier, ordered finite fallback rule $\mathcal{L}_t(h_t)$, transition kernel, and stopping rule. For $N_t > 1$, a desired multiplier $\lambda_t \in (1, N_t]$ can be set exactly by

$$\rho_t = \frac{N_t/\lambda_t - 1}{N_t - 1}, \qquad \kappa_t = \lambda_t.$$

For $N_t = 1$, use $\rho_t = 0$ and $\kappa_t = \lambda_t = 1$. An attacked comparison may choose Δ_t from the pre-block history, but must commit it before the current proposal block and cannot alter another program input. Use separate proposal, attack, and transition random streams. Do not expose the proposal or transition stream to the shift rule. Evaluate that rule on an isolated copy of the pre-block history. For each of M independent nominal trajectories, set $\Delta_t = 0$. At history h_t, compute $m_t(h_t)$ and draw the complete correlated proposal block. Apply the common candidate map and verifier in index order. If every proposal fails, inspect $\mathcal{L}_t(h_t)$ in its declared order. Record $\perp$ and stop without a system transition when the library is exhausted. Otherwise execute the selected verified action through the common transition kernel. Record the selected proposal or fallback index, every system outcome, and any abstention.

Let X_i indicate whether trace i satisfies E, and let p_L be a valid one-sided lower confidence bound for $P_0(E)$. Iid initial-law draws permit an exact Clopper–Pearson bound based on $\sum_i X_i$. A fixed finite cohort with one independent draw at each support point permits a Hoeffding bound for the uniform cohort average. For target $q \in (0,1)$, define

$$\mu_{\mathrm{cert}} = \Phi^{-1}(p_L) - \Phi^{-1}(q)$$

only when $p_L \geq q$. No radius is certified when $p_L < q$. For constant $\sigma_t = \sigma$ and $\kappa_t = \kappa$, return $B_{\mathrm{cert}} = \sigma\mu_{\mathrm{cert}}/\sqrt{\kappa}$. With varying schedules, retain the certified ellipsoid $\sum_t \kappa_t \|\Delta_t\|_2^2/\sigma_t^2 \leq \mu_{\mathrm{cert}}^2$. The event must be fixed independently of these nominal samples or covered by simultaneous inference. The energy condition must hold for every reachable history and attack seed. One observed trace cannot establish this condition. Algorithm 2 gives the complete certification procedure. The supplement includes tested implementations of the proposal block, totalized selector, finite trajectory rollout, energy calculation, and count-to-radius calculation.

Algorithm 2 Capped feasible rollout and trajectory-event certification

Fixed program horizon T, event E, schedules (N_t, σ_t, ρ_t), controller m_t
verifier V_t, ordered fallback library $\mathcal{L}_t$, dynamics, stopping rule
Statistical inputs trials M, valid lower-bound rule $\mathcal{B}$, error δ, target probability q

for independent nominal trials $i = 1, \ldots, M$ **do**
  for $t = 0, \ldots, T-1$ **do**
    Compute $m_t(h_t^{(i)})$ and draw the complete correlated proposal block
    Select the first proposal accepted by $V_t(h_t^{(i)}, \cdot)$
    If none is accepted, inspect $\mathcal{L}_t(h_t^{(i)})$ in its fixed order
    If no fallback is accepted, record $\perp$ and stop without a transition
    Otherwise execute the selected action, record the outcome, and update the history
$X_i \leftarrow \mathbf{1}\{\text{public trace } i \in E\}$ for every i
$p_L \leftarrow \mathcal{B}(X_1, \ldots, X_M, \delta)$
if $p_L < q$ **then**
  return no positive certificate
$\mu_{\mathrm{cert}} \leftarrow \Phi^{-1}(p_L) - \Phi^{-1}(q)$
return predictable shifts satisfying
  $\sum_t \kappa_t \|\Delta_t\|_2^2/\sigma_t^2 \leq \mu_{\mathrm{cert}}^2$ on every reachable history

I.1 Learned-controller evaluation

We trained three PPO-Lagrangian controllers for 10 million environment steps each and retained their terminal actor and observation-normalizer files. The controller seeds are 1120275774, 1346831933, and 658990255. Evaluation uses SafetyPointGoal2-v0 in Safety-Gymnasium 1.0.0 (Ji et al., 2023). One observation-only verifier restricts the two action coordinates. A deterministic clipped-controller action is used when no proposal passes. The verifier is a fixed heuristic and is not a native-cost oracle.

An earlier study used goal-before-cost as its event and did not certify probability 0.5. Those outcomes do not enter this calculation. A 16-seed pilot then informed the event below. The final 256-seed cohort was specified before evaluation. The audit reconstructs these seeds and verifies that they are disjoint from the earlier confirmation cohort and all 36 development labels.

The certified event is no positive native cost until the first goal or 500 executed steps. A positive cost, an early environment stop, or a nonfinite transition is failure. Reaching the goal stops the event, so no claim concerns later states. A timeout without cost is success. The target initial law is the uniform mixture over the 256 fixed reset seeds. We draw one independent Gaussian trajectory for each seed and arm. If S trajectories succeed, the simultaneous lower bound is

$$p_L = \left[\frac{S}{256} - \sqrt{\frac{\log(6/0.001)}{2 \cdot 256}} \right]_+ .$$

Hoeffding's inequality applies to the independent, nonidentical Bernoulli outcomes and targets the uniform-cohort average. It does not extend the result to other reset distributions.

The correlated-four arm has $N = 4$, $\sigma = 0.2$, $\rho = 59/63$, and $\kappa = 1.05$. The matched-one arm has $N = 1$, $\sigma = 0.2/\sqrt{1.05}$, $\rho = 0$, and $\kappa = 1$. Both have Gaussian energy charge $\kappa/\sigma^2 = 26.25$. The protocol records 0.2 and 1.05 as shared base values. The post-run integrity record distinguishes the literal schedules above. The executed schedules, energy calculations, and radii use the same charge.

At each step, the tested shift points opposite the clipped controller center and has norm $0.02/\sqrt{500}$. The empirical shifted counts in Table 1 are descriptive. The certificate uses the nominal lower

bound and Theorem 14, hence it covers every predictable proposal-center sequence with cumulative Euclidean budget at most $B = 0.02$ under the same causal program.

Table 1: SafetyPointGoal2 trajectory evaluation. Safe is the certified event count. Goal and timeout partition that count. Cost is the nominal positive-cost count. $L_{.02}$ is the familywise lower event probability at $B = 0.02$. Shift safe is the descriptive count under the tested shift. C1, C2, and C3 follow the controller-seed order in the text.

Ctrl	Proposals	Safe	Goal	Timeout	Cost	p_L	B_{cert}	$L_{.02}$	Shift safe
C1	one	211	78	133	45	.694	.0989	.657	213
C1	four	210	85	125	46	.690	.0968	.653	214
C2	one	187	60	127	69	.600	.0495	.560	180
C2	four	189	59	130	67	.608	.0535	.568	187
C3	one	205	107	98	51	.670	.0861	.633	204
C3	four	207	106	101	49	.678	.0903	.641	202

All six arms certify event probability above 0.5 at $B = 0.02$ with familywise error at most 0.001. Only 59–107 nominal trajectories per arm reach the goal before cost. Most certified successes are therefore timeouts. This result does not certify task completion, observation perturbations, changes to the verifier or dynamics, or forward invariance.

The raw result stores aggregate trajectory records rather than full per-step traces. A post-run audit checks the record schema, all $3{,}072$ arm keys, every reported summary, seed separation, finite energy, method schedules, and source closure. It cannot replay each selection from the stored aggregate records. The theorem's public trace is the arm-blind controller, selection, fallback, dynamics, and stopping output. Attack budgets and energy diagnostics are separate experimental metadata. The supplement includes the protocol, results, audit, checksums, and environment versions.

J Relation of the formal results to prior work

Table 2 identifies the earlier result used by each formal step and its application to conditional laws, their decision rules, and center-dependent occupancy.

Table 2: Result-level relation to prior work. The middle column gives the earlier statement used here. The last column gives the result established in this paper.

Current result	Used here	Established here
Occupancy divergences	Fixed-support Rényi log-partition identities and truncated-family KL and skew-Jensen identities (Nielsen & Nock, 2011; Nielsen, 2022)	Gaussian occupancy specialization followed by the equivalence between concave log occupancy and the stated ordered-pair KL and positive-order Rényi rates
Covariance control	Log-normalizer differentiation supplies standard moment and Fisher identities	Their specialization gives the finite-path KL integral, local Rényi rate, and top-covariance diagnostic used here
Corollary 2	Strongly log-concave tradeoff comparison of Gopi et al. (2022, Theorem 13)	Intrinsic affine-hull specialization with displacement scale $\|a - b\|/\sigma$
Proposition 3	Neyman–Pearson ordering for the monotone likelihood ratio between two conditioned Gaussian centers	Necessary and sufficient projected-mass inequalities for every measurable retained event under an arbitrary fixed set

Current result	Used here	Established here
Theorem 7 and Algorithm 1	Joint selected-label and selected-or-rejection masses (Sheikholeslami et al., 2022, Theorem 3.1)	Explicit competitor bounds and their simultaneous combination with the rejection-complement bound for arbitrary fixed filters
Limits of conditional votes	The cited joint-event certificate does not provide a radius from conditional votes alone	Sharpness of the joint-mass radius and the support-free conditional-only impossibility result
Proposition 9	Categorical Rényi reversal bound (Li et al., 2019, Lemma 1), data processing, Popoviciu's inequality, and likelihood-ratio oscillation	Transfer to conditional laws under a uniform covariance bound, with the R/α restriction when that bound holds only in a ball of radius R
Theorem 14	Adaptive trajectory and multi-step smoothing (Kumar et al., 2022; Lyu et al., 2024), and fully adaptive Gaussian composition (Smith & Thakurta, 2022; Koskela et al., 2023)	Exact Mahalanobis cost for correlated proposal blocks followed by one common causal selector, verifier, fallback, dynamics, and stopping program

K Experimental protocol and additional results

K.1 Released-model recertification

We audit Theorem 2 in Section 5.2 and its proof in Appendix A.2 of the retained AuditVotes v3 manuscript. The public source[1] is identified by its exact revision in the supplement. The supplement also authenticates the manuscript, released CIFAR-10 checkpoint, and inspected `core.py`. The upstream repository has no license file, so neither its source nor its checkpoint is redistributed.

The retained manuscript defines the image predictor by conditional label probabilities in Eq. (15). Its Theorem 2 uses bounds on those probabilities in the Gaussian radius, while Appendix A.2 applies the Gaussian halfspace comparison to joint retention-and-label indicators. In the released `core.py`, lines 97–103 of `Smooth._sample_noise` filter predictions by confidence before counting them. Lines 47–55 of `Smooth.certify` then replace the requested proposal count by the sum of retained class counts, compute a binomial lower bound with that denominator, and insert the bound into `sigma*norm.ppf`. The bound therefore estimates a conditional probability, whereas the joint indicator used in the halfspace comparison has nominal mass $\zeta_\sigma(a)p_y(a)$. This trace concerns the Gaussian image extension, not the separate graph certificate. The released multiclass routine uses the one-sided radius $\sigma[\Phi^{-1}(\underline{p}_A)]_+$, whereas r_{cond} in Section 2 is the top-versus-runner population expression. For $p_A(a) > 1/2$, the one-sided population radius is no larger than $r_{\mathrm{cond}}(a)$ because $p_B(a) \le 1 - p_A(a)$. All learned-model plots and endpoint tests use the released one-sided calculation.

We use PyTorch 2.3.1, torchvision 0.18.1, and one NVIDIA RTX 4070 Ti. The full test run uses 100 proposals per image in the selection phase and $10{,}000$ in the estimation phase, using batches of 1024. Random streams are derived from a declared base seed and image index. The run takes 7562 seconds. The supplement includes all $10{,}000$ per-image records and the analysis script. We do not call this an exact reproduction of the published table. The released code does not seed the PyTorch proposal stream and the original stream is unavailable. Instead, this is a seeded full-test evaluation of the released checkpoint and filtering rule under the published sample sizes.

The executed input path follows the pinned image code. CIFAR-10 test images are converted to tensors in $[0, 1]^d$. Isotropic noise is added in these pixel coordinates before the model's fixed channel normalization. The noisy tensor is not clipped. The ResNet logits determine the class by their first maximum, and a proposal is retained only when its largest softmax probability is strictly greater than 0.9. The audited source functions are `Smooth.certify` and `Smooth._sample_noise` in

[1] `https://github.com/Yuni-Lai/AuditVotes-Certified-Robustness`

the pinned `core.py`. Our evaluator reproduces that proposal law and tie rule while using explicit seeded streams and retaining both filtered and unfiltered counts.

Table 3 separates three sound finite-sample calculations. The explicit calculation reserves half of the error probability for the selected lower bound and half for simultaneous competitor upper bounds. The rejection-complement calculation reserves half for the selected lower bound and half for the selected-or-rejected lower bound. These two calculations are preselected separately. Algorithm 1 combines both runner bounds and assigns one third of the error probability to each confidence-event family.

Table 3: Additional full-test results. Positive is the fraction receiving a strictly positive radius. Median is over all test images. The remaining columns are correct fractions with radius strictly above the displayed value.

Calculation	Positive	Median	0.3	0.5	0.7
Conditional substitution	.8941	.6668	.6835	.6007	.4406
Joint complement	.8868	.3342	.5163	.3221	.1416
Joint explicit	.8872	.3293	.5114	.3137	.1323
Joint simultaneous	.8875	.3330	.5153	.3202	.1394
Unfiltered Gaussian	.9122	.3924	.5525	.3917	.2403

The selected-label accuracy is 0.7603 after filtering and 0.7867 without filtering. The mean retained fraction is 0.493386. All reported radius thresholds use a strict comparison.

The confidence statements above are mathematical real-arithmetic statements. The evaluation uses conventional SciPy beta and Gaussian quantiles. We record the numerical scope rather than claim adversarial floating-point soundness (Voráček & Hein, 2023). Every count file and post-processing script is retained so a validated quantile implementation can be substituted without reevaluating the network.

K.2 CIFAR-10.1 EVALUATION

The CIFAR-10.1 v4 evaluation protocol was specified before model inference. This version contains $2{,}021$ images and was the first version on which its creators evaluated classifiers (Recht et al., 2019). The supplement records the exact dataset revision and file checksums. The image and label files are not redistributed.

The protocol specifies the full evaluation, a hash-ranked attack cohort, the random streams, and all confirmation tests. The model, filter, noise scale, sample sizes, batching, and no-clipping proposal law match the CIFAR-10 evaluation. The run uses an independent base seed and took 1636 seconds on the same GPU with PyTorch 2.2.2 and torchvision 0.17.2. The supplement includes the protocol, implementation, per-image records, and summary.

Table 4: Full CIFAR-10.1 v4 results. Positive is the fraction receiving a strictly positive radius. Median is over all $2{,}021$ images. The remaining columns are correct fractions with radius strictly above the displayed value.

Calculation	Positive	Median	0.3	0.5	0.7
Conditional substitution	.8461	.5193	.5270	.4414	.2796
Joint simultaneous	.8362	.2122	.3449	.1752	.0643
Joint complement	.8362	.2125	.3459	.1757	.0663
Joint explicit	.8352	.2095	.3414	.1707	.0579
Unfiltered Gaussian	.8733	.2503	.3909	.2395	.1192

Table 4 reports the full calculation. The filtered and unfiltered selected-label accuracies are 0.6319 and 0.6502. The mean retained fraction is 0.351061. All methods use the same selection and estimation proposal batches within each image.

Table 5: Fresh confirmation for every positive screen in the fixed cohort. The last two columns give the declared winner's advantage over the other listed label as a percentage of all retained proposals. Every row is verified.

Cohort	ID	labels	$\|\Delta\|_2$	r_L	$\|\Delta\|_2/r_L$	nominal	shifted
3	174	$6 \to 2$	.1641	.1852	.886	57.47	18.19
11	1586	$3 \to 2$	.3205	.3707	.864	88.38	29.30
15	879	$2 \to 6$	.2744	.2912	.942	77.59	29.93
16	1060	$7 \to 4$	.4674	.4739	.986	94.80	23.82
19	759	$1 \to 9$	.2920	.3014	.969	78.27	5.80
37	215	$0 \to 9$	.1325	.1496	.885	47.36	3.88
51	1986	$8 \to 0$	.1649	.1896	.870	57.37	11.17
61	366	$9 \to 2$	.3198	.3421	.935	84.24	21.36
79	775	$1 \to 9$	.1799	.1940	.927	57.96	2.99
84	289	$9 \to 8$	.3499	.3687	.949	87.21	19.66
99	1267	$8 \to 0$	.2453	.2544	.964	71.19	15.82
125	791	$8 \to 1$	.3082	.3530	.873	86.34	7.77

K.3 Fixed-cohort CIFAR-10.1 attack

The protocol orders all $2{,}021$ indices by the SHA-256 hash of a fixed salt and the index, then takes the first 128. This selection does not use a model output or image label. All cohort members remain in the primary denominator. The released radius is positive for 107 images and nonpositive for 21.

Every positive-radius image receives four projected-gradient restarts with 150 steps, 256 Gaussian proposals per step, and temperatures 1, 0.5, and 0.25. Candidate centers are projected to $[0,1]^d$, while the Gaussian proposals supplied to the model remain unclipped. The best and final centers from each restart give eight candidates per positive-radius image. Each candidate receives a disjoint $50{,}000$-proposal screen. The center with the largest target margin is retained for that image. The search takes 8852 seconds and produces twelve positive screens. The supplement contains every search attempt and the preconfirmation summary.

Each positive screen then receives $250{,}000$ new proposals at the nominal and shifted endpoints. One Bonferroni family reserves nine pairwise winner tests at each endpoint and one nominal conditional-probability lower bound for every possible cohort member. Its size is $128[2(10-1)+1] = 2{,}432$, its familywise error is 0.001, and its per-inference error is 4.112×10^{-7}. The confirmation takes 466 seconds. Its result and row file are included in the supplement.

All twelve screened candidates have verified opposite endpoint labels and lie below fresh lower bounds on the corresponding population substituted radii. The largest endpoint test value among them is 5.27×10^{-11}. Table 5 gives the displacement, fresh radius bound, and retained-share effect sizes. The primary fixed-cohort attack yield is $12/128$. Nine of the twelve verified cases occur among the 80 cohort images that are correctly classified and have a positive released radius. These are results for the declared attack and finite cohort. The generic output also contains a binomial interval for the count. We do not use it because deterministic hash-ranking does not define an independent probability sample from a model population. Candidate tensors derived from CIFAR-10.1 images are not redistributed.

K.4 Output-selected boundary confirmation

The boundary protocol specifies twelve candidate images, the search procedure, and disjoint search, screen, and confirmation seeds. Candidate images came from the opened full-test output and were correctly classified examples whose filtered and unfiltered selected labels differed. Four projected-gradient restarts used 256 proposals per step. A fresh $50{,}000$-proposal screen selected at most six positive alternative-minus-original differences. The selected order, labels, and image tensors were then fixed before confirmation. Candidate centers were projected to $[0,1]^d$, and Gaussian proposals remained unclipped.

Each endpoint uses one million fresh proposals under the released Gaussian law and confidence filter. For a declared winner w and competitor c, condition on the number of retained proposals carrying either label. Under the null $s_w \le s_c$, the winner count is binomial with success probability at most $1/2$. Its upper tail therefore gives an exact one-sided test. One Bonferroni family contains all $6 \times 2 \times 9 = 108$ endpoint tests and six nominal one-sided Clopper–Pearson lower bounds. The per-inference error is $0.001/114$. If L_A is a nominal conditional-probability lower bound, $r_L = \sigma\Phi^{-1}(L_A)$ lower bounds the binary population radius used by the released formula. Verification requires opposite endpoint winners and $\|\Delta\|_2 < r_L$. Table 6 reports every endpoint test.

Table 6: Independent endpoint confirmation. Count pairs give original-label and alternative-label retained counts among one million proposals. A verified row has both declared endpoint winners and a shift below r_L.

ID	labels	$\|\Delta\|_2$	r_L	nominal counts	shifted counts	verified
4924	$7 \to 5$	.4577	.5097	52295/965	7182/17611	yes
5089	$6 \to 3$	.3116	.3585	32339/2345	9196/17483	yes
8654	$2 \to 3$	.3316	.4141	8736/358	2052/4782	yes
6665	$4 \to 3$	.2843	.3685	18244/1172	5741/7099	yes
634	$7 \to 4$	.3068	.4013	36793/1873	11556/12135	no
5211	$7 \to 3$	.2570	.3775	27299/1245	7754/7157	no

All nominal endpoint maximum p-values are below 10^{-300}. The largest shifted-endpoint p-value among verified rows is 2.09×10^{-33}. Pair 634 has a shifted value 8.65×10^{-5}, above the simultaneous cutoff, and pair 5211 does not reverse its empirical label ordering. Numerical underflowed values are not reported as exact zeros. Four population substitution violations are verified with familywise error at most 0.001. Each verified shift is also below the finite-sample radius stored by the released calculation during candidate selection. The stronger confirmation statement uses fresh nominal samples to lower bound the corresponding population substituted radius. The raw confirmation records, analysis, and execution log are included in the supplement. Since the candidates were selected from opened output, this result establishes existence but does not estimate prevalence.

K.5 Confidence-threshold selection

The threshold protocol was specified before any multithreshold result was generated. The declared grid is 0, 0.5, 0.7, 0.8, 0.85, 0.9, 0.925, 0.95, and 0.975. Each network evaluation is shared across every threshold. The proposal counts, error probability, model, checkpoint, and noise scale match the full-test study, while the base random seed is independent. The first $1{,}000$ ordered test indices form the development set and the remaining $9{,}000$ form the confirmation set. The supplement includes the per-image record and summary.

The objective averages explicit-joint certified accuracy over radii 0.25, 0.5, and 0.75. The development set selects threshold zero. Table 7 reports every candidate and confirms the same ordering on the untouched indices. This study addresses the released checkpoint only and does not rule out a benefit from confidence filtering after different training.

Table 7: Threshold selection. Dev and confirm are mean explicit-joint certified accuracies over the three declared radii. Retained is the mean confirmation proposal fraction.

Threshold	0	.5	.7	.8	.85	.9	.925	.95	.975
Dev	.400	.395	.376	.355	.344	.324	.308	.283	.240
Confirm	.399	.394	.376	.358	.343	.320	.304	.279	.237
Retained	1.000	.876	.698	.607	.555	.493	.455	.407	.338

K.6 Population feasible-action reanalysis

The reach–avoid reanalysis uses the already opened protocol-v2 population. It contains seven deterministic reference states from each of 64 layouts. Analytic products of one-dimensional Gaussian CDF differences give occupancy and label-event masses for each rectangular union. A bracketed scalar root locates the first vertical label boundary with Brent tolerances 2×10^{-13}. These are population calculations with standard floating-point Gaussian CDF evaluations and contain no Monte Carlo confidence interval. The smallest substituted-radius excess over the located boundary is 4.19×10^{-4}. The smallest boundary excess over the joint radius is 4.72×10^{-4}. The crossing statements use a 2×10^{-10} numerical margin. They are not outward-interval certificates.

The substituted conditional radius crosses the located boundary at all 448 states. The joint-mass radius is positive at all states and crosses none. Its median, fifth percentile, and ninety-fifth percentile are 0.2267, 0.0197, and 0.4719. The corresponding substituted values are 0.3307, 0.0320, and 0.6540. The joint radius divided by the boundary distance has median 0.7708 and maximum 0.8628. The joint radius divided by the substituted radius has median 0.6872.

K.7 Geometric breadth

A separate synthetic study crosses dimensions $2, 5, 10, 20$, component counts $4, 8, 16$, two geometry families, two Gaussian scales, two normalized separations, and paired volume settings. The 6,144 cells reuse 384 independently seeded base geometries and are not independent observations. Figure 5 reports the geometry-level comparison.

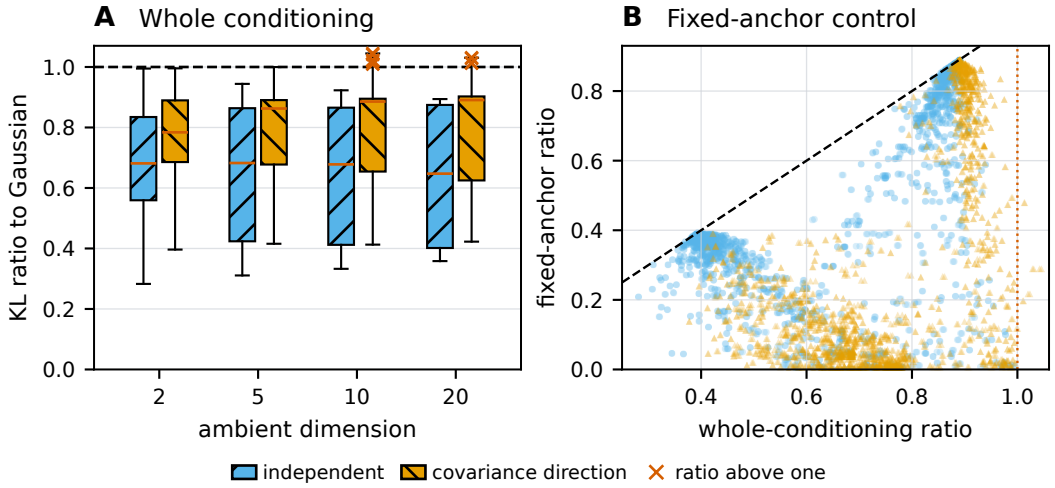

Figure 5: Product-box study. Panel A finds ten covariance-direction KL ratios above one from five base geometries. The maximum is 1.0448, and no independent-direction ratio exceeds one. Every fixed-anchor control in Panel B stays below one. The study does not estimate prevalence.

K.8 Held-out covariance replication

The replication protocol was specified after the earlier study identified the coherent 16-box, balanced-volume family and before drawing the new geometry and direction seeds. It uses dimensions 10 and 20, scales 0.25 and 0.5, normalized distances 0.05 and 0.15, and 64 base geometries per dimension. Each geometry is reused for the fixed scale, distance, and direction comparisons. The 128 base geometries, not the 1,024 factor rows, are independent analysis units.

Fourteen geometries exceed the Gaussian KL rate in the anchor top-covariance direction and none does so in the independently seeded direction. The exact one-sided sign test on the paired exceedance indicators has value 6.10×10^{-5}. The covariance-direction maximum exceeds the independent maximum in every geometry, with median difference 0.06629. The minimum fixed-cell Spearman correlation between anchor directional covariance and finite KL ratio is 0.9941. Direct KL and the

covariance path integral agree within 1.77×10^{-15}. Figure 6 gives the full local diagnostic. The paired maxima are computed from the complete row file, which is included with the result summary in the supplement.

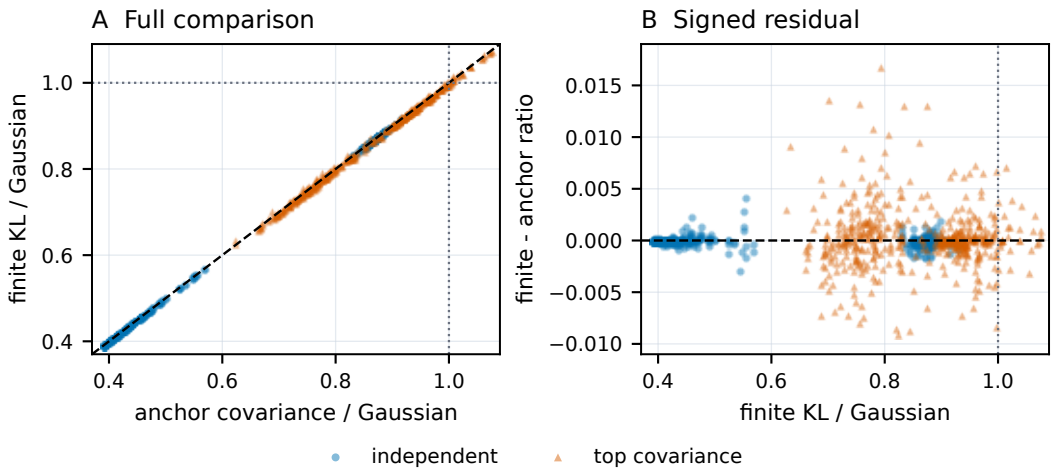

Figure 6: Held-out covariance replication. Panel A compares anchor directional covariance and finite KL over all 1,024 factor rows. The diagonal is equality and dotted lines mark the Gaussian rate of one. Panel B shows the signed finite-minus-anchor ratio, exposing the finite-scale discrepancy hidden by the near-diagonal scatter. Twenty-seven repeated top-covariance rows from 14 base geometries exceed one. No independent row does. The 128 base geometries, not factor rows, are the independent analysis units. Direct KL and the path integral agree within 1.77×10^{-15}.

The family was selected from earlier results, so this is a targeted replication rather than a prevalence study. The top-covariance direction is an adaptive stress direction. The local covariance ratio and finite KL ratio test Proposition 4. They do not certify a uniform covariance bound or predict a label change.

K.9 External projected-band confirmation

This study evaluates the geometry-controlled certificate in Proposition 9 on a learned image classifier. The filter was selected using CIFAR-10 training images and fixed before CIFAR-10.2 was accessed. It retains a raw noisy proposal z when

$$0.0625 \le |\langle u, z\rangle + 0.1072130481| \le 0.25, \qquad \sigma = 0.25.$$

The unit vector u and its exact representation are included in the supplement. For $T = \langle u, z\rangle + b$, retention implies $T \in [-\beta, \beta]$. Popoviciu gives $\mathrm{Var}(T \mid R = 1) \le \beta^2 \le \sigma^2$. The orthogonal Gaussian coordinates are unchanged and independent of the retention event. Hence $\mathrm{Cov}(Q_c) \preceq \sigma^2 I$ for every center c. This is a global analytic premise, not an empirical covariance estimate.

The official CIFAR-10.2 test set (Lu et al., 2020) contains 2,000 images with 200 from each class. The supplement records its exact revision and file checksum. Images are evaluated in stored order. The AuditVotes source, checkpoint, and architecture are those in Appendix K.1. The supplement records the dataset checksum and contains the confirmation protocol and runner.

Each image receives four independent streams. Filtered and unfiltered label selection use 10,000 raw proposals each. Their independent estimation streams use 100,000 proposals each. Gaussian proposals are not clipped. The filter is applied before model inference and rejected proposals are not replaced. All images remain in the denominator. One Bonferroni family assigns error 0.001 across 64,000 one-sided Clopper–Pearson bounds. These cover conditional label probabilities, joint retained-label masses, the independently selected joint runner used for an upper radius bound, and the unfiltered probabilities.

Table 8: CIFAR-10.2 results by class. Retained is the mean estimation-stream fraction. F-acc and U-acc are filtered and unfiltered selected-label accuracy. KL and Joint are correct certified fractions at radius 0.2 in the original forward-KL analysis. Strong counts correct images whose forward-KL lower radius exceeds the joint-mass upper radius by at least 0.01.

Class	Retained	F-acc	U-acc	KL	Joint	Strong
airplane	.420	.505	.505	.280	.250	26
automobile	.418	.640	.640	.420	.385	35
bird	.437	.505	.510	.275	.210	34
cat	.424	.545	.540	.250	.190	44
deer	.432	.585	.585	.340	.275	34
dog	.415	.660	.655	.490	.420	34
frog	.437	.665	.665	.470	.410	35
horse	.425	.745	.740	.520	.490	32
ship	.437	.770	.770	.585	.535	29
truck	.430	.695	.690	.515	.460	26
all	.428	.632	.630	.415	.363	329

Filter selection and a 200-image replication used CIFAR-10 training images only. These development results do not enter the reported CIFAR-10.2 evidence. The supplement contains the development records, the final protocol, and the complete execution history.

At radius 0.2, the original forward-KL, joint-mass, and unfiltered calculations correctly certify $829/2000$, $725/2000$, and $948/2000$ images. Forward KL certifies 104 images that joint mass does not, with none in the reverse direction. The paired exact one-sided value is 4.93×10^{-32}. This value is secondary evidence rather than sampling uncertainty for the finite census. Among correctly selected filtered labels, 329 have $r_{\mathrm{cov},L} - r_{\mathrm{mass},U} \geq 0.01$ under the same simultaneous confidence event. This compares two certificate formulas. It is not a lower bound on a difference between true robust radii.

The mean retained fraction is 0.427570. Filtered and unfiltered selected-label accuracies are 0.6315 and 0.6300. The filtered streams make $94{,}064{,}290$ model calls from 220 million proposals. The unfiltered streams make 220 million calls. The run took $26{,}992$ seconds on one RTX 4070 Ti with PyTorch 2.2.2. The supplement contains the manifest, complete row file, and summary. Ordinary unfiltered smoothing remains stronger at radius 0.2. This finite census covers one checkpoint, noise scale, and training-selected filter, not other models, filters, or systems.

K.10 Conditional divergence reanalysis

We applied Proposition 9 to the same stored counts after the original experiment. The filter, predictions, and confidence bounds are unchanged. The finite order set is

$$\{1\} \cup \{1 + k/20 \mid k = 1, \ldots, 80\} \cup \{6, 8, 12, 16, 32, 64\}.$$

For each image, we maximize the valid radius over these orders using its original conditional lower and upper probability bounds. This choice adds no confidence events or model evaluations. The covariance premise is global, so $R = \infty$ and $\Lambda = 1$. This is a deterministic reanalysis, not a new independent confirmation.

Table 9 separates the effect of reversing KL from the additional improvement obtained by optimizing the Rényi order. The forward-KL radius cannot exceed $0.25\sqrt{2 \log 2} = 0.29435\ldots$ for this covariance bound. Reverse KL removes that ceiling. At radius 0.2, Rényi certifies 167 correct images that joint mass does not, with none in the reverse direction. For 877 correctly selected labels, the Rényi lower radius exceeds the joint-mass upper radius by at least 0.01. These bounds share the original familywise error 0.001 and compare certificate formulas, not true robust radii.

Outward numerical validation. We recomputed the full-count conditional Rényi, reverse-KL, joint-mass, and unfiltered certificates using Arb interval arithmetic at 128-bit precision. Each binomial endpoint is accepted only after its tail inequality is proved against the exact rational error

Table 9: Correctly certified CIFAR-10.2 images out of 2,000. All filtered methods use identical counts and selected labels. The unfiltered predictor uses its separate original streams. Matched calls uses a random subset of those streams with the same model-call count as the filtered methods.

Certificate	$r \geq 0.2$	$r \geq 0.3$	$r \geq 0.5$
Conditional forward KL	829	0	0
Conditional reverse KL	888	698	329
Conditional Rényi	892	713	381
Joint mass	725	417	0
Unfiltered, original	948	814	529
Unfiltered, matched calls	945	806	522

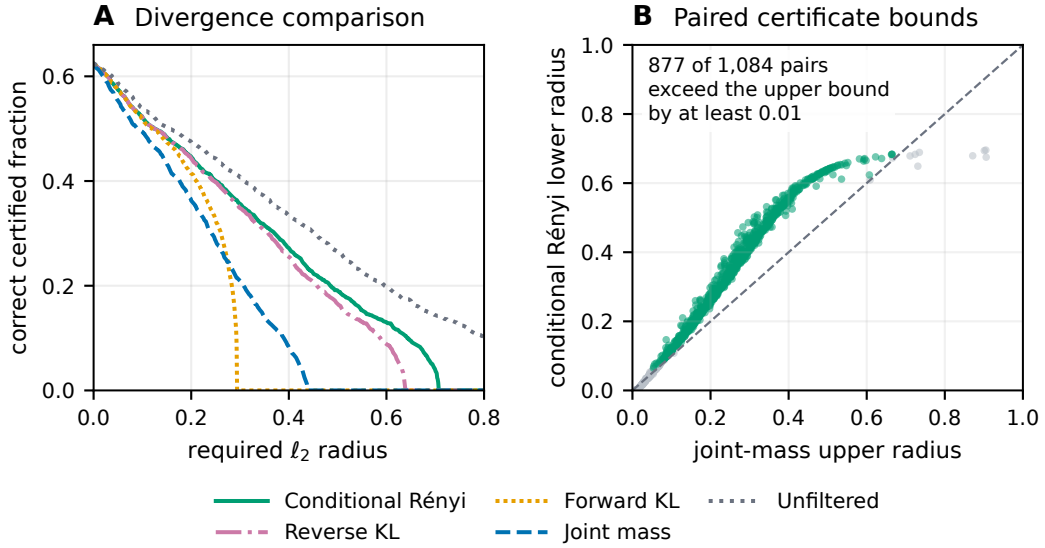

Figure 7: Conditional divergence certificates on the retained CIFAR-10.2 counts. Panel A includes all 2,000 images and displays the forward-KL ceiling and its removal by reverse KL. Panel B shows all 1,084 correctly selected labels with finite paired bounds. Green points have a conditional Rényi lower radius at least 0.01 above their joint-mass upper radius. The dashed diagonal denotes equality.

$1/64{,}000{,}000$. A positive-term binomial recurrence bounds its remaining terms by a geometric series. Gaussian quantiles and the finite-order divergence formulas are rounded outward. The certified counts at 0.2, 0.3, and 0.5, and all 877 separation statements, are unchanged. The largest change to a conditional probability endpoint is below 3.51×10^{-12}. Reporting thresholds are compared as exact rational numbers. This validates inference from the stored integer counts. It does not validate noise generation, filter evaluation, neural inference, the matched-call calculation, or the separate AuditVotes endpoint study.

Matching model calls. The filtered streams evaluate only retained proposals. To compare equal model call counts, we draw multivariate-hypergeometric subsets of each original unfiltered count vector. Each subset size equals the corresponding filtered stream's model-call count, giving 94,064,290 calls in total. Selection and estimation subsets are drawn separately with seed 20260924. Their sizes depend on independent filtered streams, not on unfiltered labels. Thus they represent uniformly sampled subsets of the unfiltered IID proposals. We reselect labels on the smaller selection batch and use the original per-tail error 1.5625×10^{-8} for 20,000 one-sided binomial bounds. Six selected labels change. Unfiltered smoothing still certifies more images at all three tabulated radii. This is one randomized count-level comparison, not a new network or timing run. Its confidence family is separate from the original evaluation. Its familywise error is at most 0.0003125, and the combined error is at most 0.0013125.