SUPPORT-AWARE COUNTERFACTUAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-grounded counterfactual methods often retrieve one or a few nearby instances with different outcomes, but finite samples may not reveal whether these instances reflect a recurring local pattern or isolated observations. We introduce **Support-Aware Counterfactual Analysis (SACA)** to characterize the surrounding evidence and determine which conclusions the data support. SACA estimates a distance–direction landscape of different-outcome probabilities and characterizes the evidence through three perspectives: *prevalence* measures how common different outcomes are nearby, *direction* identifies where they concentrate, and *reach* measures how far one must look to accumulate a prescribed amount of evidence. It reports regional estimates only when finite-sample bounds meet the requested accuracy and otherwise abstains. These support deficits guide acquisition of additional outcomes from a pool of unlabeled cases. We establish a near-optimal rate for landscape estimation, finite-sample guarantees for reach and reported regions, and approximation guarantees for acquisition. Experiments on six real-world datasets and five synthetic settings show that SACA reduces landscape estimation error by 22%, makes about 19% more regions reportable, and selects cases 22 times faster.

1 INTRODUCTION

Counterfactual analysis asks a central question in data-driven decision-making (Wachter et al., 2018): *which nearby alternatives would receive a different outcome?* It supports model explanation (Miller, 2019), actionable recourse (Karimi et al., 2021), and system auditing (Lee, 2021), with applications in healthcare (Prosperi et al., 2020), finance (Gan et al., 2021), and law (Zhang et al., 2023). Data-grounded methods address this question using observed cases with different outcomes (Poyiadzi et al., 2020; Kenny & Keane, 2021; Chen et al., 2026). Counterfactual methods typically return one or a few alternatives selected for proximity (Wachter et al., 2018), plausibility (Kanamori et al., 2020), actionability (Ustun et al., 2019), or diversity (Mothilal et al., 2020).

Figure 1 shows why nearby examples may not represent the surrounding evidence. Across three scenarios, a denied applicant has an approved case at the same nearest distance, yet approvals are spread across many directions in (a), concentrated in two directional groups in (b), and isolated in (c). Thus, the same nearest-counterfactual distance can correspond to different evidence *patterns*, defined by how common different outcomes are nearby, where they concentrate, and how far one must look to find them. Interpreting such a pattern also requires *support*: enough reference cases in the relevant region for accurate estimation. One approval among many nearby denials indicates that approvals are locally rare, whereas one approval among only a few observations remains inconclusive. Counterfactual analysis must therefore characterize the surrounding pattern and determine which conclusions the finite sample supports.

We introduce **Support-Aware Counterfactual Analysis** (SACA), which estimates a *counterfactual landscape* describing how the probability of a different outcome varies with distance and direction from the query. SACA characterizes the surrounding evidence through three complementary perspectives: *prevalence* measures how common different outcomes are nearby, *direction* identifies where they concentrate, and *reach* measures how far one must look in a given direction to accumulate a prescribed amount of evidence rather than relying on a single nearest case. Together, these perspectives distinguish the evidence patterns in Figure 1.

SACA also assesses which findings the available data support. It reports a regional estimate (i.e., how common different outcomes are within a region near the query) only when its error bound meets

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

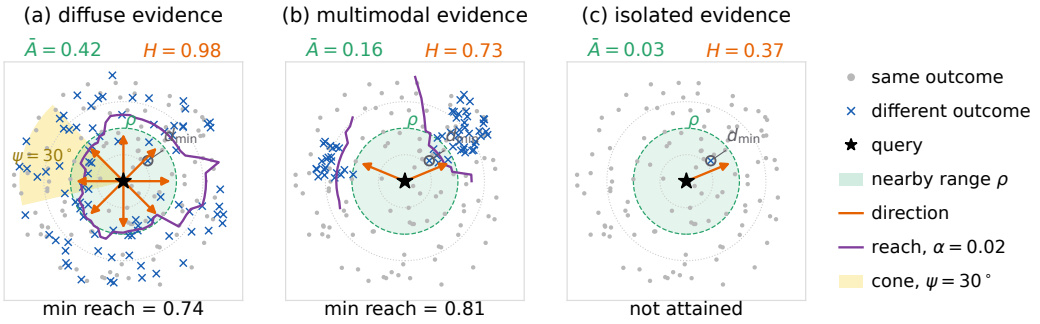


Figure 1: Here $\bar{A}$ is the different-outcome probability averaged over the nearby distance range and uniformly across directions, H is normalized directional entropy (higher values indicate broader dispersion), and $d_{\min}$ is nearest-counterfactual distance. The green disk marks the nearby range ρ; equal-length orange arrows mark directions carrying evidence; and the yellow sector shows a cone with half-angle $\psi = 30°$. Purple curves show empirical reach at $\alpha = 0.02$: in each direction, the smallest radius within which different-outcome cases in the cone account for 2% of the reference sample. "Min reach" is the smallest attained reach across directions; no curve is drawn where the level is unattained. All panels have $d_{\min} = 0.6$. Panel (a) has high $\bar{A}$ and H with reach across many directions; (b) has lower values with reach near two directional groups; and (c) has low $\bar{A}$ and unattained reach despite a nearby different-outcome case.

the requested accuracy and otherwise abstains. Abstention exposes support deficits, the shortfalls between regional case counts and a sufficient reporting count. Given a limited budget and a pool of cases with known features but unknown outcomes (Lewis & Gale, 1994; Sener & Savarese, 2018; Bae et al., 2025), SACA selects cases that add observations to these under-supported regions. For a denied applicant, SACA can report how common approvals are among comparable applicants, in which feature-space directions they concentrate, and how far one must look to accumulate sufficient approval evidence. If comparable cases are too scarce, it abstains and prioritizes additional cases that reduce these support deficits, e.g., through credit investigation of comparable applicants.

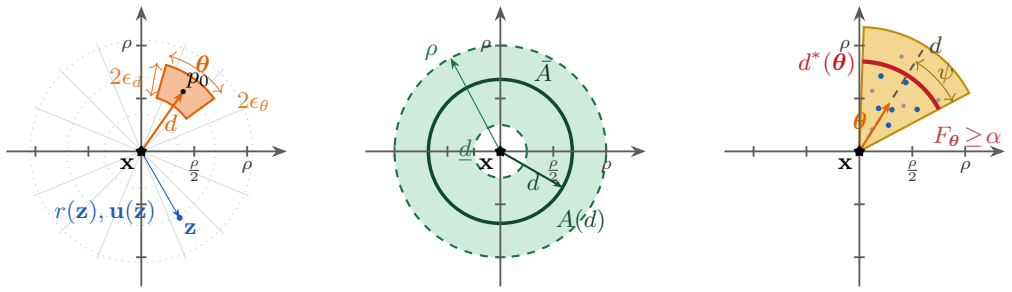
Our theoretical analysis establishes a convergence rate for the landscape estimation error that is minimax-optimal up to logarithmic factors, finite-sample reach bounds under a crossing condition, and simultaneous accuracy guarantees for reported regional estimates. The submodular acquisition objective yields a $(1 - 1/e)$ approximation under a fixed budget and a logarithmic approximation to the minimum number of acquisitions when all support deficits can be eliminated. Experiments demonstrate accurate landscape recovery, reliable regional reporting, and efficient support-guided acquisition that increases the number of reportable regions. Together, these results move counterfactual analysis beyond finding a nearby alternative to **characterizing the surrounding evidence**, **assessing what the data support**, and **identifying what to observe next**.

2 Problem Formulation

Let the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ contain n labeled reference cases, where each $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^m$ has m features and $y_i \in \mathcal{Y}$ is its outcome. Let $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ be a query. Outcomes are either observed labels or predictions of a fixed rule, with the same convention for the query and reference cases. The nearest-counterfactual distance is $d_{\min}(\mathbf{x}) = \min_{i:y_i \neq y} \|\mathbf{x}_i - \mathbf{x}\|_2$, with value $+\infty$ if no reference case has a different outcome. Tables 7 and 8 collect the notation and terminology.

Distance–direction regions. Existing counterfactual methods use the dataset or a fitted model to return one or a few alternatives (Section 6); they do not characterize how different outcomes are distributed around the query or assess whether the local sample supports an accurate conclusion. We therefore organize the surrounding evidence by distance and direction. For $\mathbf{z} \in \mathcal{X} \setminus \{\mathbf{x}\}$, define

$$r(\mathbf{z}) = \|\mathbf{z} - \mathbf{x}\|_2, \qquad \mathbf{u}(\mathbf{z}) = \frac{\mathbf{z} - \mathbf{x}}{\|\mathbf{z} - \mathbf{x}\|_2} \in \mathbb{S}^{m-1}. \tag{1}$$



(a) probe $\gamma = (d, \boldsymbol{\theta})$, region $\mathcal{R}_\gamma$, grid $\mathcal{G}$

(b) $A(d)$: the circle at radius d. $\bar{A}$: the annulus $[\underline{d}, \rho]$

(c) reach $d^*(\boldsymbol{\theta})$: first radius at which the cone mass reaches α

Figure 2: Distance–direction geometry of the counterfactual landscape around the query $\mathbf{x}$, with radial ticks in multiples of the local scale ρ. (a) The orange probe region $\mathcal{R}_\gamma$ corresponds to $\gamma = (d, \boldsymbol{\theta})$; faint rings and spokes form the grid $\mathcal{G}$; the blue reference case $\mathbf{z}$ has distance $r(\mathbf{z})$ and direction $\mathbf{u}(\mathbf{z})$. (b) Prevalence: $A(d)$ averages the landscape over directions along the dark-green circle, while $\bar{A}$ averages $A(d)$ over the light-green annulus from $\underline{d}$ to ρ. (c) Reach: within the yellow ψ-cone around $\boldsymbol{\theta}$ (orange arrow), $F_{\boldsymbol{\theta}}(d)$ accumulates blue different-outcome cases up to the dashed radius d, excluding gray ones; the red arc marks the first radius $d^*(\boldsymbol{\theta})$ at which it reaches α.

Here $r(\mathbf{z})$ is the Euclidean distance from the query and $\mathbf{u}(\mathbf{z})$ is the unit displacement direction on the sphere $\mathbb{S}^{m-1}$. A probe $\gamma = (d, \boldsymbol{\theta})$, with $d > 0$ and $\boldsymbol{\theta} \in \mathbb{S}^{m-1}$, examines the region

$$\mathcal{R}_\gamma = \{\mathbf{z} \in \mathcal{X} \setminus \{\mathbf{x}\} : |r(\mathbf{z}) - d| \le \epsilon_d,\ \angle(\mathbf{u}(\mathbf{z}), \boldsymbol{\theta}) \le \epsilon_\theta\}. \tag{2}$$

The angle $\angle$ is measured between unit vectors, and $\epsilon_d, \epsilon_\theta > 0$ set the radial and angular resolutions. Regions include cases regardless of outcome, allowing different-outcome cases to be compared with all reference cases in the same region.[1] We evaluate these regions on a finite grid

$$\mathcal{G} = \{(d_j, \boldsymbol{\theta}_t) : j = 1, \ldots, n_d,\ t = 1, \ldots, T\}, \tag{3}$$

where $\underline{d}$ and $\overline{d}$ are the chosen distance limits, with $0 < \underline{d} \le d_j \le \overline{d} < \infty$, and each $\boldsymbol{\theta}_t \in \mathbb{S}^{m-1}$.[2]

The grid and resolutions are fixed before examining the sample and remain unchanged during estimation and acquisition. A denser grid evaluates more locations, while narrower regions resolve finer patterns with fewer reference cases (Appendix E.6). Figure 2 illustrates the distance–direction geometry of the counterfactual landscape: (a) shows a probe $\gamma = (d, \boldsymbol{\theta})$, its region $\mathcal{R}_\gamma$ with radial and angular half-widths ϵ_d and ϵ_θ, and the grid $\mathcal{G}$; (b) and (c) illustrate the geometry underlying the three perspectives of Section 3.1: prevalence, direction, and reach.

From estimation to acquisition. Given $\mathcal{D}$, $(\mathbf{x}, y)$, and $\mathcal{G}$, the first task is to estimate the different-outcome probability within each $\mathcal{R}_\gamma$. For a requested accuracy $\varepsilon > 0$ and confidence level $1 - \delta$, with $\delta \in (0, 1)$, SACA reports an estimate only when its error bound is at most ε, so that all reported regional estimates meet the requested accuracy simultaneously with probability at least $1-\delta$, without requiring any assumption on how outcomes vary over the feature space. The second task is *support-guided acquisition*. Given a pool $\mathcal{U}$ of cases with known features but unknown outcomes and a budget B, SACA selects $S \subseteq \mathcal{U}$, $|S| \le B$, to reduce regional support deficits, defined as the gaps between their case counts and a sufficient reporting count (Section 3.3).

3 Support-Aware Counterfactual Analysis

To estimate the evidence surrounding a query, we define a *counterfactual landscape* describing how the probability of a different outcome varies with distance and direction.[3] Let $(\mathbf{X}, Y)$ denote a

[1] Appendix C.7 further analyzes regions restricted to feasible cases (e.g., immutable features, action costs).

[2] Under feature permutations or rotations, probe directions transform accordingly; under uniform scaling, d and ϵ_d are rescaled. Corresponding probe regions contain the same cases, with ϵ_θ unchanged (Appendix B.2).

[3] Throughout this section, the query $(\mathbf{x}, y)$ is fixed and omitted from the notation, and sums exclude reference cases with $\mathbf{x}_i = \mathbf{x}$. Implementation details, pseudocode, and complexity analysis appear in Appendix C.

random instance–outcome pair from the underlying data distribution. The *pointwise landscape* is

$$p_0(d, \boldsymbol{\theta}) = \Pr(Y \neq y \mid r(\mathbf{X}) = d,\ \mathbf{u}(\mathbf{X}) = \boldsymbol{\theta})\,, \tag{4}$$

which gives the conditional probability of a different outcome at $\mathbf{x} + d\boldsymbol{\theta}$.

For each probe region $\mathcal{R}_\gamma$, we define the *regional counterpart* $p_\gamma = \Pr(Y \neq y \mid \mathbf{X} \in \mathcal{R}_\gamma)$. It averages the different-outcome probability over the region and converges to $p_0(d, \boldsymbol{\theta})$ as $(\epsilon_d, \epsilon_\theta)$ shrink to zero under the conditions in Appendix B.1. Thus, p_0 is the pointwise target, while p_γ enables finite-sample assessment of regional evidence.

Since real datasets $\mathcal{D}$ rarely contain observations exactly at a given probe, we estimate p_0 by pooling nearby cases using a kernel estimator:

$$\widehat{p}(d, \boldsymbol{\theta}) = \frac{\sum_i K(r(\mathbf{x}_i) - d)\, L(\mathbf{u}(\mathbf{x}_i), \boldsymbol{\theta})\, \mathbb{I}[y_i \neq y]}{\sum_i K(r(\mathbf{x}_i) - d)\, L(\mathbf{u}(\mathbf{x}_i), \boldsymbol{\theta})}. \tag{5}$$

Here $\mathbb{I}[\cdot]$ is the indicator function, K is a radial kernel, and L is a directional kernel on the unit sphere. Their product gives greater weight to cases closer to the probe in distance and direction, allowing the estimate to vary smoothly across probes. The three perspectives in Section 3.1 use this smooth kernel estimate to integrate the landscape over distance and direction, even when individual regions are sparse. The regional proportion serves another role: together with its case count, it yields an error bound that determines whether support is sufficient for reporting or requires abstention (Section 3.2), while the count guides acquisition (Section 3.3).

3.1 Three Perspectives on Counterfactual Evidence

Building on the landscape, we characterize counterfactual evidence through three complementary perspectives: *prevalence*, *direction*, and *reach*. The analyst specifies a distance scale $\rho \in (\underline{d}, \overline{d}]$ to define what counts as nearby, with smaller values emphasizing cases more similar to the query.

Prevalence: how common is nearby evidence? Prevalence averages the pointwise landscape uniformly over directions and then over the nearby distance range:

$$A(d) = \mathbb{E}_{\boldsymbol{\theta}}[p_0(d, \boldsymbol{\theta})], \qquad \bar{A} = \frac{1}{\rho - \underline{d}} \int_{\underline{d}}^{\rho} A(t)\, dt, \tag{6}$$

where $\boldsymbol{\theta}$ is uniform on the unit sphere. A high $\bar{A}$ indicates that counterfactual evidence is common on average across the nearby range. Unlike the empirical fraction of nearby different-outcome cases, $\bar{A}$ averages the landscape uniformly over distance and direction, so it summarizes counterfactual evidence across the region rather than the concentration of observed cases. Figure 2(b) visualizes $A(d)$ along the dark-green circle and $\bar{A}$ over the light-green annulus from $\underline{d}$ to ρ.

Direction: where is nearby evidence concentrated? For each direction $\boldsymbol{\theta}$, we aggregate the pointwise landscape over distance, giving greater weight to closer evidence:

$$q(\boldsymbol{\theta}) \propto \int_{\underline{d}}^{\overline{d}} e^{-t/\rho}\, p_0(t, \boldsymbol{\theta})\, dt. \tag{7}$$

Here ρ controls the distance decay, with smaller values emphasizing closer evidence. When the total weighted evidence is positive, q is normalized over the unit sphere; on the fixed direction grid $\{\boldsymbol{\theta}_t\}_{t=1}^T$, the direction weights are normalized to obtain q_t. Larger $q(\boldsymbol{\theta})$ indicates greater relative concentration along $\boldsymbol{\theta}$. The direction with the largest q_t is the dominant direction, while the normalized entropy $H = -\sum_{t=1}^T q_t \log q_t / \log T$ measures dispersion: low H indicates concentration in a few directions, and high H indicates a broad distribution. The orange arrow in Figure 2(c) marks a direction $\boldsymbol{\theta}$ that serves as the axis of the reach cone defined below.

Reach: how far away is sufficient evidence? Reach tracks how counterfactual evidence accumulates along a direction. For a cone of half-angle ψ centered at $\boldsymbol{\theta}$, define the cumulative different-outcome mass up to distance d as

$$F_{\boldsymbol{\theta}}(d) = \Pr(Y \neq y,\ r(\mathbf{X}) \leq d,\ \angle(\mathbf{u}(\mathbf{X}), \boldsymbol{\theta}) \leq \psi)\,. \tag{8}$$

At an evidence level $\alpha \in (0, 1)$,[4] reach is the first distance at which the cumulative mass attains α: $d^*(\boldsymbol{\theta}) = \inf\{d \geq 0 : F_{\boldsymbol{\theta}}(d) \geq \alpha\}$, with $d^*(\boldsymbol{\theta}) = \infty$ if the level is never attained. In Figure 2(c), the

[4]We fix α and ψ before examining the sample and assess their sensitivity in Appendix E.3.

yellow sector shows the ψ-cone, the blue different-outcome cases within the dashed radius contribute to $F_{\boldsymbol{\theta}}(d)$, and the red arc marks the first radius $d^*(\boldsymbol{\theta})$ at which the accumulated mass reaches α. Empirically, we use the fraction of reference cases satisfying these conditions (Appendix C.3). At $\alpha = 1/n$, empirical reach reduces to the distance to the nearest different-outcome case in the cone. In general, it depends on accumulated evidence rather than mere existence and is infinite unless at least $\lceil n\alpha \rceil$ different-outcome cases lie in the cone. Under a crossing condition, Theorem 4 bounds the error of a finite empirical reach by a term that decreases with n. A short reach places sufficient evidence nearby, while an infinite reach indicates that the level is unattained.

For a denied loan applicant, these perspectives distinguish three situations in Figure 1. In (a), high prevalence and high directional entropy indicate that approvals are common across many directions, while short reach places accumulated approval evidence nearby. In (b), lower prevalence and two directional modes indicate that approvals are concentrated in particular combinations of feature differences; reach shows how far one must look along those directions to accumulate the required evidence. In (c), low prevalence and unattained reach at $\alpha = 0.02$ show that the nearest approved case lacks comparable accumulated evidence. These distinctions show whether a nearby approval reflects a broader pattern and where supporting cases lie, while the certificates in Section 3.2 determine which regional conclusions the sample supports.

3.2 Reporting Only Supported Estimates

Conventional counterfactual methods identify individual alternatives (Wachter et al., 2018; Keane & Smyth, 2020), but these alone do not show which patterns in the surrounding evidence are supported by the sample or when the sample is too small to support a reliable conclusion. SACA uses its two estimators for distinct purposes. The kernel estimator $\widehat{p}$ in Eq. (5) provides the smooth landscape underlying the three perspectives, but its error depends on the unknown smoothness and local density of the data (Györfi et al., 2002), so it cannot provide a sample-computable certificate for a particular estimate. For the report-or-abstain decision, SACA instead estimates the regional probability p_γ using the proportion of different-outcome cases in $\mathcal{R}_\gamma$ and bounds its error using only the observed sample. A sufficiently small bound supports reporting; otherwise, SACA abstains because the regional evidence is insufficient for the requested accuracy.

For each probe $\gamma \in \mathcal{G}$ with $|\mathcal{R}_\gamma|$ reference cases in its region, the regional estimator is

$$\widehat{p}_\gamma^{\,\mathrm{bin}} = \frac{1}{|\mathcal{R}_\gamma|} \sum_{i:\, \mathbf{x}_i \in \mathcal{R}_\gamma} \mathbb{I}[y_i \neq y]. \tag{9}$$

Its empirical-Bernstein confidence width (Audibert et al., 2009) is

$$w_\gamma = \sqrt{\frac{2\widehat{p}_\gamma^{\,\mathrm{bin}}(1-\widehat{p}_\gamma^{\,\mathrm{bin}})\ln(3|\mathcal{G}|/\delta)}{|\mathcal{R}_\gamma|}} + \frac{3\ln(3|\mathcal{G}|/\delta)}{|\mathcal{R}_\gamma|}. \tag{10}$$

The first term reflects observed outcome variability, while the second prevents the width from vanishing when all observed outcomes agree. The case count controls both terms, and the logarithmic factor accounts for simultaneous coverage across probes. The empirical-Bernstein width requires only i.i.d. reference cases and no model for how outcomes vary over the feature space; by adapting to the observed variance, it can be tighter than a Hoeffding bound (Hoeffding, 1963) when outcomes mostly agree. SACA reports the regional estimate when $|\mathcal{R}_\gamma| \geq 2$ and $w_\gamma \leq \varepsilon$; otherwise, it abstains. Theorem 5 guarantees that, with probability at least $1-\delta$, all reported regional estimates simultaneously have absolute error at most ε.

The width also gives a sufficient case-count condition. Since $\widehat{p}_\gamma^{\,\mathrm{bin}}(1-\widehat{p}_\gamma^{\,\mathrm{bin}}) \leq 1/4$, for $0 < \varepsilon \leq 1$ the reporting criterion is guaranteed whenever $|\mathcal{R}_\gamma| \geq \left\lceil \frac{9}{2}\,\varepsilon^{-2}\ln(3|\mathcal{G}|/\delta) \right\rceil$. Because this condition depends only on the case count, it can guide acquisition before candidate outcomes are known (Section 3.3). It is sufficient rather than necessary: a region with fewer cases may still satisfy $w_\gamma \leq \varepsilon$. Proposition 4 extends these guarantees to prevalence and directional summaries, which SACA reports only when their error bounds are at most ε.

3.3 Support-Guided Acquisition

To make more regional estimates reportable, SACA targets the support deficits revealed by abstention. Let $\mathcal{U}$ be a pool of cases with known features and unknown outcomes, B the acquisition

budget, and $N_\gamma(S)$ the number of cases in $S \subseteq \mathcal{U}$ that fall in $\mathcal{R}_\gamma$. Using the sufficient reporting count derived from Eq. (10), SACA selects $|S| \leq B$ to maximize

$$f(S) = \sum_{\gamma \in \mathcal{G}} \min\Big\{ N_\gamma(S),\, \big[\big\lceil \tfrac{9}{2}\, \varepsilon^{-2} \ln(3|\mathcal{G}|/\delta) \big\rceil - |\mathcal{R}_\gamma| \big]_+ \Big\}. \tag{11}$$

Here $[a]_+ = \max\{a, 0\}$. A case may belong to multiple probe regions, so acquiring its outcome increases the case count in each and can reduce several shortfalls at once. For each region, f counts only the cases needed to reach the sufficient reporting count; further cases add no value for that region. This directs the budget toward cases that help multiple regions still below the threshold, without rewarding redundant observations.

Coverage-based methods seek to represent the pool (Yehuda et al., 2022), while uncertainty-based methods seek to improve prediction (Bae et al., 2025). Neither directly targets support deficits around a query. In SACA, a candidate's value depends only on its location and current counts. Either outcome reduces the shortfall in the under-supported regions it covers, helping more regions reach a count sufficient for reporting. Although selection is outcome-independent, acquired labels are incorporated into subsequent regional estimates and reporting decisions.

By selecting at each step the candidate with the largest marginal gain in f, SACA achieves a $(1 - 1/e)$-approximation to the maximum deficit reduction under budget B because f is monotone submodular (Proposition 5). Appendix C.5 provides a lazy implementation, a batched variant, and a label-aware early-stopping extension.

4 Theoretical Analysis

Our analysis addresses three questions: Can the evidence structure in Section 3.1 be recovered from finite data? When does the sample support an accurate regional estimate (Section 3.2)? Can acquisition efficiently improve regional support (Section 3.3)? Formal theorems, assumptions, and proofs appear in Appendix B, together with extensions to estimated probe subspaces, multiple queries, adaptive stopping, and time-uniform reporting; Figure 4 gives an illustrative roadmap.

Recovering the evidence structure. Theorem 2 shows that the kernel estimator in Eq. (5) recovers the pointwise landscape uniformly:

$$\sup_{(d,\boldsymbol{\theta})} |\widehat{p}(d, \boldsymbol{\theta}) - p_0(d, \boldsymbol{\theta})| = O_{\mathbb{P}}\Big((\log n/n)^{1/(m+2)} \Big). \tag{12}$$

The rate is governed by the effective local sample size $nh\kappa^{-(m-1)/2}$, the number of cases effectively contributing to each probe estimate, where h and κ are the radial bandwidth and directional concentration in Eq. (5); finer resolution reveals more details but uses fewer nearby observations.

Theorem 3 gives a matching lower bound of order $n^{-1/(m+2)}$, so the rate is optimal up to logarithmic factors. Proposition 3 carries a uniform landscape error ε_p to the perspectives of Section 3.1: prevalence $\bar{A}$ is estimated within ε_p, and the directional distribution q and entropy H within $O(\varepsilon_p/\mu)$, where μ lower-bounds the total nearby evidence, so these perspectives converge at the landscape rate. Existing scores of local structure, such as data-complexity measures (Ho & Basu, 2002), hubness (Radovanović et al., 2010), and reachable-set scores (Kothari et al., 2024), describe the neighborhood of a query but come with no estimation rate for the counterfactual landscape.

Reach is the distance at which $F_{\boldsymbol{\theta}}$ first attains α, so its error depends on how steeply $F_{\boldsymbol{\theta}}$ crosses that level; let $\lambda > 0$ be this crossing slope near $d^*(\boldsymbol{\theta})$. Theorem 4 shows that the reach error satisfies

$$\Big| \widehat{d}^*(\boldsymbol{\theta}) - d^*(\boldsymbol{\theta}) \Big| \leq \frac{\varepsilon_F}{\lambda}, \qquad \varepsilon_F = \sqrt{\frac{\log(2T/\delta)}{2n}}, \tag{13}$$

simultaneously over all fixed directions with such a slope, with probability at least $1 - \delta$: an $n^{-1/2}$ bound, tighter for steeper crossings.

Knowing when the data are sufficient. The rate above describes how the error shrinks as n grows, but reporting a regional estimate requires a bound computable from the sample at hand. Theorem 5 establishes that the empirical-Bernstein width w_γ in Eq. (10) bounds the error of every eligible regional estimate simultaneously: with probability at least $1 - \delta$,

$$\Big| \widehat{p}_\gamma^{\,\mathrm{bin}} - p_\gamma \Big| \leq w_\gamma \qquad \text{for all } \gamma \in \mathcal{G} \text{ with } |\mathcal{R}_\gamma| \geq 2. \tag{14}$$

Proposition 4 turns the regional widths into sample-computable certificates for the perspectives of Section 3.1: a prevalence estimate aggregated from regional estimates has error at most the weighted average of their widths w_γ, and the directional distribution has total-variation error at most the total width relative to the total estimated evidence. Both bounds shrink as the aggregated regions gain cases or their outcomes become more consistent, and SACA reports a perspective only when its bound is at most ε. In sum, prevalence and direction have both a convergence rate and a sample-computable certificate, and reach has a finite-sample bound.

Turning insufficient support into data acquisition. The regional guarantee in Eq. (14), together with the width in Eq. (10), yields a sufficient reporting count. SACA leverages Eq. (11) to extract the cases that effectively reduce the resulting support deficits. Proposition 5 shows that f is monotone submodular: a case can support overlapping regions, while its marginal value decreases as their shortfalls are filled. Selecting the case with the largest marginal gain, i.e., the largest reduction in total support deficit, at each step under budget B yields a set S_B satisfying

$$f(S_B) \geq (1 - 1/e) \max_{S \subseteq \mathcal{U},\ |S| \leq B} f(S). \tag{15}$$

If the pool can meet every sufficient reporting count, the same rule does so within a logarithmic factor of the minimum number of acquisitions.

5 Experiments

We evaluate SACA on six real-world OpenXAI benchmarks (*Adult*, *COMPAS*, *German*, *HELOC*, *Heart*, and *Pima*) and five synthetic Gaussian-mixture settings (*iso*, *aniso*, *multi*, *ring*, and *imb*) with closed-form posteriors for controlled evaluation (Agarwal et al., 2022; Wielopolski et al., 2024; Nemecek et al., 2025). Following established counterfactual benchmarks (Pawelczyk et al., 2021; Guidotti, 2024), we use training cases as the reference sample and test cases as queries. We ask four questions: **(RQ1)** Do the evidence perspectives reveal structure beyond nearest-counterfactual distance? **(RQ2)** How accurately does SACA recover the landscape, and does its report-or-abstain rule identify reliable regional estimates? **(RQ3)** How do direction and reach differ in endpoint selection? **(RQ4)** How effectively and efficiently does support-guided acquisition increase the proportion of reportable regions? Appendix D details the experimental protocols, baselines, hyperparameters, and prespecified settings, while Appendix E provides per-dataset results and further analyses.

RQ1: Characterization of Local Evidence Structures. We examine what prevalence, direction, and reach reveal beyond nearest-counterfactual distance in real-world data. Figure 3(a–c) illustrates these perspectives on a two-feature projection of HELOC: prevalence shows where different outcomes are common, direction identifies where they concentrate, and existence-level reach shows how far one must look to find them. Figure 3(d–f) compares three queries with nearly identical nearest-counterfactual distances, yet their prevalence ranges from nearly zero to 0.3, and their radial profiles differ in level and shape. The leading ranks of their sorted directional profiles differ, although later ranks converge. Thus, the same nearest-counterfactual distance can correspond to an approval supported by many comparable cases or an isolated approval with little supporting evidence. Table 1 shows that this distinction persists across datasets. German has about 21 times Adult's mean prevalence despite similar existence-level reach attainment, while COMPAS attains reach in more directions than German despite much lower prevalence. Attaining reach in many directions therefore does not imply that different outcomes are common. Appendix E.1 provides visualizations for all six datasets, comparisons with classical data-profiling measures, and query-level comparisons with existing counterfactual methods.

RQ2: Landscape Estimation Accuracy and Regional Reporting Reliability. We evaluate two targets: recovery of the pointwise landscape p_0 and reliability of regional estimates. For landscape recovery, we compare SACA with a binned proxy based on regional sample proportions, radial-only smoothing, broad radial smoothing, and logistic regression. We measure MAE against the closed-form p_0 of five synthetic mixtures (Appendix D.2). In Table 2, SACA has the lowest aggregate MAE, improving over the binned proxy by about 22%. Both smoothing ablations have much larger errors, especially in *multi* and *ring*; logistic regression performs well on simple landscapes but poorly in the multimodal and ring settings. For regional reporting, the regional proportion instead estimates p_γ, the different-outcome probability within a fixed probe region, and its empirical-Bernstein width determines whether to report or abstain. Table 3 shows that abstained probes have several times

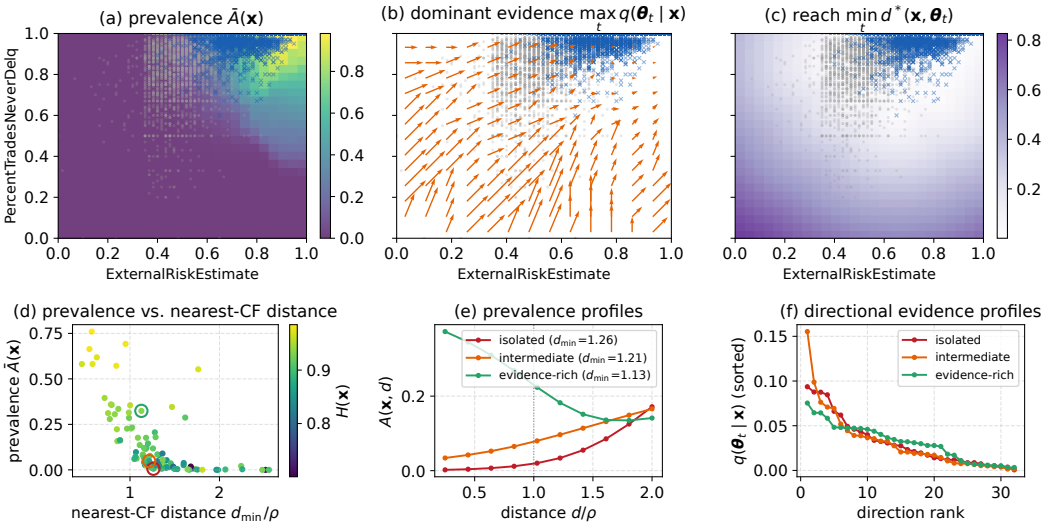


Figure 3: Evidence structures in HELOC. (a)–(c) show prevalence, dominant directional evidence, and existence-level reach on a two-feature projection, using labels from the full-space classifier. Shading represents each quantity; crosses mark different-outcome cases and dots mark same-outcome cases. In (d), each query is plotted by its normalized nearest-counterfactual distance $d_{\min}/\rho$ and prevalence, with color indicating directional entropy H. The three circled queries have similar $d_{\min}/\rho$; (e) shows their radial prevalence profiles and (f) their sorted directional evidence profiles.

Table 1: Dataset-level evidence profiles.

Dataset	n	m	$\bar{A}$	H	Att.
Adult	36,177	13	0.034	0.71	0.66
HELOC	7,896	23	0.140	0.90	0.73
COMPAS	4,937	7	0.100	0.79	0.75
Heart	2,926	15	0.066	0.82	0.53
German	800	60	0.717	1.00	0.65
Pima	614	8	0.147	0.91	0.69

Table 2: Landscape MAE vs. closed-form p_0.

Method	*iso*	*aniso*	*multi*	*ring*	*imb*	Mean
SACA	0.030	**0.022**	**0.059**	**0.025**	**0.011**	**0.029**
Binned	0.039	0.031	0.072	0.030	0.015	0.037
Radial	0.211	0.227	0.281	0.275	0.070	0.212
Broad	0.104	0.134	0.224	0.267	0.080	0.162
Logistic	**0.006**	0.032	0.379	0.390	0.015	0.164

Table 3: Report-or-abstain rule: reported vs. abstained probes.

ε	n	Reported	Rep. MAE	Abst. MAE
0.10	2,000	0.01	0.004	0.044
0.10	8,000	0.25	0.008	0.027
0.10	32,000	0.70	0.007	0.022
0.20	2,000	0.10	0.012	0.047
0.20	8,000	0.57	0.012	0.036
0.20	32,000	0.86	0.009	0.032

the MAE of reported probes against reference regional probabilities. Larger samples increase the reported fraction, while a looser tolerance permits more reports with higher average error. Theorem 5 guarantees simultaneous high-probability accuracy for reported estimates under its conditions, without making an accuracy claim about abstained estimates. Additional estimators, coverage and violation results, ablations, and perspective-level comparisons appear in Appendix E.2.

RQ3: Complementary Properties of Direction and Reach. SACA direction selects the dominant direction of the distance-weighted landscape, while SACA reach selects the direction with the shortest finite crossing. We compare them with a uniformly random direction (Marsaglia, 1972), local support (Gemou et al., 2026), path geometry (Sokol et al., 2025), feature responsiveness (Cheon et al., 2025), and a posterior-gradient oracle computed from the known posterior (Baehrens et al., 2010). When a selected direction attains $\alpha = 0.02$, its first crossing serves as the endpoint. We measure the endpoint's different-outcome posterior and distance normalized by ρ, and report attainment as the fraction of query–seed evaluations in which the selected direction reaches the level. Table 4 shows that SACA direction yields the highest endpoint posterior. Its paired gains are modest but consistent, including against the posterior-gradient oracle, which uses a local gradient rather than aggregating the landscape over distance. SACA reach yields shorter finite crossings but lower endpoint posteriors. As α increases from $1/n$ to 0.10, fitted crossings become steeper and estimated-reach MAE falls at every sample size, while reference-direction attainment drops by more than 80% (Table 5). The association between steeper crossings and lower error is consistent with Theorem 4; a fitted slope alone, however, does not establish the theorem's required lower bound throughout a neighborhood of the crossing. Per-scenario results and paired comparisons appear in Appendix E.3.

Table 4: Endpoint quality at evidence level $\alpha = 0.02$.

Method	Post.↑	Dist./ρ ↓	Att.↑
Random dir.	0.755	1.65	0.50
Local support	0.764	1.48	**0.87**
Path geometry	0.718	1.56	0.80
Respons.	0.780	1.50	0.86
Posterior-gradient oracle	0.770	1.50	0.82
SACA direction	**0.791**	1.49	0.84
SACA reach	0.760	**1.45**	**0.87**

Table 5: Reach error by evidence level α.

			Reach MAE with n		
α	Slope↑	Att.↑	2,000	8,000	32,000
$1/n$	0.003	0.83	0.183	0.136	0.101
0.01	0.027	0.64	0.077	0.043	0.023
0.02	0.043	0.57	0.065	0.037	0.019
0.05	0.075	0.39	0.057	0.035	0.016
0.10	0.107	0.14	0.055	0.031	0.015

Table 6: Support-guided acquisition.

Method	Cov.↑	AUC ↑	Time↓
SACA	**0.407**	**0.294**	**0.0053**
EWR	0.343	0.279	0.1142
Unc. Herding	0.237	0.219	6.1237
BADGE	0.236	0.218	0.2238
ProbCover	0.239	0.220	0.0570
TypiClust	0.238	0.217	0.1170
Random	0.237	0.218	0.0068

RQ4: Coverage and Efficiency of Support-Guided Acquisition. Table 6 tests whether acquiring cases targeted to regional support deficits makes more probes reportable. On each benchmark, we hide labels for a pool of up to $5{,}000$ training cases and acquire $B = 300$ of them. We compare SACA with eight baselines: Uncertainty Herding (Bae et al., 2025), BADGE (Ash et al., 2020), and classifier margin (Lewis & Gale, 1994), which use classifier-guided selection signals; TypiClust (Hacohen et al., 2022), ProbCover (Yehuda et al., 2022), and Facility (Gonzalez, 1985), which prioritize feature-space representativeness or coverage; EWR, which adapts expected-reduction selection (Roy & McCallum, 2001) to empirical-Bernstein width (Maurer & Pontil, 2009); and random acquisition. We report width-qualified coverage, the fraction of probes whose observed width is at most ε, its normalized AUC over the acquisition budget, and selection time. Meeting this width threshold is an operational reporting criterion, not a post-selection confidence guarantee. SACA qualifies about 19% more probes than EWR and about 72% more than the other strategies, while selecting cases about 22 times faster than EWR. SACA targets regions with unfilled support deficits and updates acquisition scores only for candidates sharing an affected region; EWR recomputes expected width reductions after each acquired label. The coverage gains are consistent with the count-deficit mechanism analyzed in Proposition 5. Additional baselines and ablations appear in Appendix E.4.

6 Related Work

Counterfactual methods optimize or retrieve alternatives (Wachter et al., 2018; Keane & Smyth, 2020), emphasizing plausibility (Poyiadzi et al., 2020; Tan et al., 2026), diversity (Mothilal et al., 2020), or actionability (Ustun et al., 2019). Related guarantees address prediction robustness under input or model perturbations or multiplicity (Maragno et al., 2024; Jiang et al., 2024a; Turbal et al., 2025), while other work connects local data support to robustness and prediction changes (Pawelczyk et al., 2020b; Gemou et al., 2026). SACA instead characterizes counterfactual evidence across distance and direction and assesses its statistical support. We establish a landscape estimation rate optimal up to logarithmic factors, finite-sample reach bounds under a crossing condition, and simultaneous accuracy guarantees for reported regional estimates, with abstention when the requested accuracy is unsupported. Beyond estimation, SACA targets support deficits in fixed query-centered regions, unlike active learning methods that prioritize predictive uncertainty or feature-space coverage (Sener & Savarese, 2018; Yehuda et al., 2022; Ash et al., 2020; Bae et al., 2025). Candidate values depend on their locations and remaining deficits rather than unknown outcomes or predictive uncertainty. The resulting monotone submodular objective yields a $(1 - 1/e)$-approximation under a fixed budget. Appendix F provides detailed comparisons with related work.

7 Conclusion

The distance to the nearest counterfactual cannot tell whether it is supported by many similar cases or stands alone. Support-Aware Counterfactual Analysis (SACA) characterizes the surrounding evidence across distance and direction and reports regional estimates only when the sample supports the requested accuracy. We establish a landscape estimation rate optimal up to logarithmic factors, reach stability under a crossing condition, simultaneous guarantees for reported regions, and a submodular objective for reducing support deficits. Experiments reveal distinct evidence structures despite similar nearest distances, accurate landscape recovery and regional reporting, and improved reportability through acquisition. Extending this observational framework to causal and actionable settings requires modeling feasible interventions and their effects.

AI USE STATEMENT

We used generative AI to help write and debug code for experiments and figures, and to improve the grammar and clarity of the manuscript. We did not use it to generate data, develop the theory or methods, or analyze and interpret results. The authors reviewed and tested all AI-assisted code, verified all edited text, and take full responsibility for the paper.

ETHICS STATEMENT

The experiments use synthetic data and OpenXAI benchmarks; acquisition reveals withheld outcomes from fixed pools. On real data, outcomes are classifier predictions, so SACA describes model behavior and may reflect biases in the models or reference data. Abstention makes weak support visible, but may occur more often for underrepresented groups. The guarantees establish statistical estimation under stated assumptions, not causality, fairness, or feasible recourse. Consequential uses require domain-specific validation, group-level assessment, and actionable constraints. Regional summaries carry no formal privacy guarantee, and any further data collection must respect consent, access restrictions, and dataset terms.

REPRODUCIBILITY STATEMENT

Sections 2–4 define SACA, describe its estimation and acquisition procedures, and summarize the theoretical results. Appendix B gives the assumptions and proofs; Appendices C and D document the algorithms and experimental setup; and Appendix E provides per-dataset and supplementary analyses. Code is available at `https://anonymous.4open.science/r/ICLR27_SACA-962A`.

REFERENCES

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15784–15799, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/65398a0eba88c9b4a1c38ae405b125ef-Abstract-Datasets_and_Benchmarks.html`.

Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. Faithful model explanations through energy-constrained conformal counterfactuals. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 10829–10837, 2024. doi: 10.1609/aaai.v38i10.28956.

Asiful Arefeen, Shovito Barua Soumma, and Hassan Ghasemzadeh. RealAC: A domain-agnostic framework for realistic and actionable counterfactual explanations. *arXiv preprint arXiv:2508.10455*, 2025.

André Artelt and Barbara Hammer. Convex density constraints for computing plausible counterfactual explanations. In *Artificial Neural Networks and Machine Learning – ICANN 2020*, volume 12396 of *Lecture Notes in Computer Science*, pp. 353–365. Springer, 2020.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations (ICLR)*, 2020.

Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Wonho Bae, Danica J. Sutherland, and Gabriel L. Oliveira. Uncertainty herding: One active learning method for all label budgets. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=UgPoHhYQ2U.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(61):1803–1831, 2010. URL https://jmlr.org/papers/v11/baehrens10a.html.

Tom Bewley, Salim I. Amoukou, Saumitra Mishra, Daniele Magazzeni, and Manuela Veloso. Counterfactual metarules for local and global recourse. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3707–3724. PMLR, 2024. URL https://proceedings.mlr.press/v235/bewley24a.html.

Aman Bilkhoo, Mehran Hosseini, Milad Kazemi, and Nicola Paoletti. CONFEX: Uncertainty-aware counterfactual explanations with conformal guarantees. *arXiv preprint arXiv:2510.19754*, 2025.

Szymon Bobek, Łukasz Bałec, and Grzegorz J. Nalepa. Actionable and diverse counterfactual explanations incorporating domain knowledge and plausibility constraints. *arXiv preprint arXiv:2511.20236*, 2025.

Dieter Brughmans, Pieter Leyman, and David Martens. NICE: an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*, 38(5):2665–2703, 2024. doi: 10.1007/s10618-023-00930-y.

Lena Marie Budde, Ayan Majumdar, Richard Uth, Markus Langer, and Isabel Valera. From universal to individualized actionability: Revisiting personalization in algorithmic recourse. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1166–1195, 2026. doi: 10.1145/3805689.3812251.

Ioannis Chatzis, Jason Liartis, Athanasios Voulodimos, and Giorgos Stamou. Optimal recourse summaries via bi-objective decision tree learning. *arXiv preprint arXiv:2605.07598*, 2026.

Hsi-Wen Chen, Jian Pei, De-Nian Yang, and Ming-Syan Chen. Counterfactual analysis on large graphs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 390–401, 2026. doi: 10.1145/3770855.3817647.

Seung Hyun Cheon, Anneke Wernerfelt, Sorelle A. Friedler, and Berk Ustun. Feature responsiveness scores: Model-agnostic explanations for recourse. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=wsWCVrH9dv.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. Counterfactual active learning for out-of-distribution generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 11362–11377. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.636.

Ahmad-Reza Ehyaei, Ali Shirali, and Samira Samadi. Collective counterfactual explanations: Balancing individual goals and collective dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 50485–50516, 2025. doi: 10.52202/085713-1685.

Oleksii Furman, Ulvi Movsum-zada, Patryk Marszałek, Maciej Zięba, and Marek Śmieja. DiCoFlex: Model-agnostic diverse counterfactuals with flexible control. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 71592–71623, 2025. doi: 10.52202/085713-2405.

Oleksii Furman, Patryk Wielopolski, Łukasz Lenkiewicz, Jerzy Stefanowski, and Maciej Zięba. Unifying perspectives: Plausible counterfactual explanations on global, group-wise, and local levels. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 571–579, 2026. doi: 10.24963/ijcai.2026/65.

Jingwei Gan, Shinan Zhang, Chi Zhang, and Andy Li. Automated counterfactual generation in financial model risk management. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pp. 4064–4068, 2021. doi: 10.1109/BigData52589.2021.9671561.

Ioanna Gemou, Matteo Gamba, Randall Balestriero, and Ritambhara Singh. A geometric view of counterfactual behavior: Interaction of boundary proximity and local support. *arXiv preprint arXiv:2606.04209*, 2026. doi: 10.48550/arXiv.2606.04209.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Margarita A. Guerrero and Cristian R. Rojas. ACE: Adapting sampling for counterfactual explanations. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference (L4DC)*, volume 331 of *Proceedings of Machine Learning Research*, pp. 242–264. PMLR, 2026. URL https://proceedings.mlr.press/v331/guerrero26a.html.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002. doi: 10.1007/b97848.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8175–8195. PMLR, 2022.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12351–12367. PMLR, 2023.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval abstractions for robust counterfactual explanations. *Artificial Intelligence*, 336:104218, 2024a.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI), Survey Track*, pp. 8086–8094, 2024b. doi: 10.24963/ijcai.2024/894.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2855–2862, 2020. doi: 10.24963/ijcai.2020/395.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1846–1870. PMLR, 2022.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Learning decision trees and forests with algorithmic recourse. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22936–22962. PMLR, 2024.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 353–362, 2021. doi: 10.1145/3442188.3445899.

Loukas Kavouras, Eleni Psaroudaki, Konstantinos Tsopelas, Dimitrios Rontogiannis, Nikolaos Theologitis, Dimitris Sacharidis, Giorgos Giannopoulos, Dimitrios Tomaras, Kleopatra Markou, Dimitrios Gunopulos, Dimitris Fotakis, and Ioannis Emiris. GLANCE: Global actions in a nutshell for counterfactual explainability. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 22545–22553, 2026. doi: 10.1609/aaai.v40i27.39414.

Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *Case-Based Reasoning Research and Development*, pp. 163–178. Springer, 2020.

Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 11575–11585, 2021.

Gunnar König, Hidde Fokkema, Timo Freiesleben, Celestine Mendler-Dünner, and Ulrike von Luxburg. Performative validity of recourse explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 139334–139370, 2025. doi: 10.52202/085713-4652.

Marcin Kostrzewa, Maciej Zięba, and Jerzy Stefanowski. A probabilistic consensus-driven approach for robust counterfactual explanations. *arXiv preprint arXiv:2604.17494*, 2026.

Avni Kothari, Bogdan Kulynych, Tsui-Wei Weng, and Berk Ustun. Prediction without preclusion: Recourse verification with reachable sets. In *International Conference on Learning Representations (ICLR)*, 2024.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2801–2807, 2019.

Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Achieving diversity in counterfactual explanations: a review and discussion. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1859–1869, 2023.

Seung C. Lee. Auditing algorithms: A rational counterfactual framework. *Journal of International Technology and Information Management*, 30(2):120–145, 2021.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994.

Dan Ley, Saumitra Mishra, and Daniele Magazzeni. GLOBE-CE: A translation based approach for global counterfactual explanations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19315–19342. PMLR, 2023.

Helena Löfström, Tuwe Löfström, Ulf Johansson, and Cecilia Sönströd. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246:123154, 2024. doi: 10.1016/j.eswa.2024.123154.

Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 5313–5322, 2022. doi: 10.1609/aaai.v36i5.20468.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, Ş Ilker Birbil, and Dick den Hertog. Finding regions of counterfactual explanations via robust optimization. *INFORMS Journal on Computing*, 36(5):1316–1334, 2024.

George Marsaglia. Choosing a point from the surface of a sphere. *The Annals of Mathematical Statistics*, 43(2):645–646, 1972. doi: 10.1214/aoms/1177692644.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 607–617, 2020.

Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Jiri Nemecek, Tomas Pevny, and Jakub Marecek. Generating likely counterfactuals using sum-product networks. In *International Conference on Learning Representations (ICLR)*, 2025.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Feasible recourse plan via diverse interpolation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4679–4698. PMLR, 2023.

Erfaun Noorani, Pasan Dissanayake, Faisal Hamman, and Sanghamitra Dutta. Counterfactual explanations for model ensembles using entropic risk measures. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1566–1575, 2025.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132, 2020a.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 809–818. PMLR, 2020b.

Martin Pawelczyk, Sascha Bielawski, Johan van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract-round1.html`.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020. doi: 10.1145/3375627.3375850.

Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12187–12198, 2020.

Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. MCCE: Monte Carlo sampling of valid and realistic counterfactual explanations for tabular data. *Data Mining and Knowledge Discovery*, 38(4):1830–1861, 2024. doi: 10.1007/s10618-024-01017-y.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 441–448, 2001.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 20–28, 2019.

Lisa Schut, Oscar Key, Rory McGrath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1756–1764. PMLR, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Michael R. Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256, 2014.

Kacper Sokol, Edward Small, and Yueqing Xuan. Navigating explanatory multiverse through counterfactual path geometry. *Machine Learning*, 114:144, 2025. doi: 10.1007/s10994-025-06769-2.

Jun Tan, Qing Guo, Zicheng Xu, Jinglin Li, Qi Fang, and Ning Gui. Density-guided robust counterfactual explanations on tabular data under model multiplicity. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. ElliCE: Efficient and provably robust algorithmic recourse via the Rashomon sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 123872–123893, 2025. doi: 10.52202/085713-4128.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 10–19, 2019.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 12976 of *Lecture Notes in Computer Science*, pp. 650–665. Springer, 2021. doi: 10.1007/978-3-030-86520-7_40.

Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. Feature-based learning for diverse and privacy-preserving counterfactual explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2211–2222, 2023. doi: 10.1145/3580305.3599343.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018. URL `https://jolt.law.harvard.edu/volumes/volume-31`.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1954–1963. PMLR, 2015.

Patryk Wielopolski, Oleksii Furman, Jerzy Stefanowski, and Maciej Zięba. Probabilistically plausible counterfactual explanations with normalizing flows. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024. doi: 10.3233/FAIA240584.

Laurence A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982. doi: 10.1007/BF02579435.

Yueqing Xuan, Kacper Sokol, Mark Sanderson, and Jeffrey Chan. Perfect counterfactuals in imperfect worlds: Modelling noisy implementation of actions in sequential algorithmic recourse. *Machine Learning*, 114(8):187, 2025.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 22354–22367, 2022.

Kun Zhang, Chong Chen, Yuanzhuo Wang, Qi Tian, and Long Bai. CFGL-LCR: A counterfactual graph learning framework for legal case retrieval. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3332–3341, 2023.

Wei Zhang, Brian Barr, and John Paisley. Tabular diffusion counterfactual explanations. *Frontiers in Artificial Intelligence*, 2026. doi: 10.3389/frai.2026.1743495.

Chenrui Zhu, Louenas Bounia, Vu Linh Nguyen, Sébastien Destercke, and Arthur Hoarau. Robust explanations through uncertainty decomposition: A path to trustworthier AI. In *Pattern Recognition*, volume 16815 of *Lecture Notes in Computer Science*, pp. 204–219. Springer, 2027. doi: 10.1007/978-3-032-31663-9_14.

A NOTATION AND TERMINOLOGY

Table 7: Notation used throughout the paper.

Symbol	Definition / Convention
Data and query geometry	
$\mathcal{D}$, n, m	Labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$; sample size; feature dimension.
$\mathbf{x}$, $\mathbf{z}$, y	Query and generic feature vectors in $\mathcal{X} \subseteq \mathbb{R}^m$; query outcome $y \in \mathcal{Y}$.
$(\mathbf{X}, Y)$	Random feature–outcome pair; outcomes are labels or fixed-rule predictions.
$r(\mathbf{z})$, $\mathbf{u}(\mathbf{z})$	Query-centered radius and unit direction for $\mathbf{z} \neq \mathbf{x}$ (Eq. (1)).
d, $\boldsymbol{\theta}$, Γ	Probe radius $d > 0$ and direction $\boldsymbol{\theta} \in \mathbb{S}^{m-1}$; smooth domain $\Gamma = [\underline{d}, \overline{d}] \times \mathbb{S}^{m-1}$.
$\underline{d}$, $\overline{d}$	Chosen lower and upper analysis distances, with $0 < \underline{d} \leq \overline{d} < \infty$.
$d_{\min}(\mathbf{x})$	Nearest observed different-outcome distance; $+\infty$ if no such point exists.
Landscape and reach	
γ, $\mathcal{R}_\gamma$	Probe $\gamma = (d, \boldsymbol{\theta})$ and its distance–direction region, also written $\mathcal{R}(\mathbf{x}, d, \boldsymbol{\theta})$ (Eq. (2)).
ϵ_d, ϵ_θ, ψ	Positive radial/angular resolutions; reach cone half-angle $\psi \in (0, \pi/2]$, chosen independently of ϵ_θ.
$p_{\epsilon_d,\epsilon_\theta}$, p_0	Regional / pointwise different-outcome probability in $[0, 1]$; $p_{\epsilon_d,\epsilon_\theta}$ is the main text's p_γ (Eq. (4)).
$\eta(\mathbf{z})$	Different-outcome posterior $\Pr(Y \neq y \mid \mathbf{X} = \mathbf{z})$; equals $p_0(\mathbf{x}, d, \boldsymbol{\theta})$ at $\mathbf{z} = \mathbf{x} + d\boldsymbol{\theta}$ (Thm. 1).
F, $\widehat{F}$	Cumulative cone mass, $F_{\boldsymbol{\theta}}(d)$ in the main text, and its empirical estimate (Eqs. (8), (22)).
d^*, $\widehat{d}^*$, α	Reach / empirical reach: first radius where F / $\widehat{F}$ attains level $\alpha \in (0, 1)$; $+\infty$ if unattained.
Estimation and certification	
$\widehat{p}^{\text{bin}}$, $\widehat{p}$	Binned regional and smooth kernel estimates, in $[0, 1]$ when defined (Section 3.2 and Eq. (5)).
K_h, L_κ	Radial/directional kernels, written K and L in the main text; bandwidth $h > 0$ and concentration $\kappa \geq 0$.
$\mathcal{I}_{\mathbf{x}}$, Ω_γ, $\omega_{\min}$	Reference indices $\mathcal{I}_{\mathbf{x}} = \{i : \mathbf{x}_i \neq \mathbf{x}\}$; expected kernel mass Ω_γ; minimum mass $\omega_{\min}(h, \kappa) = \inf_{\gamma \in \Gamma} \Omega_\gamma > 0$.
$\mathcal{G}$, T, n_d, Q	Finite probe grid; numbers of directions $T \geq 2$, radii $n_d \geq 1$, and prespecified queries $Q \geq 1$.
ε, δ, w_γ	Reporting tolerance $\varepsilon \in (0, 1]$, failure probability $\delta \in (0, 1)$, and certificate width w_γ (Eq. (10)); $w_\gamma = \infty$ for fewer than two observations.
ε_p, ε_F	Nonnegative uniform-error bounds for $\widehat{p}$ and $\widehat{F}$, respectively.
λ, r_0, μ	Reach crossing margin $\lambda > 0$ and radius $r_0 > 0$; evidence floor $\mu \in (0, 1]$.
P, $\widehat{P}$, k, ξ	Target/estimated rank-k projectors; error bound $\|\widehat{P} - P\|_{\text{op}} \leq \xi < 1$.
$\widehat{Z}$, E	Estimated total regional directional evidence and propagated error; certification of normalized perspectives requires $\widehat{Z} > E$.
Evidence perspectives and adaptive evaluation	
ρ	Local scale $\rho \in (\underline{d}, \overline{d}]$ that defines the nearby distance range (Section 3.1).
A, $\bar{A}$	Direction-averaged prevalence profile and its radial average over $[\underline{d}, \rho]$, in $[0, 1]$ (Eq. (6)).
A_T, $\bar{A}_T$, IA_T	Grid prevalence profile/scalar; inverse prevalence $\text{IA}_T = 1/\bar{A}_T$ for $\bar{A}_T > 0$.
q, q_T, H_T	Distance-weighted directional density, grid probability vector, and normalized entropy $H_T \in [0, 1]$ (written H in the main text) (Eqs. (7)–(21)).
$R_1 < \cdots < R_{n_R}$	Prespecified positive search radii; n_R is the schedule length.
$\mathcal{N}$, $\bar{A}_{\text{ball}}(\mathbf{x}; R)$	Retrieved neighborhood / in-ball different-outcome proportion in $[0, 1]$.
Support-guided acquisition	
$\mathcal{U}$, S, B	Unlabeled pool with known features; acquired subset $S \subseteq \mathcal{U}$; budget $B \in \mathbb{N}_0$ with $\|S\| \leq B$.
$N_{\mathcal{D}}(\gamma)$, $N_S(\gamma)$	Reference and acquired regional counts ($\|\mathcal{R}_\gamma\|$ and $N_\gamma(S)$ in the main text).
n_ε, D_γ, D	Count threshold ($n_\varepsilon^{\text{suff}} = \lceil \frac{9}{2}\varepsilon^{-2} \ln(3\|\mathcal{G}\|/\delta) \rceil$ in the theory); deficit $D_\gamma = [n_\varepsilon - N_{\mathcal{D}}(\gamma)]_+$; total $D = \sum_\gamma D_\gamma$.
$f(S)$, $C(S)$	Deficit-reduction objective and count-qualified fraction (Eqs. (11), (24)).
$C_{\text{width}}(S)$, b	Width-qualified fraction in $[0, 1]$; batch size $b \in \mathbb{N}$ between gain updates.

Availability. Regional quantities require positive conditioning support; kernel estimates require positive kernel mass; directional normalization requires positive total evidence. Prevalence and direction summarize p_0; the reach summarizes cumulative mass F.

Table 8: Terminology used throughout the paper, with the section in which each term is introduced.

Term	Meaning	Where
Query	The instance–outcome pair $(\mathbf{x}, y)$ under analysis.	Sec. 2
Reference sample	The labeled dataset $\mathcal{D}$ against which the query is compared.	Sec. 2
Different-outcome (counterfactual) evidence	Reference cases whose outcome differs from y.	Sec. 2
Probe	A distance–direction pair $\gamma = (d, \boldsymbol{\theta})$ at which evidence is examined.	Sec. 2
Region	The set $\mathcal{R}_\gamma$ of cases within ϵ_d of distance d and within angle ϵ_θ of $\boldsymbol{\theta}$.	Sec. 2
Resolution	The radial and angular half-widths $(\epsilon_d, \epsilon_\theta)$ defining a region.	Sec. 2
Probe grid	The prespecified finite set $\mathcal{G}$ of n_d distances and T directions.	Sec. 2
Distance scale	The analyst-chosen ρ determining which distances count as nearby.	Sec. 3.1
Counterfactual landscape	The probability of a different outcome as a function of probe: regional (p_γ) or pointwise (p_0).	Sec. 3.1
Kernel / binned estimator	$\widehat{p}$, the smooth estimate of p_0; $\widehat{p}^{\,\mathrm{bin}}_\gamma$, the regional proportion estimating p_γ.	Sec. 3.1
Prevalence	The direction-averaged landscape $A(d)$ and its nearby average $\bar{A}$; inverse prevalence $\mathrm{IA} = 1/\bar{A}$.	Sec. 3.1
Direction	The direction distribution q of nearby evidence and its normalized entropy H.	Sec. 3.1
Reach	The first distance $d^*(\boldsymbol{\theta})$ at which the cumulative different-outcome mass $F_{\boldsymbol{\theta}}$ in a cone attains level α.	Sec. 3.1
Existence level	The choice $\alpha = 1/n$, under which the reach equals the distance to the nearest different-outcome case in the cone.	Sec. 3.1
Width	The empirical-Bernstein half-width w_γ of a probe's confidence interval.	Sec. 3.2
Certified / reported	A probe or summary whose width or propagated bound is at most the accuracy ε.	Sec. 3.2
Abstain	The decision not to report a probe or summary whose bound exceeds ε.	Sec. 3.2
Support	The number $\lvert\mathcal{R}_\gamma\rvert$ of reference cases in a region.	Sec. 3.2
Support deficit	$D_\gamma = [n_\varepsilon - \lvert\mathcal{R}_\gamma\rvert]_+$, the shortfall relative to the sufficient reporting count n_ε.	Sec. 3.3
Acquisition	Selecting a set $S \subseteq \mathcal{U}$ of unlabeled cases, $\lvert S\rvert \le B$, whose outcomes are queried to reduce deficits.	Sec. 3.3

B Detailed Proof

This appendix gives the assumptions, proofs, and extensions of the results summarized in Section 4. We make the query $(\mathbf{x}, y)$ explicit in $p_{\epsilon_d,\epsilon_\theta}$, $\widehat{p}$, $\widehat{p}^{\,\mathrm{bin}}$, F, and d^*, so that $p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta})$ and $F(\mathbf{x}, d, \boldsymbol{\theta})$ denote the main text's p_γ and $F_{\boldsymbol{\theta}}(d)$; write $N_{\mathcal{D}}(\gamma)$, $N_S(\gamma)$, and $n_\varepsilon^{\mathrm{suff}}$ for the main text's $\lvert\mathcal{R}_\gamma\rvert$, $N_\gamma(S)$, and sufficient reporting count $\lceil \frac{9}{2}\varepsilon^{-2}\ln(3\lvert\mathcal{G}\rvert/\delta)\rceil$, respectively; and abbreviate $r_i = r(\mathbf{x}_i)$, $\mathbf{u}_i = \mathbf{u}(\mathbf{x}_i)$, and $\mathcal{I}_{\mathbf{x}} = \{i : \mathbf{x}_i \neq \mathbf{x}\}$ for the reference cases that do not coincide with the query. Figure 4 summarizes the complete set of theoretical results and is read top to bottom. The three estimators at the top, the kernel landscape $\widehat{p}$, the empirical cumulative mass $\widehat{F}$, and the binned proportion $\widehat{p}^{\,\mathrm{bin}}$, feed four lanes: *estimable* (how fast each is recovered from data), *interpretable* (what the landscape and its perspectives mean in the population), *certifiable* (which finite-sample bounds hold at the sample in hand), and *acquisition* (how the bounds turn into a selection rule). Solid boxes are core results, dashed boxes extensions, arrows show which result feeds which, and the orange badges L, P, D, and R mark which quantity of Figure 1 a result focuses on: the landscape, prevalence, direction, or reach.

B.1 Population Geometry

Theorem 1 (Population geometry)**.** *Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(\mathbf{X}, Y)$. Suppose $d > 0$, $\Pr(Y = c) = \pi_c$, and $\mathbf{X} \mid Y = c$ has a class-conditional density f_c that is continuous in a neighborhood of $\mathbf{z}_0 = \mathbf{x} + d\boldsymbol{\theta}$. Assume that the marginal density $g(\mathbf{z}_0) = \sum_c \pi_c f_c(\mathbf{z}_0)$ is positive, and define*

$$\eta(\mathbf{z}) = \Pr(Y \neq y \mid \mathbf{X} = \mathbf{z}) = \frac{\sum_{c \neq y} \pi_c f_c(\mathbf{z})}{\sum_{c \in \mathcal{Y}} \pi_c f_c(\mathbf{z})}.$$

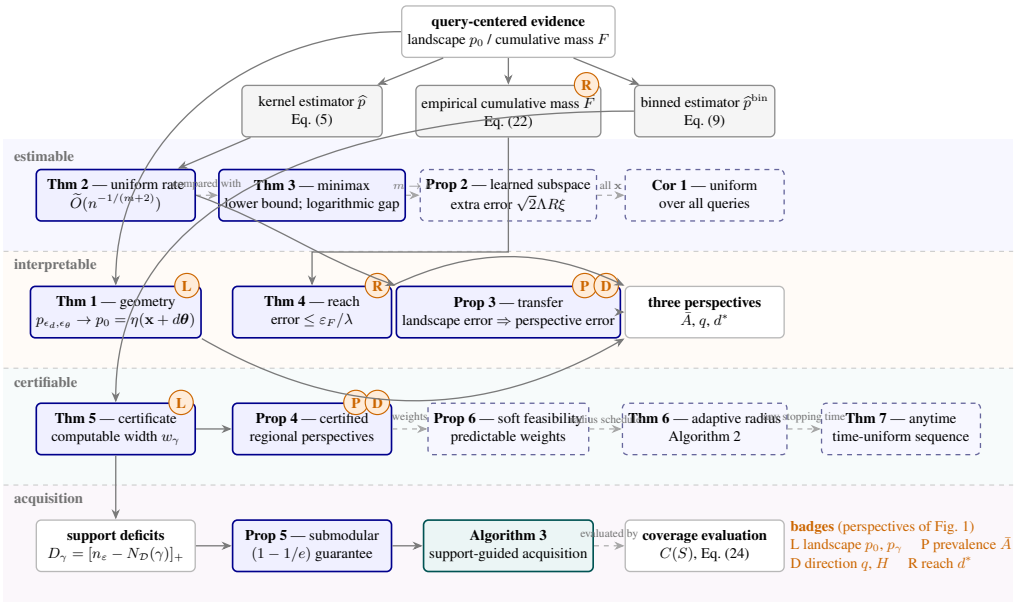


Figure 4: Theoretical roadmap, read top to bottom. The three estimators at the top feed four lanes: estimable, interpretable, certifiable, and acquisition. Solid blue boxes are core results, dashed boxes extensions, and teal boxes algorithms; arrows show which result feeds which. Orange badges name the object of Figure 1 that a result interprets (interpretable lane) or certifies (certifiable lane): L the landscape, P prevalence, D direction, R reach. Reach carries its own finite-sample certificate in Theorem 4, so its R badge sits in the interpretable lane.

Then the pointwise landscape satisfies

$$p_0(\mathbf{x}, d, \boldsymbol{\theta}) = \eta(\mathbf{x} + d\boldsymbol{\theta}).$$

Let $p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta})$ denote the regional target p_γ of Section 3 for the region in Eq. (2), with its dependence on the two resolutions made explicit. Then

$$\lim_{\epsilon_d,\epsilon_\theta \to 0} p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta}) = \eta(\mathbf{x} + d\boldsymbol{\theta}).$$

For each fixed pair of positive resolutions with nonzero regional probability,

$$\widehat{p}^{\text{bin}}(\mathbf{x}, d, \boldsymbol{\theta}) \xrightarrow[n\to\infty]{\text{a.s.}} p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta}).$$

If, in addition, the class-conditional densities are continuous and the marginal density is positive throughout a neighborhood of $\mathbf{x}$, and η is twice continuously differentiable there, then, after taking the resolution limit, as $d \to 0$,

$$\eta(\mathbf{x} + d\boldsymbol{\theta}) = \eta(\mathbf{x}) + d\nabla\eta(\mathbf{x})^\top\boldsymbol{\theta} + \frac{d^2}{2}\boldsymbol{\theta}^\top\nabla^2\eta(\mathbf{x})\boldsymbol{\theta} + o(d^2).$$

When $\nabla\eta(\mathbf{x}) \neq 0$, the unit direction maximizing the first-order term is $\nabla\eta(\mathbf{x})/\|\nabla\eta(\mathbf{x})\|_2$.

Proof. Let $\mathcal{R}_\epsilon = \mathcal{R}(\mathbf{x}, d, \boldsymbol{\theta})$, where $\epsilon = (\epsilon_d, \epsilon_\theta)$, and set $\mathbf{z}_0 = \mathbf{x} + d\boldsymbol{\theta}$. Bayes' rule gives

$$\Pr(Y \neq y \mid \mathbf{X} = \mathbf{z}) = \frac{\sum_{c\neq y}\pi_c f_c(\mathbf{z})}{\sum_{c\in\mathcal{Y}}\pi_c f_c(\mathbf{z})} = \eta(\mathbf{z})$$

whenever the denominator is positive. The denominator remains positive in a neighborhood of $\mathbf{z}_0$, and η is continuous there. Because $d > 0$, radial–angular coordinates are locally regular at $\mathbf{z}_0$. Consequently, for $V = (r(\mathbf{X}), \mathbf{u}(\mathbf{X}))$, a regular conditional version satisfies $\mathbb{E}[\mathbb{I}[Y \neq y] \mid V = (d, \boldsymbol{\theta})] = \eta(\mathbf{z}_0)$, which is the asserted identity for p_0.

For all sufficiently small ϵ, $\Pr(\mathbf{X} \in \mathcal{R}_\epsilon) > 0$. The tower property yields

$$p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta}) = \mathbb{E}[\eta(\mathbf{X}) \mid \mathbf{X} \in \mathcal{R}_\epsilon].$$

Every $\mathbf{z} \in \mathcal{R}_\epsilon$ satisfies

$$\|\mathbf{z} - \mathbf{z}_0\|_2 \le \epsilon_d + (d + \epsilon_d)\epsilon_\theta.$$

It follows that

$$|p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta}) - \eta(\mathbf{z}_0)| \le \sup_{\mathbf{z}\in\mathcal{R}_\epsilon} |\eta(\mathbf{z}) - \eta(\mathbf{z}_0)| \longrightarrow 0$$

as $\epsilon_d, \epsilon_\theta \to 0$. This proves the resolution limit.

For fixed ϵ, define $I_i = \mathbb{I}[\mathbf{X}_i \in \mathcal{R}_\epsilon]$ and $J_i = I_i \mathbb{I}[Y_i \neq y]$. The strong law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} I_i \xrightarrow{\text{a.s.}} \Pr(\mathbf{X} \in \mathcal{R}_\epsilon) > 0, \qquad \frac{1}{n}\sum_{i=1}^{n} J_i \xrightarrow{\text{a.s.}} \Pr(Y \neq y, \mathbf{X} \in \mathcal{R}_\epsilon).$$

The ratio of these limits is $p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, d, \boldsymbol{\theta})$, proving the asserted almost-sure convergence.

For the local expansion, the strengthened density assumptions make the preceding identity valid throughout a sufficiently small punctured neighborhood of $\mathbf{x}$. Define $\eta_{\boldsymbol{\theta}}(t) = \eta(\mathbf{x} + t\boldsymbol{\theta})$. The assumed C^2 regularity implies

$$\eta'_{\boldsymbol{\theta}}(0) = \nabla\eta(\mathbf{x})^\top \boldsymbol{\theta}, \qquad \eta''_{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}^\top \nabla^2 \eta(\mathbf{x}) \boldsymbol{\theta}.$$

Taylor's theorem gives the displayed expansion. If $\nabla\eta(\mathbf{x}) \neq 0$, the Cauchy–Schwarz inequality shows that the first-order term is maximized over the unit sphere by $\boldsymbol{\theta} = \nabla\eta(\mathbf{x})/\|\nabla\eta(\mathbf{x})\|_2$. □

B.2 Invariance of the Probe Geometry

Proposition 1 (Invariance under isometries and a common change of units)**.** *Let $U \in \mathbb{R}^{m\times m}$ be orthogonal and let $c > 0$. Transform the query, the reference cases, and the candidate pool by $\mathbf{z} \mapsto cQ\mathbf{z}$, map the probe grid by $(d, \boldsymbol{\theta}) \mapsto (cd, U\boldsymbol{\theta})$, and set $\epsilon'_d = c\,\epsilon_d$, $\epsilon'_\theta = \epsilon_\theta$, $h' = ch$, and $\kappa' = \kappa$. Write $\gamma' = (cd, U\boldsymbol{\theta})$ for the image of a probe $\gamma = (d, \boldsymbol{\theta})$. Then*

- *(i) $r'(\mathbf{z}') = c\,r(\mathbf{z})$ and $\angle(\mathbf{u}'(\mathbf{z}'), U\boldsymbol{\theta}) = \angle(\mathbf{u}(\mathbf{z}), \boldsymbol{\theta})$, so $\mathcal{R}'_{\gamma'} = cQ\,\mathcal{R}_\gamma$;*
- *(ii) $|\mathcal{R}'_{\gamma'}| = |\mathcal{R}_\gamma|$ and $\widehat{p}^{\,\text{bin}}_{\gamma'} = \widehat{p}^{\,\text{bin}}_{\gamma}$, hence $w'_{\gamma'} = w_\gamma$ and exactly the same probes are reported at any (ε, δ);*
- *(iii) $\widehat{p}'(cd, U\boldsymbol{\theta}) = \widehat{p}(d, \boldsymbol{\theta})$, the direction distribution satisfies $q'(U\boldsymbol{\theta}) = q(\boldsymbol{\theta})$, the normalized entropy H is unchanged, and the reach satisfies $\widehat{d}^{\,*\prime}(U\boldsymbol{\theta}) = c\,\widehat{d}^{\,*}(\boldsymbol{\theta})$;*
- *(iv) if the distance scale is estimated equivariantly, as the median k-th nearest-neighbor distance is, then $\rho' = c\rho$, and prevalence $\bar{A}$ together with every radius expressed in units of ρ is unchanged.*

Relabeling the features is the case in which U is a permutation matrix and $c = 1$, and a common change of units is the case $U = I$.

Proof. Because U is orthogonal, $\|cQ\mathbf{v}\| = c\|\mathbf{v}\|$ for every $\mathbf{v}$, so $r'(\mathbf{z}') = c\,r(\mathbf{z})$ and the unit displacement transforms as $\mathbf{u}'(\mathbf{z}') = U\,\mathbf{u}(\mathbf{z})$. Orthogonality also preserves inner products, $\langle U\mathbf{u}, U\boldsymbol{\theta}\rangle = \langle \mathbf{u}, \boldsymbol{\theta}\rangle$, so the angular coordinate is unchanged. Membership in a region is the conjunction of $|r - d| \le \epsilon_d$ and $\angle(\mathbf{u}, \boldsymbol{\theta}) \le \epsilon_\theta$; the first becomes $|cr - cd| \le c\epsilon_d$, which is the same condition, and the second is unchanged. This proves (i), and (ii) follows because the region contains the same reference cases with the same outcomes, so the count, the different-outcome count, and therefore the empirical-Bernstein width of Eq. (10) are all unchanged.

For (iii), with the kernels K_h and L_κ of Appendix C.2, the kernel weights depend on the data only through $(r_i - d)/h$ and $\langle \mathbf{u}_i, \boldsymbol{\theta}\rangle$. The first is unchanged because both the radius and the bandwidth scale by c, and the second is unchanged by orthogonality, so every weight in Eq. (5) is unchanged and so is the ratio that defines $\widehat{p}$. The directional weights $e^{-t/\rho}$ in Eq. (7) are evaluated at $t' = ct$ with $\rho' = c\rho$, hence also unchanged, so q is transported by U and its normalized entropy, a function of the values of q alone, is invariant. The empirical cumulative mass satisfies $\widehat{F}'_{U\boldsymbol{\theta}}(cd) = \widehat{F}_{\boldsymbol{\theta}}(d)$ by the same argument, so its first crossing of a fixed level α occurs at $c\,\widehat{d}^{\,*}(\boldsymbol{\theta})$. Part (iv) follows since the k-th nearest-neighbor distances scale by c, so $\rho' = c\rho$, the averaging range $[\underline{d}, \rho]$ in Eq. (6) maps to $[c\underline{d}, c\rho]$, and $\bar{A}$ is an average of unchanged values over the image range. □

B.3 Uniform Landscape Estimation

Theorem 2 (Uniform landscape estimation rate)**.** *Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(\mathbf{X}, Y)$. Let $m \geq 2$, let $V = (r, \mathbf{u}) \in [0, \infty) \times \mathbb{S}^{m-1}$, and set $Y' = \mathbb{I}[Y \neq y]$. Define*

$$p_0(\mathbf{x}, \gamma) = \mathbb{E}[Y' \mid V = \gamma], \qquad \gamma = (d, \boldsymbol{\theta}),$$

to be a version of the pointwise landscape. Let $\Gamma = [\underline{d}, \overline{d}] \times \mathbb{S}^{m-1}$, where $0 < \underline{d} \leq \overline{d} < \infty$, and equip the radial–angular space with

$$\mathrm{dist}_\Gamma\big((r, \mathbf{u}), (d, \boldsymbol{\theta})\big) = |r - d| + \angle(\mathbf{u}, \boldsymbol{\theta}).$$

*For a fixed query $\mathbf{x}$, assume the following conditions: **(A1)** $p_0(\mathbf{x}, \cdot)$ is Λ-Lipschitz on $[0, \infty) \times \mathbb{S}^{m-1}$; **(A2)** V has a density f_V with respect to $dr\, d\mathbf{u}$ that is bounded above by $f_{\max}$ everywhere and bounded below by $f_{\min} > 0$ on $[\underline{d} - h_0, \overline{d} + h_0] \times \mathbb{S}^{m-1}$; **(A3)** $K_h(t) = \exp\{-t^2/(2h^2)\}$ and $L_\kappa(\mathbf{u}, \boldsymbol{\theta}) = \exp\{\kappa(\mathbf{u}^\top \boldsymbol{\theta} - 1)\}$, where $0 < h \leq h_0 \leq \min\{1, \underline{d}/2\}$ and $\kappa \geq 1$; and **(A4)** for some fixed $a > 0$, $h^{-1} + \kappa + \omega_{\min}(h, \kappa)^{-1} \leq n^a$, where*

$$\omega_\gamma(r, \mathbf{u}) = K_h(r - d) L_\kappa(\mathbf{u}, \boldsymbol{\theta}), \qquad \Omega_\gamma = \mathbb{E}[\omega_\gamma(V)], \qquad \omega_{\min}(h, \kappa) = \inf_{\gamma \in \Gamma} \Omega_\gamma.$$

There exists a constant $c_m > 0$, depending only on m, such that

$$\omega_{\min}(h, \kappa) \geq c_m f_{\min} h \kappa^{-(m-1)/2}.$$

Moreover, there exist constants $C_1, C_2, C_3 > 0$, depending only on the stated regularity constants and the probe domain, such that for every $\delta \in (0, 1)$, if $n\omega_{\min}(h, \kappa) \geq C_3 \log(n/\delta)$, then, with probability at least $1 - \delta$,

$$\sup_{\gamma \in \Gamma} |\widehat{p}(\mathbf{x}, \gamma) - p_0(\mathbf{x}, \gamma)| \leq C_1(h + \kappa^{-1/2}) + C_2 \sqrt{\frac{\log(n/\delta)}{n\omega_{\min}(h, \kappa)}}.$$

For the regional landscape, the same event also satisfies

$$\sup_{\gamma \in \Gamma} |\widehat{p}(\mathbf{x}, \gamma) - p_{\epsilon_d, \epsilon_\theta}(\mathbf{x}, \gamma)| \leq C_1(h + \kappa^{-1/2}) + C_2 \sqrt{\frac{\log(n/\delta)}{n\omega_{\min}(h, \kappa)}} + \Lambda(\epsilon_d + \epsilon_\theta),$$

whenever the conditioning regions have positive probability. Consequently, $\widehat{p}$ is uniformly consistent in probability for p_0 whenever $h \to 0$, $\kappa \to \infty$, and $nh\kappa^{-(m-1)/2}/\log n \to \infty$. In particular, choosing $h \asymp \kappa^{-1/2} \asymp (\log n/n)^{1/(m+2)}$ yields uniform error $O_{\mathbb{P}}((\log n/n)^{1/(m+2)})$.

Proof. Fix $\gamma = (d, \boldsymbol{\theta})$, abbreviate $\widehat{p}(\gamma) = \widehat{p}(\mathbf{x}, \gamma)$, and define

$$Y_i' = \mathbb{I}[Y_i \neq y], \qquad \widehat{C}_\gamma = \frac{1}{n} \sum_{i=1}^n \omega_\gamma(r_i, \mathbf{u}_i) Y_i', \qquad \widehat{\Omega}_\gamma = \frac{1}{n} \sum_{i=1}^n \omega_\gamma(r_i, \mathbf{u}_i).$$

Let $C_\gamma = \mathbb{E}[\widehat{C}_\gamma]$, $\Omega_\gamma = \mathbb{E}[\widehat{\Omega}_\gamma]$, and $p_{h,\kappa}(\gamma) = C_\gamma/\Omega_\gamma$. Because $\mathbb{E}[Y_i' \mid r_i, \mathbf{u}_i] = p_0(\mathbf{x}, r_i, \mathbf{u}_i)$, the function $p_{h,\kappa}$ is the population kernel average of p_0. Thus,

$$|\widehat{p}(\gamma) - p_0(\mathbf{x}, \gamma)| \leq |\widehat{p}(\gamma) - p_{h,\kappa}(\gamma)| + |p_{h,\kappa}(\gamma) - p_0(\mathbf{x}, \gamma)|. \tag{16}$$

We first bound the effective local mass. For $|r - d| \leq h$, $K_h(r - d) \geq e^{-1/2}$. If $\phi = \angle(\mathbf{u}, \boldsymbol{\theta}) \leq \kappa^{-1/2}$, then $\cos\phi \geq 1 - \phi^2/2$, and hence $L_\kappa(\mathbf{u}, \boldsymbol{\theta}) \geq e^{-1/2}$. The surface measure of this spherical cap is bounded below by $c_m' \kappa^{-(m-1)/2}$. Assumptions (A2)–(A3) therefore imply, uniformly over $\gamma \in \Gamma$,

$$\Omega_\gamma \geq f_{\min} e^{-1} \int_{d-h}^{d+h} dr \int_{\angle(\mathbf{u}, \boldsymbol{\theta}) \leq \kappa^{-1/2}} d\mathbf{u} \geq c_m f_{\min} h \kappa^{-(m-1)/2}.$$

This proves the asserted lower bound on $\omega_{\min}$.

For the bias, (A1) gives

$$|p_{h,\kappa}(\gamma) - p_0(\mathbf{x}, \gamma)| \leq \frac{\Lambda}{\Omega_\gamma} \mathbb{E}[\omega_\gamma(r, \mathbf{u})\{|r - d| + \angle(\mathbf{u}, \boldsymbol{\theta})\}].$$

The radial Gaussian moments and density bounds yield

$$\frac{\mathbb{E}[\omega_\gamma |r-d|]}{\Omega_\gamma} \le C \frac{f_{\max}}{f_{\min}} h.$$

For the angular component, the density bounds and polar coordinates centered at $\boldsymbol{\theta}$ yield, up to a constant depending on $f_{\max}/f_{\min}$, the ratio

$$\frac{\int_0^\pi e^{\kappa(\cos\phi - 1)} \phi \sin^{m-2}\phi \, d\phi}{\int_0^\pi e^{\kappa(\cos\phi-1)} \sin^{m-2}\phi \, d\phi}.$$

The inequalities $\cos\phi - 1 \le -c\phi^2$ and $\sin\phi \le \phi$ show that the numerator is $O(\kappa^{-m/2})$. Restricting the denominator to $0 \le \phi \le \kappa^{-1/2}$ gives a lower bound of order $\kappa^{-(m-1)/2}$. Consequently,

$$\sup_{\gamma\in\Gamma} |p_{h,\kappa}(\gamma) - p_0(\mathbf{x},\gamma)| \le C_1(h + \kappa^{-1/2}). \tag{17}$$

Turning to concentration at a fixed probe, for fixed γ, both ω_γ and $\omega_\gamma Y'$ take values in $[0,1]$, and

$$\mathrm{Var}(\omega_\gamma) \le \mathbb{E}[\omega_\gamma^2] \le \Omega_\gamma, \qquad \mathrm{Var}(\omega_\gamma Y') \le \Omega_\gamma.$$

Bernstein's inequality therefore implies that, for every $u > 0$, with probability at least $1 - 4e^{-u}$,

$$\max\{|\widehat{\Omega}_\gamma - \Omega_\gamma|, |\widehat{C}_\gamma - C_\gamma|\} \le \beta_\gamma(u) := \sqrt{\frac{2\Omega_\gamma u}{n}} + \frac{2u}{3n}.$$

If $n\Omega_\gamma \ge 32u$, then $\beta_\gamma(u) \le \Omega_\gamma/2$ and $\widehat{\Omega}_\gamma \ge \Omega_\gamma/2$. Since $0 \le p_{h,\kappa}(\gamma) \le 1$,

$$\begin{aligned} |\widehat{p}(\gamma) - p_{h,\kappa}(\gamma)| &= \frac{|(\widehat{C}_\gamma - C_\gamma) - p_{h,\kappa}(\gamma)(\widehat{\Omega}_\gamma - \Omega_\gamma)|}{\widehat{\Omega}_\gamma} \\ &\le C\sqrt{\frac{u}{n\Omega_\gamma}} \le C\sqrt{\frac{u}{n\omega_{\min}}}. \end{aligned}$$

To make the bound uniform, note that the kernel weights satisfy

$$\|\omega_\gamma - \omega_{\gamma'}\|_\infty \le h^{-1}|d - d'| + \kappa\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \tag{18}$$

because $\sup_t |K_h'(t)| \le h^{-1}$ and the directional kernel is κ-Lipschitz in its probe direction. Set

$$\tau_n = \frac{\omega_{\min}}{16n(h^{-1} + \kappa)}$$

and let $\{\gamma_j\}_{j=1}^J$ be a τ_n-net of Γ. Compactness of $[\underline{d}, \overline{d}] \times \mathbb{S}^{m-1}$ implies $J \le C_\Gamma \tau_n^{-m}$. Assumption (A4) then gives

$$\log J \le C_{m,a} \log n. \tag{19}$$

Apply the fixed-probe bound at every net point with $u = \log(4J/\delta)$. After increasing C_3 if necessary, the mass condition ensures that $n\Omega_{\gamma_j} \ge 32u$. A union bound then gives simultaneous control over all net points with probability at least $1 - \delta$.

For an arbitrary γ, choose a net point γ_j within τ_n. By Eq. (18), $\|\omega_\gamma - \omega_{\gamma_j}\|_\infty \le \omega_{\min}/(16n)$. On the preceding event, $\widehat{\Omega}_{\gamma_j} \ge \Omega_{\gamma_j}/2 \ge \omega_{\min}/2$, and hence $\widehat{\Omega}_\gamma \ge 7\omega_{\min}/16$. The ratio inequality

$$\left|\frac{a}{b} - \frac{a'}{b'}\right| \le \frac{|a-a'| + |b-b'|}{\min\{b, b'\}}, \qquad 0 \le a \le b, \quad 0 \le a' \le b',$$

shows that replacing γ by γ_j changes either $\widehat{p}$ or $p_{h,\kappa}$ by at most C/n. Combining the net bound, Eq. (19), and the mass condition yields

$$\sup_{\gamma\in\Gamma} |\widehat{p}(\gamma) - p_{h,\kappa}(\gamma)| \le C_2\sqrt{\frac{\log(n/\delta)}{n\omega_{\min}}}. \tag{20}$$

Equations (16), (17), and (20) prove the finite-sample assertion. For a regional target, the tower property and (A1) give

$$|p_{\epsilon_d,\epsilon_\theta}(\mathbf{x},\gamma) - p_0(\mathbf{x},\gamma)| \le \mathbb{E}[|p_0(\mathbf{x},V) - p_0(\mathbf{x},\gamma)| \,|\, \mathbf{X} \in \mathcal{R}(\mathbf{x},d,\boldsymbol{\theta})] \le \Lambda(\epsilon_d + \epsilon_\theta).$$

The triangle inequality proves the asserted regional bound.

Under the stated consistency conditions, both terms converge to zero. With $b_n = (\log n/n)^{1/(m+2)}$, $h \asymp b_n$, and $\kappa \asymp b_n^{-2}$, the effective mass is of order b_n^m, while the bias and stochastic terms are both $O(b_n)$. This establishes the stated rate. □

B.4 Minimax Optimality

Theorem 3 (Minimax lower bound). *Let $\mathcal{P}(\Lambda, f_{\min}, f_{\max})$ denote the nonempty class of all distributions satisfying (A1)–(A2) of Theorem 2, with $\Lambda > 0$. There exist constants $c, c' > 0$, depending only on the regularity parameters and the probe domain, such that, for all sufficiently large n, every estimator $\widetilde{p}_n$ based on n i.i.d. observations satisfies*

$$\sup_{P \in \mathcal{P}(\Lambda, f_{\min}, f_{\max})} \Pr_P\left(\sup_{\gamma \in \Gamma} |\widetilde{p}_n(\gamma) - p_{0,P}(\gamma)| \geq c n^{-1/(m+2)} \right) \geq c'.$$

Together with Theorem 2, this result implies that the kernel estimator with $h \asymp \kappa^{-1/2} \asymp (\log n/n)^{1/(m+2)}$ is minimax rate-optimal up to a logarithmic factor.

Proof. A two-point reduction suffices. Let $V = (r, \mathbf{u})$ denote the radial–angular coordinate. Fix an interior probe $\gamma_0 = (d_0, \boldsymbol{\theta}_0) \in \Gamma$ and a common design density f_0 satisfying (A2).

Let $\chi : [0, \infty) \to [0, 1]$ be 1-Lipschitz, with $\chi(0) = 1$ and $\chi(t) = 0$ for $t \geq 1$. For $b > 0$ and $\varepsilon_n \leq 1/4$, define

$$\eta_0(\gamma) = \frac{1}{2}, \qquad \eta_1(\gamma) = \frac{1}{2} + \varepsilon_n \chi\left(\frac{\mathrm{dist}_\Gamma(\gamma, \gamma_0)}{b} \right).$$

These regression functions take values in $[1/4, 3/4]$, satisfy $\|\eta_1 - \eta_0\|_\infty = \varepsilon_n$, and have Lipschitz constants zero and ε_n/b, respectively. The choice $b = \varepsilon_n/\Lambda$ therefore places both induced distributions in the model class for all sufficiently small ε_n.

For $j \in \{0, 1\}$, let P_j be the joint distribution under which $V \sim f_0$ and $Y' \mid V \sim \mathrm{Bernoulli}(\eta_j(V))$. For $|t| \leq 1/4$,

$$\mathrm{kl}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} + t \right) \leq 4t^2.$$

The perturbation is supported on a radial interval of length $O(b)$ and an angular cap of surface measure $O(b^{m-1})$. Hence

$$\begin{aligned} \mathrm{KL}(P_0^{\otimes n} \| P_1^{\otimes n}) &= n\, \mathbb{E}_{f_0}[\mathrm{kl}(\eta_0(V) \| \eta_1(V))] \\ &\leq C_m f_{\max} n \varepsilon_n^2 b^m = C_m f_{\max} \Lambda^{-m} n \varepsilon_n^{m+2}. \end{aligned}$$

Set $\varepsilon_n = c_0 n^{-1/(m+2)}$, where $c_0 > 0$ is sufficiently small that the preceding display is at most $1/2$. For all sufficiently large n, $\varepsilon_n \leq 1/4$ and the support of the perturbation lies in the interior of the probe domain. Pinsker's inequality then gives $\mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq 1/2$. Consequently, every test φ of P_0 against P_1 satisfies

$$\max_{j \in \{0,1\}} P_j^{\otimes n}(\varphi \neq j) \geq \frac{1 - \mathrm{TV}(P_0^{\otimes n}, P_1^{\otimes n})}{2} \geq \frac{1}{4}.$$

Given an estimator $\widetilde{p}_n$, define the minimum-distance test

$$\varphi \in \arg\min_{j \in \{0,1\}} \|\widetilde{p}_n - \eta_j\|_\infty,$$

using an arbitrary deterministic rule to break ties. Because $\|\eta_1 - \eta_0\|_\infty = \varepsilon_n$, the event $\|\widetilde{p}_n - \eta_j\|_\infty < \varepsilon_n/2$ implies $\varphi = j$. Therefore,

$$\max_{j \in \{0,1\}} P_j^{\otimes n}\left(\|\widetilde{p}_n - \eta_j\|_\infty \geq \frac{\varepsilon_n}{2} \right) \geq \frac{1}{4}.$$

Absorbing $c_0/2$ into c proves the lower bound.

The upper rate in Theorem 2 differs from this lower bound by $(\log n)^{1/(m+2)}$. □

B.5 Learned-Subspace Perturbation

Proposition 2 (Learned-subspace perturbation). *Let P and $\widehat{P}$ be orthogonal projectors onto k-dimensional subspaces, with $\|\widehat{P} - P\|_{\mathrm{op}} \leq \xi < 1$. For every unit vector $\boldsymbol{\theta} \in \mathrm{range}(P)$, define*

$$\widehat{\boldsymbol{\theta}} = \frac{\widehat{P}\boldsymbol{\theta}}{\|\widehat{P}\boldsymbol{\theta}\|_2}.$$

If η is Λ-Lipschitz and $0 \le d \le R$ for $R > 0$, then

$$|\eta(\mathbf{x} + d\widehat{\boldsymbol{\theta}}) - \eta(\mathbf{x} + d\boldsymbol{\theta})| \le \sqrt{2}\Lambda R\xi.$$

Consequently, for $\varepsilon_k \ge 0$, any estimator satisfying $|\widehat{\eta}(\mathbf{x}, d, \widehat{\boldsymbol{\theta}}) - \eta(\mathbf{x} + d\widehat{\boldsymbol{\theta}})| \le \varepsilon_k$ also satisfies

$$|\widehat{\eta}(\mathbf{x}, d, \widehat{\boldsymbol{\theta}}) - \eta(\mathbf{x} + d\boldsymbol{\theta})| \le \varepsilon_k + \sqrt{2}\Lambda R\xi.$$

Proof. Fix a unit vector $\boldsymbol{\theta} \in \mathrm{range}(P)$. Because $P\boldsymbol{\theta} = \boldsymbol{\theta}$,

$$\|(I - \widehat{P})\boldsymbol{\theta}\|_2 = \|(P - \widehat{P})\boldsymbol{\theta}\|_2 \le \xi.$$

Let $a = \|\widehat{P}\boldsymbol{\theta}\|_2$. Orthogonality of $\widehat{P}$ gives

$$a^2 = 1 - \|(I - \widehat{P})\boldsymbol{\theta}\|_2^2 \ge 1 - \xi^2 > 0,$$

so $\widehat{\boldsymbol{\theta}}$ is well defined. Moreover, $\langle \boldsymbol{\theta}, \widehat{P}\boldsymbol{\theta} \rangle = a^2$, and hence

$$\begin{aligned} \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 = 2 - 2\frac{\langle \boldsymbol{\theta}, \widehat{P}\boldsymbol{\theta} \rangle}{a} &= 2(1 - a) \\ &\le 2\big(1 - \sqrt{1 - \xi^2}\big) \le 2\xi^2. \end{aligned}$$

The Lipschitz property of η therefore gives

$$|\eta(\mathbf{x} + d\widehat{\boldsymbol{\theta}}) - \eta(\mathbf{x} + d\boldsymbol{\theta})| \le \Lambda d\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 \le \sqrt{2}\Lambda R\xi.$$

The second assertion follows by applying the triangle inequality to the estimation error at the learned probe and the preceding approximation error. □

B.6 Reach Stability

Theorem 4 (Reach stability)**.** *Fix a query $(\mathbf{x}, y)$, a direction $\boldsymbol{\theta}$, and an evidence level $\alpha \in (0, 1)$, and let $\widehat{d}^*(\mathbf{x}, \boldsymbol{\theta})$ be the plug-in reach obtained by replacing F with a nondecreasing estimate $\widehat{F}$ in the definition of d^* following Eq. (8). Suppose $d^* = d^*(\mathbf{x}, \boldsymbol{\theta}) \in (0, \infty)$ and the following margin condition holds: there exist $\lambda > 0$ and $r_0 \in (0, d^*]$ such that*

$$F(\mathbf{x}, d^* + t, \boldsymbol{\theta}) \ge \alpha + \lambda t, \qquad F(\mathbf{x}, d^* - t, \boldsymbol{\theta}) \le \alpha - \lambda t$$

for all $t \in [0, r_0]$. If

$$\sup_{|d - d^*| \le r_0} |\widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}) - F(\mathbf{x}, d, \boldsymbol{\theta})| \le \varepsilon_F$$

with $\varepsilon_F < \lambda r_0$, then

$$|\widehat{d}^*(\mathbf{x}, \boldsymbol{\theta}) - d^*(\mathbf{x}, \boldsymbol{\theta})| \le \frac{\varepsilon_F}{\lambda}.$$

Thus the reach error is bounded by the local estimation error divided by the crossing margin.

For the empirical estimator $\widehat{F}$ in Eq. (22), let the query $(\mathbf{x}, y)$, cone half-angle ψ, and directions $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$ be fixed independently of an i.i.d. reference sample. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\max_{1 \le t \le T} \sup_{d \ge 0} |\widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}_t) - F(\mathbf{x}, d, \boldsymbol{\theta}_t)| \le \sqrt{\frac{\log(2T/\delta)}{2n}}.$$

Consequently, the reach bound holds simultaneously over every direction $\boldsymbol{\theta}_t$ whose margin condition holds with constants $(\lambda_t, r_{0,t})$ satisfying $\sqrt{\log(2T/\delta)/(2n)} < \lambda_t r_{0,t}$.

Proof. Fix $(\mathbf{x}, y)$ and $\boldsymbol{\theta}$, and write

$$F(d) = F(\mathbf{x}, d, \boldsymbol{\theta}), \qquad \widehat{F}(d) = \widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}).$$

Let $d^* = \inf\{d \ge 0 : F(d) \ge \alpha\}$ and $\widehat{d}^* = \inf\{d \ge 0 : \widehat{F}(d) \ge \alpha\}$. Set $t_\varepsilon = \varepsilon_F/\lambda$. The condition $\varepsilon_F < \lambda r_0$ implies $t_\varepsilon < r_0$.

At the right comparison point, the local uniform bound and margin condition give

$$\widehat{F}(d^* + t_\varepsilon) \ge F(d^* + t_\varepsilon) - \varepsilon_F \ge \alpha + \lambda t_\varepsilon - \varepsilon_F = \alpha.$$

Therefore $\widehat{d}^* \le d^* + t_\varepsilon$.

For the converse, let $0 \le d < d^* - t_\varepsilon$. If $d \in [d^* - r_0, d^* - t_\varepsilon)$, write $d = d^* - t$ with $t \in (t_\varepsilon, r_0]$. The local uniform bound and margin condition imply

$$\widehat{F}(d) \le F(d) + \varepsilon_F \le \alpha - \lambda t + \varepsilon_F < \alpha.$$

If $d < d^* - r_0$, monotonicity of $\widehat{F}$ gives

$$\widehat{F}(d) \le \widehat{F}(d^* - r_0) \le F(d^* - r_0) + \varepsilon_F \le \alpha - \lambda r_0 + \varepsilon_F < \alpha.$$

Thus no nonnegative point below $d^* - t_\varepsilon$ belongs to $\{d \ge 0 : \widehat{F}(d) \ge \alpha\}$, and $\widehat{d}^* \ge d^* - t_\varepsilon$. Combining the two bounds gives

$$|\widehat{d}^* - d^*| \le t_\varepsilon = \frac{\varepsilon_F}{\lambda}.$$

For the empirical assertion, fix $\boldsymbol{\theta}_t$ and define the extended real variable $V_i = r_i$ when $i \in \mathcal{I}_{\mathbf{x}}$, $y_i \ne y$, and $\angle(\mathbf{u}_i, \boldsymbol{\theta}_t) \le \psi$, and $V_i = +\infty$ otherwise. Then $F(\mathbf{x}, d, \boldsymbol{\theta}_t) = \Pr(V_i \le d)$ and $\widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}_t) = n^{-1} \sum_{i=1}^n \mathbb{I}[V_i \le d]$. The Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) gives

$$\Pr\left(\sup_{d \ge 0} |\widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}_t) - F(\mathbf{x}, d, \boldsymbol{\theta}_t)| > u \right) \le 2e^{-2nu^2}.$$

Setting $u = \sqrt{\log(2T/\delta)/(2n)}$ and applying a union bound over the T prespecified directions proves the simultaneous statement. Applying the deterministic result direction by direction completes the proof. □

B.7 Transfer to Smooth Evidence Perspectives

On the probe grid $\mathcal{G}$ of Eq. (3), the continuous perspectives of Section 3.1 are replaced by their grid counterparts. Prevalence at distance d is $A_T(\mathbf{x}, d) = T^{-1} \sum_{t=1}^T p_0(\mathbf{x}, d, \boldsymbol{\theta}_t)$, and $\bar{A}_T(\mathbf{x})$ is obtained by substituting A_T for A in Eq. (6); the inverse prevalence is $\mathrm{IA}_T(\mathbf{x}) = 1/\bar{A}_T(\mathbf{x})$ whenever $\bar{A}_T(\mathbf{x}) > 0$. For the direction perspective, let $M_t = \int_{\underline{d}}^{\overline{d}} e^{-r/\rho} p_0(\mathbf{x}, r, \boldsymbol{\theta}_t)\, dr$ and $q_{T,t} = M_t / \sum_{s=1}^T M_s$, so that $q_T = (q_{T,1}, \ldots, q_{T,T})$ is the grid direction distribution; with $W_\Gamma = \int_{\underline{d}}^{\overline{d}} e^{-r/\rho}\, dr$, the normalized total directional evidence is $Z_T = (TW_\Gamma)^{-1} \sum_{t=1}^T M_t$. For $T \ge 2$, the normalized entropy of q_T is

$$H_T(\mathbf{x}) = -\frac{1}{\log T} \sum_{t=1}^T q_{T,t} \log q_{T,t}, \tag{21}$$

with $0 \log 0 = 0$ and effective direction count T^{H_T}. The main text omits the grid subscript. Appendix C.3 describes the radial discretization.

Proposition 3 (Transfer to smooth evidence perspectives)**.** *Consider the grid-discretized evidence perspectives defined above, computed over $T \ge 2$ probe directions. Let $\Gamma_T = [\underline{d}, \overline{d}] \times \{\boldsymbol{\theta}_t\}_{t=1}^T$, and define the plug-in estimates of A_T, $\bar{A}_T$, q_T, H_T, and IA_T by replacing p_0 with $\widehat{p}$. Suppose*

$$\sup_{(d,\boldsymbol{\theta}) \in \Gamma_T} |\widehat{p}(\mathbf{x}, d, \boldsymbol{\theta}) - p_0(\mathbf{x}, d, \boldsymbol{\theta})| \le \varepsilon_p.$$

Suppose that $\mu \in (0, 1]$ lower-bounds both the grid prevalence $\bar{A}_T(\mathbf{x})$ and the normalized total directional evidence $Z_T = (TW_\Gamma)^{-1} \sum_{t=1}^T M_t$ defined above. Then, whenever $\varepsilon_p \le \mu/4$, for every $d \in [\underline{d}, \overline{d}]$,

$$|\widehat{A}_T(\mathbf{x}, d) - A_T(\mathbf{x}, d)| \le \varepsilon_p, \qquad |\widehat{\bar{A}}_T(\mathbf{x}) - \bar{A}_T(\mathbf{x})| \le \varepsilon_p, \qquad \|\widehat{q}_T - q_T\|_1 \le \frac{4\varepsilon_p}{\mu}.$$

Moreover, let $\tau = 2\varepsilon_p/\mu$ and $h_b(\tau) = -\tau \log \tau - (1 - \tau) \log(1 - \tau)$, with $h_b(0) = h_b(1) = 0$ by continuity. Then

$$|\widehat{H}_T(\mathbf{x}) - H_T(\mathbf{x})| \le \frac{\tau \log(T - 1) + h_b(\tau)}{\log T}, \qquad \frac{|\widehat{\mathrm{IA}}_T(\mathbf{x}) - \mathrm{IA}_T(\mathbf{x})|}{\mathrm{IA}_T(\mathbf{x})} \le \frac{2\varepsilon_p}{\mu}.$$

Proof. Work throughout on the event $\sup_{(d,\boldsymbol{\theta})\in\Gamma_T} |\widehat{p}(\mathbf{x}, d, \boldsymbol{\theta}) - p_0(\mathbf{x}, d, \boldsymbol{\theta})| \le \varepsilon_p$.

For every $d \in [\underline{d}, \overline{d}]$,

$$|\widehat{A}_T(\mathbf{x}, d) - A_T(\mathbf{x}, d)| = \left| \frac{1}{T} \sum_{t=1}^{T} \{\widehat{p}(\mathbf{x}, d, \boldsymbol{\theta}_t) - p_0(\mathbf{x}, d, \boldsymbol{\theta}_t)\} \right| \le \varepsilon_p.$$

Integrating this inequality over $[\underline{d}, \rho]$ and dividing by $\rho - \underline{d}$ gives $|\widehat{\bar{A}}_T(\mathbf{x}) - \bar{A}_T(\mathbf{x})| \le \varepsilon_p$.

For each direction, recall M_t and define its plug-in counterpart,

$$M_t = \int_{\underline{d}}^{\overline{d}} e^{-r/\rho} p_0(\mathbf{x}, r, \boldsymbol{\theta}_t)\, dr, \qquad \widehat{M}_t = \int_{\underline{d}}^{\overline{d}} e^{-r/\rho} \widehat{p}(\mathbf{x}, r, \boldsymbol{\theta}_t)\, dr.$$

Let $Z = \sum_t M_t$ and $\widehat{Z} = \sum_t \widehat{M}_t$. Uniform control of the landscape implies

$$|\widehat{M}_t - M_t| \le \varepsilon_p W_\Gamma, \qquad \|\widehat{M} - M\|_1 \le T\varepsilon_p W_\Gamma, \qquad |\widehat{Z} - Z| \le T\varepsilon_p W_\Gamma.$$

By assumption, $Z \ge TW_\Gamma\mu$, and hence $\widehat{Z} \ge TW_\Gamma(\mu - \varepsilon_p) > 0$. Writing $q_T = M/Z$ and $\widehat{q}_T = \widehat{M}/\widehat{Z}$ gives

$$\|\widehat{q}_T - q_T\|_1 \le \frac{\|\widehat{M} - M\|_1}{\widehat{Z}} + \frac{|\widehat{Z} - Z|}{\widehat{Z}} \le \frac{2\varepsilon_p}{\mu - \varepsilon_p} \le \frac{4\varepsilon_p}{\mu},$$

where the final inequality follows from $\varepsilon_p \le \mu/4$.

Let $v = \mathrm{TV}(\widehat{q}_T, q_T) = \frac{1}{2}\|\widehat{q}_T - q_T\|_1$. Then $v \le \tau := 2\varepsilon_p/\mu \le 1/2$. Because $T \ge 2$, we also have $\tau \le 1 - 1/T$, so the finite-alphabet entropy continuity inequality of Audenaert (2007) yields

$$|\widehat{H}_T - H_T| \le \frac{\tau \log(T-1) + h_b(\tau)}{\log T}.$$

For inverse prevalence, $\bar{A}_T \ge \mu$ and $\widehat{\bar{A}}_T \ge \mu - \varepsilon_p \ge 3\mu/4$. Therefore,

$$\frac{|\widehat{\mathrm{IA}}_T - \mathrm{IA}_T|}{\mathrm{IA}_T} = \frac{|\widehat{\bar{A}}_T - \bar{A}_T|}{\widehat{\bar{A}}_T} \le \frac{\varepsilon_p}{\mu - \varepsilon_p} \le \frac{2\varepsilon_p}{\mu}.$$

□

B.8 Uniformity Across Queries

Corollary 1 (Uniform estimation across queries)**.** *Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(\mathbf{X}, Y)$. Let $\mathcal{X}_0 \subset \mathcal{X}$ be compact and fix the target outcome y. Assume that (A1)–(A4) of Theorem 2 hold uniformly over $\mathbf{x} \in \mathcal{X}_0$, with (A1) strengthened to joint Λ-Lipschitz continuity of $p_0(\mathbf{x}, d, \boldsymbol{\theta})$ in $(\mathbf{x}, d, \boldsymbol{\theta})$. For $\gamma = (d, \boldsymbol{\theta})$ and $\mathbf{z} \ne \mathbf{x}$, let $r_{\mathbf{x}}(\mathbf{z}) = \|\mathbf{z} - \mathbf{x}\|$, $\mathbf{u}_{\mathbf{x}}(\mathbf{z}) = (\mathbf{z} - \mathbf{x})/r_{\mathbf{x}}(\mathbf{z})$, and*

$$\omega_{\mathbf{x},\gamma}(\mathbf{z}) = K_h(r_{\mathbf{x}}(\mathbf{z}) - d) L_\kappa(\mathbf{u}_{\mathbf{x}}(\mathbf{z}), \boldsymbol{\theta}),$$

with $\omega_{\mathbf{x},\gamma}(\mathbf{x}) = 0$. Define

$$\omega_{\min}(h, \kappa) = \inf_{\mathbf{x}\in\mathcal{X}_0,\, \gamma\in\Gamma} \mathbb{E}[\omega_{\mathbf{x},\gamma}(\mathbf{X})].$$

Assume in addition that the query-indexed weight class

$$\mathcal{W}_{h,\kappa} = \{\mathbf{z} \mapsto \omega_{\mathbf{x},\gamma}(\mathbf{z}) : (\mathbf{x}, \gamma) \in \mathcal{X}_0 \times \Gamma\}$$

has polynomial uniform entropy: for constants $v, c_0 > 0$ and a sequence A_n satisfying $\log \mathsf{A}_n \le c_0 \log n$,

$$\log \mathsf{N}(\tau, \mathcal{W}_{h,\kappa}, \|\cdot\|_\infty) \le v \log(\mathsf{A}_n/\tau), \qquad 0 < \tau < 1.$$

There exist constants $C_1, C_2, C_3 > 0$, depending only on the stated regularity and entropy constants, such that for every $\delta \in (0, 1)$, if $n\omega_{\min}(h, \kappa) \ge C_3 \log(n/\delta)$, then, with probability at least $1 - \delta$,

$$\sup_{\mathbf{x}\in\mathcal{X}_0} \sup_{\gamma\in\Gamma} |\widehat{p}(\mathbf{x}, \gamma) - p_0(\mathbf{x}, \gamma)| \le C_1(h + \kappa^{-1/2}) + C_2\sqrt{\frac{\log(n/\delta)}{n\omega_{\min}(h, \kappa)}}.$$

Consequently, for each fixed direction grid, the conclusions of Proposition 3 hold simultaneously over all queries in any subset where $\bar{A}_T$ and the normalized total direction-grid evidence are uniformly bounded below by $\mu > 0$.

Proof. Index the kernel numerator and denominator by $s = (\mathbf{x}, d, \boldsymbol{\theta}) \in \mathcal{X}_0 \times \Gamma$, and write $\omega_s = \omega_{\mathbf{x},\gamma}$ for $\gamma = (d, \boldsymbol{\theta})$. Let $Y' = \mathbb{I}[Y \neq y]$, $\Omega_s = \mathbb{E}[\omega_s(\mathbf{X})]$, and $p_{h,\kappa}(s) = \mathbb{E}[\omega_s(\mathbf{X})Y']/\Omega_s$. By the bias calculation in the proof of Theorem 2, the uniform versions of (A1)–(A3) imply

$$\sup_s |p_{h,\kappa}(s) - p_0(s)| \le C_1(h + \kappa^{-1/2}), \qquad \inf_s \mathbb{E}[\omega_s(\mathbf{X})] \ge \omega_{\min}(h, \kappa).$$

Set $\tau_n = \omega_{\min}/(16n)$ and choose a τ_n-cover of $\mathcal{W}_{h,\kappa}$ in the supremum norm. Its cardinality J satisfies

$$\log J \le v \log(16n\mathsf{A}_n/\omega_{\min}) \le C \log n,$$

where the final inequality follows from the entropy assumption and (A4). Apply the fixed-probe Bernstein argument from Theorem 2 to each representative weight with failure probability δ/J. The lower bound on $n\omega_{\min}$ ensures that all empirical denominators are at least $\omega_{\min}/2$ on the resulting event. A union bound therefore yields

$$\max_{1 \le j \le J} |\widehat{p}(s_j) - p_{h,\kappa}(s_j)| \le C\sqrt{\frac{\log(J/\delta)}{n\omega_{\min}}}$$

with probability at least $1 - \delta$.

For arbitrary s, select a representative s_j whose weight differs by at most τ_n in supremum norm. The empirical numerator and denominator, as well as their expectations, then differ by at most τ_n. The ratio inequality used in the proof of Theorem 2, together with the denominator lower bound, gives

$$|\widehat{p}(s) - \widehat{p}(s_j)| + |p_{h,\kappa}(s) - p_{h,\kappa}(s_j)| \le C/n.$$

Because $\log J = O(\log n)$, the stochastic term absorbs the C/n term. Adding the uniform bias bound proves the landscape assertion.

The lower-evidence condition is deterministic and uniform over the stated query subset. Applying Proposition 3 to every query on the same landscape event proves the simultaneous bounds for the derived evidence perspectives. □

B.9 Regional Empirical-Bernstein Certificate

Theorem 5 (Regional empirical-Bernstein certificate)**.** *Let $(\mathbf{X}_i, Y_i)_{i=1}^n$ be independent and identically distributed observations. Fix a query $(\mathbf{x}, y)$ and any fixed measurable feature-defined region $\mathcal{R}$ with $\Pr(\mathbf{X} \in \mathcal{R}) > 0$, and define*

$$p_{\mathcal{R}} = \Pr(Y \neq y \mid \mathbf{X} \in \mathcal{R}), \qquad N = \sum_{i=1}^n \mathbb{I}[\mathbf{X}_i \in \mathcal{R}], \qquad \widehat{p}_{\mathcal{R}} = \frac{1}{N}\sum_{i=1}^n \mathbb{I}[\mathbf{X}_i \in \mathcal{R}]\mathbb{I}[Y_i \neq y].$$

For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following bound holds whenever $N \ge 2$:

$$|\widehat{p}_{\mathcal{R}} - p_{\mathcal{R}}| \le \sqrt{\frac{2\,\widehat{p}_{\mathcal{R}}(1 - \widehat{p}_{\mathcal{R}})\ln(3/\delta)}{N}} + \frac{3\ln(3/\delta)}{N}.$$

In particular, the certificate width depends only on observed quantities, and $N \ge \frac{9}{2}\varepsilon^{-2}\ln(3/\delta)$ in-region observations suffice to certify $p_{\mathcal{R}}$ within additive error $\varepsilon \in (0, 1]$. For a finite probe grid $\mathcal{G}$ fixed independently of the sample, applying the bound to each region $\mathcal{R}_\gamma$ with failure probability $\delta/|\mathcal{G}|$ and taking a union bound gives, with probability at least $1 - \delta$, $|\widehat{p}_\gamma^{\,\mathrm{bin}} - p_{\epsilon_d,\epsilon_\theta}(\mathbf{x}, \gamma)| \le w_\gamma$ simultaneously for all $\gamma \in \mathcal{G}$ with $N_{\mathcal{D}}(\gamma) \ge 2$, where w_γ is the width in Eq. (10); in particular, $N_{\mathcal{D}}(\gamma) \ge n_\varepsilon^{\mathrm{suff}} = \lceil \frac{9}{2}\varepsilon^{-2}\ln(3|\mathcal{G}|/\delta)\rceil$ implies $w_\gamma \le \varepsilon$.

Proof. Let $I_i = \mathbb{I}[\mathbf{X}_i \in \mathcal{R}]$ and condition on the membership vector $(I_1, \ldots, I_n)$. For any realization with $\mathcal{S} = \{i : I_i = 1\}$ and $|\mathcal{S}| = N \ge 2$, the variables

$$Y_i' = \mathbb{I}[Y_i \neq y], \qquad i \in \mathcal{S},$$

are independent Bernoulli variables with common mean

$$\bar{p} = \Pr(Y \neq y \mid \mathbf{X} \in \mathcal{R}) = p_{\mathcal{R}}.$$

Moreover, $\widehat{p}_{\mathcal{R}} = N^{-1}\sum_{i \in \mathcal{S}} Y_i'$.

Apply Theorem 1 of Audibert et al. (2009) with $x = \ln(3/\delta)$. The empirical variance is

$$\widehat{\sigma}_N^2 = \frac{1}{N}\sum_{i\in\mathcal{S}}(Y_i' - \widehat{p}_{\mathcal{R}})^2 = \widehat{p}_{\mathcal{R}}(1-\widehat{p}_{\mathcal{R}}).$$

Conditional on the membership vector, the theorem gives probability at least $1 - 3e^{-x} = 1 - \delta$ that

$$|\widehat{p}_{\mathcal{R}} - \bar{p}| \le \sqrt{\frac{2\widehat{p}_{\mathcal{R}}(1-\widehat{p}_{\mathcal{R}})\ln(3/\delta)}{N}} + \frac{3\ln(3/\delta)}{N}.$$

Because this conditional statement holds for every membership realization with $N \ge 2$, integrating over the membership vector proves the asserted unconditional event.

For the sufficient count, let $\ell = \ln(3/\delta)$ and assume $N \ge \frac{9}{2}\ell/\varepsilon^2$. Since $\widehat{p}_{\mathcal{R}}(1-\widehat{p}_{\mathcal{R}}) \le 1/4$ and $\varepsilon \le 1$,

$$\sqrt{\frac{2\widehat{p}_{\mathcal{R}}(1-\widehat{p}_{\mathcal{R}})\ell}{N}} \le \sqrt{\frac{\ell}{2N}} \le \frac{\varepsilon}{3}, \qquad \frac{3\ell}{N} \le \frac{2\varepsilon^2}{3} \le \frac{2\varepsilon}{3}.$$

The two terms sum to at most ε, proving the final assertion. The constant is the smallest of this form: writing $N = c\varepsilon^{-2}\ell$, the requirement becomes $(2c)^{-1/2} + 3\varepsilon/c \le 1$, whose worst case $\varepsilon = 1$ holds with equality at $c = 9/2$ and fails for every smaller c. □

B.10 Certified Regional Evidence Perspectives

Proposition 4 (Certified regional evidence perspectives)**.** *Let $\mathcal{J}$ be the finite set of regional probes contributing to an evidence perspective, and suppose that, on a common event, $|\widehat{p}_\gamma^{\mathrm{bin}} - p_\gamma| \le w_\gamma$ for every $\gamma \in \mathcal{J}$. For nonnegative prevalence weights a_γ summing to one,*

$$\left|\sum_{\gamma\in\mathcal{J}} a_\gamma \widehat{p}_\gamma^{\mathrm{bin}} - \sum_{\gamma\in\mathcal{J}} a_\gamma p_\gamma\right| \le \sum_{\gamma\in\mathcal{J}} a_\gamma w_\gamma.$$

For directional weights $b_{\gamma t} \ge 0$, where $t \in \{1,\dots,T\}$ and $T \ge 2$, define $\widehat{M}_t = \sum_{\gamma\in\mathcal{J}} b_{\gamma t}\widehat{p}_\gamma^{\mathrm{bin}}$, $M_t = \sum_{\gamma\in\mathcal{J}} b_{\gamma t}p_\gamma$, $e_t = \sum_{\gamma\in\mathcal{J}} b_{\gamma t}w_\gamma$, $\widehat{Z} = \sum_t \widehat{M}_t$, $Z = \sum_t M_t$, and $E = \sum_t e_t$. If $\widehat{Z} > E$, then $Z > 0$, and the normalized vectors $\widehat{q}_t = \widehat{M}_t/\widehat{Z}$ and $q_t = M_t/Z$ satisfy $\mathrm{TV}(\widehat{q}, q) \le E/(\widehat{Z}-E)$. The corresponding normalized entropy error is at most $\{\tau\log(T-1) + h_b(\tau)\}/\log T$, where $\tau = \min\{E/(\widehat{Z}-E), 1-1/T\}$ and $h_b(\tau) = -\tau\log\tau - (1-\tau)\log(1-\tau)$, with $h_b(0) = h_b(1) = 0$ by continuity. Thus, $\widehat{Z} > E$ also certifies positive total directional evidence.

Proof. On the simultaneous event in the proposition,

$$\left|\sum_{\gamma\in\mathcal{J}} a_\gamma(\widehat{p}_\gamma^{\mathrm{bin}} - p_\gamma)\right| \le \sum_{\gamma\in\mathcal{J}} a_\gamma w_\gamma,$$

which proves the prevalence bound.

For each direction t, nonnegativity of the directional weights gives

$$|\widehat{M}_t - M_t| \le \sum_{\gamma\in\mathcal{J}} b_{\gamma t}w_\gamma = e_t.$$

Therefore, $\|\widehat{M} - M\|_1 \le E$ and $|\widehat{Z} - Z| \le E$, where $Z = \sum_t M_t$. The condition $\widehat{Z} > E$ implies $Z \ge \widehat{Z} - E > 0$, so both normalized vectors are well defined. Using $\widehat{q} = \widehat{M}/\widehat{Z}$ and $q = M/Z$,

$$\begin{aligned}\|\widehat{q} - q\|_1 &\le \left\|\frac{\widehat{M}-M}{Z}\right\|_1 + \left\|\widehat{M}\left(\frac{1}{\widehat{Z}} - \frac{1}{Z}\right)\right\|_1 \\ &= \frac{\|\widehat{M}-M\|_1 + |\widehat{Z}-Z|}{Z} \le \frac{2E}{\widehat{Z}-E}.\end{aligned}$$

Dividing by two proves $\mathrm{TV}(\widehat{q}, q) \le E/(\widehat{Z}-E)$.

If $E/(\widehat{Z}-E) \le 1-1/T$, the entropy bound follows directly from Audenaert (2007). Otherwise, setting $\tau = 1-1/T$ gives $\tau\log(T-1) + h_b(\tau) = \log T$, so the normalized bound equals one. Finally, $Z \ge \widehat{Z} - E > 0$ establishes the certified positivity of the total directional evidence. □

B.11 Certified Adaptive Stopping

Theorem 6 (Certified adaptive stopping)**.** *Under the i.i.d. reference-sample setting of Theorem 5, fix a query $(\mathbf{x}, y)$ and a radius schedule $R_1 < \cdots < R_{n_R}$ independently of that sample, together with a requested accuracy $\varepsilon \in (0, 1]$ and a failure probability $\delta \in (0, 1)$. Let $\bar{A}_{\mathrm{ball}}(\mathbf{x}; R) = \Pr(Y \neq y \mid \|\mathbf{X} - \mathbf{x}\| \leq R)$ denote the in-ball different-outcome proportion used by the stopping criterion, and let N_s be the number of reference cases within radius R_s. For each s, let $\widehat{A}_{\mathrm{ball},s}$ be the corresponding empirical proportion. Suppose Algorithm 2 stops at the first index $\widehat{s}$ whose empirical-Bernstein width with failure probability δ/n_R is at most ε, or sets $\widehat{s} = n_R$ if no earlier radius satisfies the criterion. Then:* (i) (validity) *with probability at least $1 - \delta$, whenever the width criterion is satisfied at $\widehat{s}$,*

$$\left|\widehat{A}_{\mathrm{ball},\widehat{s}} - \bar{A}_{\mathrm{ball}}(\mathbf{x}; R_{\widehat{s}})\right| \leq \varepsilon;$$

(ii) (adaptivity) *if an index satisfying the following condition exists, then the algorithm deterministically stops no later than the first such index s°:*

$$N_{s^\circ} \geq \tfrac{9}{2}\varepsilon^{-2} \ln(3n_R/\delta).$$

Consequently, it uses no reference case outside radius R_{s°.

Proof. For each $s \in \{1, \ldots, n_R\}$, the ball $\mathcal{B}_s = \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}\|_2 \leq R_s\}$ is fixed independently of the reference outcomes. Applying Theorem 5 to $\mathcal{B}_s$ with failure probability δ/n_R gives

$$|\widehat{A}_{\mathrm{ball},s} - \bar{A}_{\mathrm{ball}}(\mathbf{x}; R_s)| \leq w_s$$

with probability at least $1 - \delta/n_R$, where

$$w_s = \sqrt{\frac{2\widehat{A}_{\mathrm{ball},s}(1 - \widehat{A}_{\mathrm{ball},s}) \ln(3n_R/\delta)}{N_s}} + \frac{3 \ln(3n_R/\delta)}{N_s}$$

for $N_s \geq 2$. A union bound makes these inequalities simultaneous over all scheduled radii with probability at least $1 - \delta$. On this event, if the width criterion is satisfied at the selected index $\widehat{s}$, then

$$|\widehat{A}_{\mathrm{ball},\widehat{s}} - \bar{A}_{\mathrm{ball}}(\mathbf{x}; R_{\widehat{s}})| \leq w_{\widehat{s}} \leq \varepsilon.$$

This proves validity at the data-dependent stopping index.

By the sufficient count condition in Theorem 5, $w_s \leq \varepsilon$ whenever

$$N_s \geq \tfrac{9}{2}\varepsilon^{-2} \ln(3n_R/\delta).$$

The counts N_s are nondecreasing because the balls are nested. Hence, if s° is the first index satisfying this inequality, the procedure stops at or before s°. It follows that $\widehat{s} \leq s^\circ$ and that no reference case outside $\mathcal{B}_{s^\circ}$ is used. □

B.12 Anytime Finite-Reference Inference

Theorem 7 (Anytime finite-reference inference)**.** *Fix a query $(\mathbf{x}, y)$, let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be independent reference cases, and fix $\delta \in (0, 1)$. Conditional on $\mathbf{X}_{1:n}$, order the observations by a permutation measurable with respect to $\mathbf{X}_{1:n}$. In this order, let $Y_i' = \mathbb{I}[Y_i \neq y]$ and*

$$\bar{\nu}_t = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}[Y_i' \mid \mathbf{X}_i].$$

Let $\widehat{Y}_i \in [0, 1]$ be predictable with respect to the revealed-label filtration, and let $u_{\delta/2}$ be a sub-exponential empirical-Bernstein uniform boundary calibrated for $[0, 1]$-bounded increments as defined by Howard et al. (2021). Then

$$I_t = \left[\frac{1}{t}\sum_{i=1}^{t} Y_i' \;\pm\; \frac{u_{\delta/2}\left(\sum_{i=1}^{t}(Y_i' - \widehat{Y}_i)^2\right)}{t}\right] \cap [0, 1]$$

satisfies

$$\Pr(\bar{\nu}_t \in I_t \text{ for every } 1 \leq t \leq n \mid \mathbf{X}_{1:n}) \geq 1 - \delta.$$

Consequently, I_ς has conditional coverage at least $1 - \delta$ for every stopping time $\varsigma \leq n$ adapted to the revealed-label filtration.

Proof. Condition throughout on $\mathbf{X}_{1:n}$ and let $\mathcal{H}_t = \sigma(Y'_1, \dots, Y'_t)$. Because the ordering is measurable with respect to the conditioned features, conditional independence of the labels implies

$$\nu_i := \mathbb{E}[Y'_i \mid \mathcal{H}_{i-1}, \mathbf{X}_{1:n}] = \mathbb{E}[Y'_i \mid \mathbf{X}_i].$$

Thus $\Delta_i = Y'_i - \nu_i$ is a bounded martingale-difference sequence, and $t\bar{\nu}_t = \sum_{i=1}^t m_i$.

Applying the empirical-Bernstein confidence-sequence theorem of Howard et al. (2021) with the predictable forecasts $\widehat{Y}_i$ gives

$$\Pr\left(\exists t \geq 1 : \sum_{i=1}^t \Delta_i > u_{\delta/2}\left(\sum_{i=1}^t (Y'_i - \widehat{Y}_i)^2\right) \,\middle|\, \mathbf{X}_{1:n}\right) \leq \frac{\delta}{2}.$$

Applying the corresponding lower-tail inequality and taking a union bound yields, with conditional probability at least $1 - \delta$, the simultaneous bounds

$$\left|\sum_{i=1}^t (Y'_i - \nu_i)\right| \leq u_{\delta/2}\left(\sum_{i=1}^t (Y'_i - \widehat{Y}_i)^2\right) \quad \text{for every } 1 \leq t \leq n.$$

Division by t is equivalent to $\bar{\nu}_t \in I_t$.

The event holds simultaneously over $1 \leq t \leq n$ and therefore at every stopping time $\varsigma \leq n$ adapted to $(\mathcal{H}_t)$. At time t, the estimand is the finite-reference conditional mean $t^{-1}\sum_{i \leq t} \nu_i$. □

B.13 Count-Deficit Acquisition

Proposition 5 (Count-deficit acquisition)**.** *Let $\mathcal{G}$ be a finite probe grid, let $n_\varepsilon \in \mathbb{N}$, and let $D_\gamma = [n_\varepsilon - N_{\mathcal{D}}(\gamma)]_+$ be the support deficit of probe γ. For $S \subseteq \mathcal{U}$, let $f(S) = \sum_{\gamma \in \mathcal{G}} \min\{N_S(\gamma), D_\gamma\}$, the objective of Eq.* (11) *with the sufficient reporting count replaced by n_ε. Then $f(\varnothing) = 0$, and f is monotone and submodular. Hence the sequential greedy set S_B satisfies*

$$f(S_B) \geq (1 - 1/e) \max_{|S| \leq B} f(S).$$

Let $D = \sum_{\gamma \in \mathcal{G}} D_\gamma$. If $D \geq 1$ and full deficit coverage is feasible, so that $f(\mathcal{U}) = D$, then sequential greedy solves $\min\{|S| : f(S) = D\}$ within a factor $(1 + \ln D)$ of the optimum.

Proof. For each $\gamma \in \mathcal{G}$, define

$$N_S(\gamma) = \sum_{\mathbf{z} \in S} \mathbb{I}[\mathbf{z} \in \mathcal{R}_\gamma], \qquad f_\gamma(S) = \min\{N_S(\gamma), D_\gamma\}.$$

If $S \subseteq S' \subseteq \mathcal{U}$ and $\mathbf{z} \notin S'$, then

$$\begin{aligned} f_\gamma(S \cup \{\mathbf{z}\}) - f_\gamma(S) &= \mathbb{I}[\mathbf{z} \in \mathcal{R}_\gamma]\mathbb{I}[N_S(\gamma) < D_\gamma] \\ &\geq \mathbb{I}[\mathbf{z} \in \mathcal{R}_\gamma]\mathbb{I}[N_{S'}(\gamma) < D_\gamma] \\ &= f_\gamma(S' \cup \{\mathbf{z}\}) - f_\gamma(S'). \end{aligned}$$

The marginal gains are nonnegative, so f_γ is monotone and submodular. Summation over γ preserves both properties, proving the first assertion. The cardinality-constrained guarantee follows from the classical greedy theorem for normalized monotone submodular maximization (Nemhauser et al., 1978).

The saturation in f is essential for the approximation guarantee. For comparison, up to the constant factor $1/|\mathcal{G}|$, the marginal contribution of probe γ to the thresholded count-qualified fraction in Eq. (24) is

$$\mathbb{I}[\mathbf{z} \in \mathcal{R}_\gamma]\mathbb{I}[N_S(\gamma) = D_\gamma - 1],$$

which can increase as S approaches the threshold. The greedy guarantee therefore applies to the deficit-reduction objective f.

For the covering statement, assume $f(\mathcal{U}) = D$ and let $S^\star$ be an optimal cover with $k = |S^\star|$. At an intermediate greedy set S, submodularity gives

$$D - f(S) \leq \sum_{\mathbf{z} \in S^\star \setminus S} \{f(S \cup \{\mathbf{z}\}) - f(S)\}.$$

Hence some element of $S^\star \setminus S$ has marginal gain at least $(D - f(S))/k$, and the greedy marginal gain is no smaller. Let $\bar{D}_j = D - f(S_j)$ be the residual before a greedy selection, and let g_j be the marginal gain of that selection. Then $g_j \geq \bar{D}_j/k$. Charge the unit cost of this selection equally to the g_j newly covered deficit units. Each such unit receives charge at most $k/\bar{D}_j$, and

$$1 \leq k \sum_{r=\bar{D}_j - g_j + 1}^{\bar{D}_j} \frac{1}{r}.$$

The residual intervals in this sum are disjoint across iterations. Summing the charges until the residual reaches zero gives

$$|S_{\text{greedy}}| \leq k \sum_{r=1}^{D} \frac{1}{r} \leq k(1 + \ln D).$$

This is the standard integer-valued submodular-cover bound (Wolsey, 1982). □

C Implementation and Algorithms

This appendix is self-contained: it restates the quantities SACA computes before describing how each is computed. Appendix A tabulates every symbol.

C.1 Setup and Notation Recalled

A query is a pair $(\mathbf{x}, y)$ with $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, and the reference sample is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Outcomes are either observed labels or the predictions of a fixed rule, under the same convention for the query and the sample. A case $\mathbf{z} \neq \mathbf{x}$ is placed relative to the query by the displacement coordinates of Eq. (1), its distance $r(\mathbf{z}) = \|\mathbf{z} - \mathbf{x}\|_2$ and its direction $\mathbf{u}(\mathbf{z}) = (\mathbf{z} - \mathbf{x})/r(\mathbf{z})$ on the unit sphere $\mathbb{S}^{m-1}$. A probe $\gamma = (d, \boldsymbol{\theta})$ names one distance and one direction; its region $\mathcal{R}_\gamma$ of Eq. (2) collects the cases within ϵ_d of d in distance and ϵ_θ of $\boldsymbol{\theta}$ in angle, and $\mathcal{G}$ of Eq. (3) is the grid of n_d radii by T directions on which everything below is evaluated. Throughout, ρ is the local scale, ψ the cone half-angle used for reach, ε the requested accuracy, and δ the failure probability.

Four quantities are computed on that grid. The pointwise landscape $p_0(d, \boldsymbol{\theta})$ of Eq. (4) is the probability that an outcome differs at distance d in direction $\boldsymbol{\theta}$, estimated by the kernel ratio $\widehat{p}$ of Eq. (5). Prevalence $\bar{A}$ of Eq. (6) averages p_0 uniformly over directions and then over radii up to ρ; the direction distribution $q(\boldsymbol{\theta})$ of Eq. (7) weights each direction by its exponentially discounted radial evidence; and the reach $d^*(\boldsymbol{\theta})$ inverts the cumulative cone mass $F_{\boldsymbol{\theta}}$ of Eq. (8) at the evidence level α. Reporting uses the regional proportion $\widehat{p}_\gamma^{\text{bin}}$ of Eq. (9) together with the empirical-Bernstein width w_γ of Eq. (10), which depends only on $|\mathcal{R}_\gamma|$, the observed proportion, $|\mathcal{G}|$, and δ. Acquisition maximizes the count-deficit objective f of Eq. (11), whose per-probe deficit is the shortfall of $|\mathcal{R}_\gamma|$ below the count n_ε that certifies width ε. Where a count must be split by its source, we follow Appendix B and write $N_{\mathcal{D}}(\gamma)$ for the reference cases in $\mathcal{R}_\gamma$ and $N_S(\gamma)$ for those contributed by an acquired set S, so that $|\mathcal{R}_\gamma| = N_{\mathcal{D}}(\gamma)$ before any acquisition.

Two execution modes recur below. Profiling evaluates the full probe grid, while the separate adaptive-radius routine stops at the first certified in-ball estimate (Theorem 6).

C.2 Representation, Kernels, and Probe Grid

Feature geometry. We fit min–max scaling to each training split and apply it to the test split; categorical features retain the benchmark encoding. Euclidean geometry defines the radius and direction in Eq. (1). Domain-specific inner-product embeddings and the feasibility restrictions in Appendix C.7 can encode alternative notions of similarity and admissible displacement.

Smoothing. The kernels in Eq. (5) are $K(t) = K_h(t) = \exp(-t^2/2h^2)$ and $L(\mathbf{u}, \boldsymbol{\theta}) = L_\kappa(\mathbf{u}, \boldsymbol{\theta}) = \exp\{\kappa(\mathbf{u}^\top \boldsymbol{\theta} - 1)\}$. Both are bounded by one; their smoothing scales are h radially and $\kappa^{-1/2}$ angularly. Experiment-specific settings appear in Appendix D.

Directional probing. SACA evaluates T fixed directions and n_d radial locations. In two dimensions, the directions are equally spaced; in higher dimensions, they are sampled once from the unit sphere.

Angular neighborhoods become sparse with dimension, as reflected by the $\kappa^{-(m-1)/2}$ effective-mass term in Theorem 2. When $m > 16$, we project displacements onto a top-10 PCA subspace. In the reported runs, the basis is fitted once to the 200 nearest different-outcome displacements of the first 20 prespecified evaluation queries, or all queries when fewer are available, using the initial labeled reference sample; it is then held fixed across the run. Directions are evaluated in this subspace, whereas radial distances remain in the full normalized space.

For nonzero projected displacements and fixed $\widehat{P}$, this hybrid representation targets the conditional mean given $V_{\mathrm{hyb}} = (\|\mathbf{X}-\mathbf{x}\|_2, \widehat{P}(\mathbf{X}-\mathbf{x})/\|\widehat{P}(\mathbf{X}-\mathbf{x})\|_2)$; zero projections form a null set under the continuous-density conditions used by the analysis. The uniform regression analysis applies under the corresponding density and smoothness conditions, while Proposition 2 bounds the posterior perturbation for a fixed target direction mapped into the learned subspace. The reported high-dimensional profiles use this fixed basis descriptively; fitting it on a split independent of the certification sample yields formal regional certificates.

C.3 Computing the Evidence Perspectives

The grid perspectives of Appendix B.7 are computed by averaging $\widehat{p}$ over directions at each radius, then over radial locations within ρ for scalar prevalence. Radial quadrature weights proportional to $e^{-d/\rho}$ are normalized over radial locations; the resulting directional evidence values are normalized across directions before computing entropy. These nonnegative sums admit the transfer bounds of Propositions 3 and 4 for their respective targets.

Empirical reach. With $\inf \varnothing = +\infty$ for both the population and the empirical quantities, the cumulative mass in Eq. (8) is estimated by

$$\widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in \mathcal{I}_{\mathbf{x}}} \mathbb{I}[y_i \neq y,\ r_i \leq d,\ \angle(\mathbf{u}_i, \boldsymbol{\theta}) \leq \psi]\,, \tag{22}$$

and the empirical reach is $\widehat{d}^*(\mathbf{x}, \boldsymbol{\theta}) = \inf\{d \geq 0 : \widehat{F}(\mathbf{x}, d, \boldsymbol{\theta}) \geq \alpha\}$.

The cone half-angle ψ is specified independently of the regional resolution ϵ_θ. On an ordered radial grid, SACA returns the first grid point at which $\widehat{F}$ reaches α. Under the conditions of Theorem 4, if the grid brackets the finite empirical crossing and has maximum spacing Δ_d, discretization adds at most Δ_d to the reach error, giving a total bound of $\varepsilon_F/\lambda + \Delta_d$.

Reach with few cases. Reach is defined through the probability mass $F_{\boldsymbol{\theta}}$ rather than a case count, but its empirical version is a count: $\widehat{F}$ increases by $1/n$ per different-outcome case in the cone, so $\widehat{d}^*$ is finite only if at least $\lceil n\alpha \rceil$ such cases lie within $\overline{d}$. When fewer cases are available, SACA reports the level as unattained ($\widehat{d}^* = \infty$) instead of extrapolating from the cases at hand, and the number of contributing cases $n\widehat{F}(\mathbf{x}, \overline{d}, \boldsymbol{\theta})$ indicates how far the sample falls short of the level. This is the intended behavior: the perspective asks where a prescribed amount of evidence has accumulated, and a sample too small to contain that amount cannot locate it.

Theorem 4 governs the two possible errors. On its high-probability event, every finite empirical reach lies within ε_F/λ of the population reach, and an unattained empirical reach implies $F_{\boldsymbol{\theta}}(\overline{d}) < \alpha + \varepsilon_F$, so a reported ∞ can only occur when the population mass within $\overline{d}$ exceeds α by less than $\varepsilon_F = \sqrt{\log(2T/\delta)/(2n)}$.

Consequently, unattained reach reflects scarce evidence rather than sampling noise once α exceeds ε_F by a margin, which requires $n\alpha^2$ to be large relative to $\log(2T/\delta)$. For smaller samples, an unattained reach should be read together with the contributing count $n\widehat{F}(\mathbf{x}, \overline{d}, \boldsymbol{\theta})$, which Eq. (22) already provides.

C.4 Indexing and Certified Adaptive Estimation

Two costs dominate a query: locating the reference cases that lie near it, and evaluating the probe grid over those cases. The implementation separates them. An offline phase runs once per reference sample and is reused by every query; an online phase runs per query and stops as soon as the

requested accuracy is reached, so a query in a dense region pays for far fewer cases than one in a sparse region.

Algorithm 1 builds a reusable k-d tree and stores the local scale and, when $m > 16$, the PCA basis fitted to the prespecified basis-query set $\mathcal{Q}_{\text{basis}}$. The tree is built on the scaled features, so a query retrieves its neighborhood by a radius query instead of scanning the sample. Both fitted objects, the local scale ρ and the basis $\widehat{P}$, are computed once from the reference sample before any query is evaluated and then frozen for the run.

Algorithm 2 implements Theorem 6 using failure probability δ/n_R at each scheduled radius, where the schedule is a fixed increasing sequence of n_R radii chosen before any outcome is read. Spending δ/n_R at each one makes the widths valid simultaneously across the schedule, which is what permits stopping at the first radius whose width is at most ε: because every radius is covered on one event, the stopping radius may be chosen after seeing the widths. If no radius qualifies, the final estimate is returned with the certification flag set to false rather than being reported.

For an open-ended neighbor stream, the implementation also provides a time-uniform Hoeffding width using $\delta_t = 6\delta/(\pi^2 t^2)$. The empirical-Bernstein sequence in Theorem 7 provides a sharper alternative. Both sequences target the running finite-reference conditional mean, whereas the scheduled-radius procedure targets the population in-ball quantity.

C.5 Support-Guided Acquisition

Count thresholds. We use either the sufficient threshold $n_\varepsilon^{\text{suff}}$ of Theorem 5 or the smaller operational threshold $n_\varepsilon^{\text{op}}$, obtained by numerically inverting Eq. (10) at $\widehat{p}^{\text{bin}} = 1/2$. Both are outcome-independent and supply a count condition for certification under a sampling design valid for the regional estimand.

Objective conventions and selection step. For multiple prespecified queries, γ additionally indexes the query and $\mathcal{G}$ is the union of their probe grids. The marginal gain of a candidate $\mathbf{z} \in \mathcal{U} \setminus S$ is

$$\Delta(\mathbf{z} \mid S) = f(S \cup \{\mathbf{z}\}) - f(S) = \big|\{\gamma \in \mathcal{G} : \mathbf{z} \in \mathcal{R}_\gamma,\ N_{\mathcal{D}}(\gamma) + N_S(\gamma) < n_\varepsilon\}\big|, \tag{23}$$

which sequential selection maximizes at every step.

Count-based coverage. For the count-threshold protocol, the fraction of probes whose support reaches the threshold after acquiring S is

$$C(S) = \frac{1}{|\mathcal{G}|} \sum_{\gamma \in \mathcal{G}} \mathbb{I}[N_{\mathcal{D}}(\gamma) + N_S(\gamma) \geq n_\varepsilon]\,. \tag{24}$$

This evaluation metric need not be submodular (Appendix B.13). The experiments also report width-qualified coverage, defined in Appendix D.4.

Candidate–probe memberships depend only on features, so they are precomputed once and stored as sparse incidence lists: for each candidate, the probes whose region contains it, and for each probe, the candidates it contains. A selection therefore updates only the probes incident to the chosen candidate, and only the candidates sharing one of those probes can change gain. The acquired outcome is retained and reused by the subsequent regional estimates.

Lazy and batched selection. All sequential runs use Minoux's lazy evaluation: submodularity makes stale gains upper bounds, so a priority queue refreshes its top candidate until its updated gain dominates the next bound. This reproduces the exact sequential rule and its $(1-1/e)$ guarantee while evaluating two orders of magnitude fewer gains. The batched variant instead selects the b highest-gain candidates before refreshing any gain, trading exactness for $\lceil B/b \rceil$ refresh rounds in place of B, with $b = 25$ in all reported runs; it can overshoot a deficit when several selected candidates serve the same probe, so Proposition 5 covers the sequential variant and the batched one is compared against it empirically. The width-targeting EWR comparison is specified in Appendix D.4.

Label-aware early stopping. The objective in Eq. (11) uses only case counts, so it treats a region as under-supported until its deficit is filled, regardless of the outcomes revealed along the way. The sufficient count is sufficient rather than necessary: the width in Eq. (10) adapts to the observed variance, so a region whose acquired outcomes mostly agree can satisfy $w_\gamma \leq \varepsilon$ with far fewer

Algorithm 1 SACA offline indexing

Require: Reference sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$; fixed basis-query set $\mathcal{Q}_{\text{basis}}$
Build spatial index $\mathcal{T}$ over $\{\mathbf{x}_i\}_{i=1}^{n}$
Estimate local scale ρ as the median 10-NN distance
if $m > 16$ **then**
  Fit a top-10 PCA basis $\widehat{P}$ to pooled different-outcome displacements from $\mathcal{Q}_{\text{basis}}$
else
  $\widehat{P} \leftarrow \varnothing$
return $(\mathcal{T}, \rho, \widehat{P})$

Algorithm 2 Certified adaptive-radius in-ball estimate

Require: Query $(\mathbf{x}, y)$; index $\mathcal{T}$; radii $R_1 < \cdots < R_{n_R}$; accuracy ε; failure probability δ
certified $\leftarrow$ **false**; $\widehat{s} \leftarrow n_R$
for $s = 1, \ldots, n_R$ **do**
  $\mathcal{N}_s \leftarrow \textsc{RangeSearch}(\mathcal{T}, \mathbf{x}, R_s)$
  $N_s \leftarrow |\mathcal{N}_s|$
  if $N_s = 0$ **then**
    $\widehat{A}_{\text{ball},s} \leftarrow \textsc{Undefined}$; $w_s \leftarrow +\infty$
  else
    $\widehat{A}_{\text{ball},s} \leftarrow N_s^{-1} \sum_{\mathbf{x}_i \in \mathcal{N}_s} \mathbb{I}[y_i \neq y]$
    if $N_s < 2$ **then**
      $w_s \leftarrow +\infty$
    else
      $w_s \leftarrow \textsc{EmpiricalBernsteinWidth}(\widehat{A}_{\text{ball},s}, N_s, \delta/n_R)$
  if $w_s \leq \varepsilon$ **then**
    $\widehat{s} \leftarrow s$; *certified* $\leftarrow$ **true**
    break
return $(\widehat{A}_{\text{ball},\widehat{s}}, R_{\widehat{s}}, w_{\widehat{s}}, \textit{certified})$

cases than $n_\varepsilon^{\text{suff}}$. For $\widehat{p}_\gamma^{\text{bin}} = 0$, the width is $3\ln(3|\mathcal{G}|/\delta)/|\mathcal{R}_\gamma|$, which needs only a $2\varepsilon/3$ fraction of the cases that the worst case at $\widehat{p}_\gamma^{\text{bin}} = 1/2$ needs.

A label-aware variant exploits this. After each acquired outcome, it recomputes w_γ for the probes incident to the new case, sets the deficit of every probe that has become reportable to zero, and continues greedy selection on the remaining deficits, stopping when no deficit remains or the budget is exhausted. Deficits only decrease under this update, so the variant never spends more than the count-only rule on any probe and reallocates the released budget to regions that are still under-supported; the incidence lists above make the update as cheap as a count update.

We keep the count-only rule as the default for two reasons. First, its selection is outcome-independent, so the acquired cases in each region remain an outcome-blind sample and Theorem 5 applies verbatim; under label-aware selection, the order in which outcomes are revealed depends on earlier outcomes, and validity must instead come from a time-uniform sequence such as Theorem 7, whose widths are wider by a constant factor and whose target is the running conditional mean of the revealed cases. Second, Proposition 5 bounds greedy selection against a fixed deficit vector; with deficits updated after every acquisition, the $(1 - 1/e)$ guarantee holds stage by stage relative to the current deficits, but we do not claim a global approximation guarantee for the adaptive variant. Establishing that guarantee and evaluating the variant empirically are natural extensions of the present acquisition rule.

Algorithm 3 Support-guided acquisition

Require: Probe grid $\mathcal{G}$; pool $\mathcal{U}$; deficits $\{D_\gamma\}_{\gamma\in\mathcal{G}}$; budget $0 \le B \le |\mathcal{U}|$
Precompute $\mathcal{G}(\mathbf{z}) = \{\gamma \in \mathcal{G} : \mathbf{z} \in \mathcal{R}_\gamma\}$ for every $\mathbf{z} \in \mathcal{U}$
$S \leftarrow \emptyset$; $N_\gamma(S) \leftarrow 0$ for every $\gamma \in \mathcal{G}$
for $j = 1, \dots, B$ **do**
  $\mathbf{z}^* \leftarrow \arg\max_{\mathbf{z}\in\mathcal{U}\setminus S} \sum_{\gamma\in\mathcal{G}(\mathbf{z})} \mathbb{I}[N_\gamma(S) < D_\gamma]$
  if the maximum gain is zero **then**
    break
  Query the outcome of $\mathbf{z}^*$ and set $S \leftarrow S \cup \{\mathbf{z}^*\}$
  for all $\gamma \in \mathcal{G}(\mathbf{z}^*)$ **do**
    $N_\gamma(S) \leftarrow N_\gamma(S) + 1$
return Acquired labeled set S

C.6 Complexity

Index construction costs $O(n \log n)$ time and $O(n)$ space. For a retrieved neighborhood $\mathcal{N}$, displacement coordinates cost $O(|\mathcal{N}|m)$ and probe-grid evaluation costs $O(|\mathcal{N}|Tn_d)$, giving $O(|\mathcal{N}|(m + Tn_d))$ time and $O(Tn_d)$ additional aggregation space. Exhaustive evaluation replaces $|\mathcal{N}|$ by n.

Coordinates $(r_i, \mathbf{u}_i)$ and outcome indicators are reused across probes. For each reach direction, sorting the in-cone observations by radius and locating the $\lceil \alpha n \rceil$-th different-outcome case costs $O(|\mathcal{N}| \log |\mathcal{N}|)$. Certification can first screen regional counts: Eq. (10) exceeds ε whenever $N_{\mathcal{D}}(\gamma) < 3\ln(3|\mathcal{G}|/\delta)/\varepsilon$, while $N_{\mathcal{D}}(\gamma) \ge n_\varepsilon^{\text{suff}}$ guarantees $w_\gamma \le \varepsilon$. Only intermediate counts require the full width computation.

A naive greedy acquisition implementation costs $O(B|\mathcal{U}||\mathcal{G}|)$. Sparse incidence lists restrict updates to the probes containing each candidate. Batch size b reduces global gain recomputations from B to $\lceil B/b \rceil$; lazy evaluation avoids unnecessary candidate refreshes as described above. Appendix E.5 measures these costs on reference samples of up to one million cases.

C.7 Feasibility-Constrained SACA

We incorporate feasibility restrictions (Ustun et al., 2019; Karimi et al., 2021; Poyiadzi et al., 2020; Mahajan et al., 2019) through admissible sets and outcome-independent weights; intervention semantics require a domain causal model.

For a query $\mathbf{x}$, let $\mathbf{\Delta} = \mathbf{z} - \mathbf{x}$. A feasibility specification consists of an admissible set $\mathcal{F}_\mathbf{x} \subseteq \mathbb{R}^m$, a cost $c_\mathbf{x} : \mathbb{R}^m \to [0, \infty)$, and an outcome-independent weight $w_\mathbf{x} : \mathbb{R}^m \to [0, 1]$. For an action budget $b_{\text{act}} \ge 0$ and temperature $\vartheta > 0$, examples include

$$\begin{aligned}
\Delta_j &= 0 && \text{for an immutable feature,}\\
\Delta_j &\ge 0 && \text{for an increase-only feature,}\\
c_\mathbf{x}(\mathbf{\Delta}) &\le b_{\text{act}} && \text{for an action budget,}\\
w_\mathbf{x}(\mathbf{\Delta}) &= \exp\{-c_\mathbf{x}(\mathbf{\Delta})/\vartheta\} && \text{for graded cost.}
\end{aligned}$$

Hard feasibility replaces a region $\mathcal{R}_\gamma$ by

$$\mathcal{R}_{\gamma,\mathcal{F}} = \{\mathbf{z} \in \mathcal{R}_\gamma : \mathbf{z} - \mathbf{x} \in \mathcal{F}_\mathbf{x}\}. \tag{25}$$

The binned and kernel estimators then use the retained observations. The feasible reach is defined for directions whose cones accumulate the prescribed feasible mass. Soft feasibility instead multiplies each regional or kernel contribution by $w_i = w_\mathbf{x}(\mathbf{x}_i - \mathbf{x})$.

If $\mathcal{F}_\mathbf{x}$ is fixed independently of the certification sample, hard filtering changes only the conditioning region. Theorem 5 therefore applies with the retained regional count. For soft costs, the following result controls the conditional weighted mean of a finite reference sequence.

Proposition 6 (Finite-reference certificate with predictable weights)**.** *Fix a query $(\mathbf{x}, y)$, let the reference cases be independent, and fix $\delta \in (0, 1)$. Set $Y_i' = \mathbb{I}[Y_i \ne y]$ and $\mathcal{H}_t = \sigma(\mathbf{X}_{1:n}, Y_1', \dots, Y_t')$. Let $w_i \in [0, 1]$ and $\widehat{Y}_i \in [0, 1]$ be $\mathcal{H}_{i-1}$-measurable, and set $W_t = \sum_{i\le t} w_i$. At every time t with*

$W_t > 0$, define

$$\widehat{p}_{w,t} = \frac{\sum_{i \le t} w_i Y_i'}{W_t}, \qquad p_{w,t} = \frac{\sum_{i \le t} w_i \mathbb{E}[Y_i' \mid \mathbf{X}_i]}{W_t}. \tag{26}$$

An empirical-Bernstein uniform boundary $u_{\delta/2}$ calibrated for $[0,1]$-bounded increments gives the simultaneous bounds

$$|\widehat{p}_{w,t} - p_{w,t}| \le \frac{u_{\delta/2}\Big(\sum_{i \le t} w_i^2 (Y_i' - \widehat{Y}_i)^2\Big)}{W_t} \quad \textit{for all } 1 \le t \le n \textit{ with } W_t > 0 \tag{27}$$

with probability at least $1 - \delta$. Ignoring logarithmic and variance-adaptation factors, the leading square-root term at time t scales as $n_{\mathrm{eff},t}^{-1/2}$, where

$$n_{\mathrm{eff},t} = \frac{W_t^2}{\sum_{i \le t} w_i^2}. \tag{28}$$

Proof. Set

$$\nu_i = \mathbb{E}[Y_i' \mid \mathbf{X}_i], \qquad \zeta_i = w_i Y_i', \qquad \widehat{\zeta}_i = w_i \widehat{Y}_i.$$

Because w_i and $\widehat{Y}_i$ are $\mathcal{H}_{i-1}$-measurable, $\widehat{\zeta}_i$ is predictable. Independence of the reference cases gives

$$\mathbb{E}[\zeta_i \mid \mathcal{H}_{i-1}] = w_i \nu_i,$$

so $\zeta_i - w_i \nu_i$ is a bounded martingale difference.

Applying the empirical-Bernstein confidence-sequence theorem of Howard et al. (2021) to ζ_i with forecast $\widehat{\zeta}_i$ gives

$$\left| \sum_{i=1}^{t} w_i (Y_i' - \nu_i) \right| \le u_{\delta/2}\left(\sum_{i=1}^{t} w_i^2 (Y_i' - \widehat{Y}_i)^2 \right) \quad \text{for every } 1 \le t \le n$$

with probability at least $1 - \delta$, after allocating $\delta/2$ to each tail. Division by W_t gives Eq. (27).

Since $(Y_i' - \widehat{Y}_i)^2 \le 1$, the intrinsic variance argument is at most $\sum_{i \le t} w_i^2$. Dividing the leading square-root term of the boundary by W_t therefore gives order

$$\frac{\sqrt{\sum_{i \le t} w_i^2}}{W_t} = n_{\mathrm{eff},t}^{-1/2}, \qquad n_{\mathrm{eff},t} = \frac{W_t^2}{\sum_{i \le t} w_i^2},$$

up to the boundary's logarithmic and lower-order terms. This proves the stated scaling. □

For outcome-independent weights measurable from the features and $W_n > 0$, Hoeffding's inequality gives a variance-free fixed-time alternative. Conditional on the features, apply it to $\sum_{i \le n} w_i (Y_i' - \nu_i)$, whose summands lie in intervals of length w_i, and divide by W_n to obtain

$$|\widehat{p}_{w,n} - p_{w,n}| \le \sqrt{\frac{\log(2/\delta)}{2 n_{\mathrm{eff}}}}, \qquad n_{\mathrm{eff}} = \frac{\big(\sum_i w_i\big)^2}{\sum_i w_i^2},$$

with probability at least $1 - \delta$. The time-uniform empirical-Bernstein result above also covers outcome-adaptive weights when they are sequentially predictable; weights learned on an independent split are fixed for this analysis.

D Experimental Setup

Theorem 5 requires the query, outcome rule, representation, and probe geometry to be fixed independently of the certification sample. Accordingly, the grid, resolutions, bandwidths, evidence level, and tolerance are prespecified for all four protocols. The synthetic certificate experiments use a fully prespecified geometry. In the real-data protocols, the local scale ρ and the PCA basis for $m > 16$ are computed once from training observations before any query is evaluated and then held fixed. Because these observations overlap the reference sample, the certificate treats these settings as fixed, which is an approximation in those runs. For joint regional statements over Q prespecified queries, we allocate failure probability $\delta/(Q|\mathcal{G}|)$ to each probe.

D.1 RQ1: Characterization of Local Evidence Structures

Datasets. We use six binary tabular benchmarks from OpenXAI (Agarwal et al., 2022). Adult (n=36,177, m=13) predicts whether annual income exceeds \$50,000; COMPAS (4,937, 7) predicts two-year recidivism; German (800, 60) and HELOC (7,896, 23) concern credit risk and repayment; and Heart (2,926, 15) and Pima (614, 8) concern heart disease and diabetes. These datasets span income, criminal justice, credit, and clinical settings, with sample sizes and dimensions that affect the effective local sample size $nh\kappa^{-(m-1)/2}$ in Theorem 2.

Following established counterfactual benchmarks (Pawelczyk et al., 2021; Guidotti, 2024), outcomes are predictions from the supplied pretrained neural classifiers. Training instances form the reference sample, and queries are test instances assigned to the negative class; preprocessing is described in Appendix C.2. The negative class denotes an adverse decision in the four nonclinical benchmarks and the absence of a condition in Heart and Pima. Thus, different-outcome cases represent favorable decisions in the former and flagged patients in the latter. Each dataset profile uses up to 100 queries from the disjoint test split.

SACA configuration. Motivated by the ten-neighbor convention in local density and plausibility measures (Smith et al., 2014; Kanamori et al., 2020; Poyiadzi et al., 2020), we set ρ to the median 10-NN distance over at most 500 training observations. Using the representation in Appendix C.2, we evaluate $T = 32$ directions and 10 distances in $[0.25\rho, 2\rho]$, with radial bandwidth $h = 0.35\rho$ and directional concentration $\kappa = 8$. Scalar prevalence is reported only when all contributing kernel estimates are finite, and inverse prevalence is capped at 999. Directional entropy is averaged over queries with total estimated directional evidence above 10^{-12}. Dataset profiles use existence-level reach with a 60° cone half-angle over all reference radii.

For the HELOC field in Figure 3, projected training references and their full-model labels define hypothetical negative queries on a 26×26 lattice over $[0.03, 0.97]^2$. We use $\rho = 0.15$, $T = 16$, $h = 0.06$, $\kappa = 32$, and distances in $[0.03, 0.35]$. The matched-distance profiles retain full-space radii and the primary top-10-PCA angular protocol.

Comparisons and metrics. We compare dataset profiles with the maximum univariate Fisher ratio F_1, 1-NN leave-one-out error N_3, and minimum-spanning-tree cross-class fraction N_1 (Ho & Basu, 2002); 10-neighbor disagreement kDN (Smith et al., 2014); and hubness skewness (Radovanović et al., 2010). Computation of N_1 and hubness uses random subsamples of at most 3,000 and 5,000 observations, respectively.

We assess profile robustness through the mean pairwise Spearman rank stability of uncapped inverse prevalence across $h/\rho \in \{0.25, 0.35, 0.50\}$ and neighborhood scales in $\{0.75\rho, \rho, 1.25\rho\}$. Each comparison includes only queries with estimable contributing probes and positive prevalence under both settings. Reach attainment is $T^{-1}\sum_{t=1}^{T}\mathbb{I}[\widehat{d}^*(\mathbf{x}, \boldsymbol{\theta}_t) < \infty]$, averaged over queries.

For matched-distance panels, we compute the median finite $d_{\min}/\rho$ among queries with defined prevalence and entropy. We expand the relative distance band through 0.12, 0.20, 0.30, and 0.50 until it contains at least three queries, then highlight those with the minimum, median, and maximum prevalence in the first adequate band.

D.2 RQ2: Landscape Estimation Accuracy and Regional Reporting Reliability

Scenarios. We use five two-class Gaussian mixtures with closed-form different-outcome posteriors, following controlled counterfactual evaluations (Wielopolski et al., 2024; Nemecek et al., 2025). They test isotropic, anisotropic, multimodal, ring-shaped, and rare different-outcome evidence. Unless specified otherwise, the class priors are equal. In two dimensions, *iso* uses means $(-1, 0)$ and $(1, 0)$ with covariance I. The *aniso* scenario retains these means but assigns covariances I and $\mathrm{diag}(0.3, 2.0)$, making evidence depend on direction at a fixed distance. The *multi* scenario combines a query-class component $\mathcal{N}(0, 0.5I)$ with three different-outcome components centered at $(2, 0)$, $(-1.5, 1.5)$, and $(0, -2.5)$; the latter have covariance $0.3I$ and mixture weights 0.4, 0.4, and 0.2. The *ring* scenario contrasts $\mathcal{N}(0, 0.4I)$ with eight equally weighted components of covariance $0.25I$ whose centers lie uniformly on a circle of radius two. Finally, *imb* uses means $(-1, 0)$ and $(1.5, 0)$, covariances I and $\mathrm{diag}(0.3, 2.0)$, and different-outcome prior $\pi_1 = 0.1$ to test reporting with sparse regional evidence.

Estimation and ground truth. Unless specified otherwise, landscape recovery, ablation, and certificate experiments use 16 directions and 10 distances in $[0.25, 2.5]$. The kernel ablation uses $n = 8{,}000$, $(h, \kappa) = (0.2, 32)$, 20 fixed queries, and 10 seeds. For each query and seed, landscape error is $\mathrm{MAE} = |\mathcal{G}|^{-1} \sum_{\gamma \in \mathcal{G}} |\widehat{p}(\mathbf{x}, \gamma) - p_0(\mathbf{x}, \gamma)|$; Table 2 averages this error across runs.

Ablations. The binned proxy is the regional average in Eq. (9), analogous to histogram regression (Györfi et al., 2002). Radial-only smoothing sets $\kappa = 0$, and broad radial smoothing sets $h = 5$. Both modify the normalized kernel estimator in Eq. (5), based on Nadaraya–Watson regression (Nadaraya, 1964; Watson, 1964); they are controlled ablations rather than published algorithms. Evaluation against p_0 also captures the binned proxy's finite-resolution approximation error.

Baselines. All pointwise comparators use the same grid, queries, and seeds:

- **k-NN proportion** (Brughmans et al., 2024; Smith et al., 2014): the different-outcome fraction among the $k = 50$ reference cases nearest to each probe center.
- **Euclidean kernel** (Artelt & Hammer, 2020; Wielopolski et al., 2024): an isotropic Gaussian Nadaraya–Watson estimator with feature-space bandwidth 0.2. With equal class-conditional bandwidths and empirical class priors, it equals the posterior obtained from class-conditional kernel densities, so we report the two formulations as one row.
- **Logistic posterior** (Kothari et al., 2024; Cheon et al., 2025): the probability predicted at each probe center by logistic regression with the default ℓ_2 penalty, fitted once per seed to the same $8{,}000$ labeled observations.
- **MLP posterior**: the corresponding probability from a multilayer perceptron with two hidden layers of 100 units, trained for at most 500 epochs.

Table 13 repeats the comparison on the ten-dimensional mixture of Appendix D.5, using eight queries and five seeds. SACA measures radii in the full space and directions in a three-dimensional subspace fitted to the 200 nearest different-outcome displacements. Each estimator is reported at its best setting over $h \in \{0.2, 0.4, 0.8, 1.2, 1.8, 2.5\}$ for the Euclidean kernel, $k \in \{25, 50, 100, 200, 400\}$ for k-NN, and $(h, \kappa) \in \{0.2, 0.4, 0.8\} \times \{8, 32, 128\}$ for SACA. Table 14 applies the fixed-grid summaries of Appendix B.7 to each estimate and the closed-form landscape. Its distance-adaptive variant sets $\kappa = (d/h)^2$ to maintain a constant angular arc length.

Certificate evaluation. Coverage uses the first 10 of the 20 fixed queries over 10 seeds. Both coverage and report-or-abstain experiments use binned regions with radial half-width 0.15 and cone half-angle 30°. Posterior landscapes are evaluated analytically; coverage targets are approximated with an independent Monte Carlo pool of size 10^6, and the abstention experiment uses a pool of size 3×10^5. Pointwise coverage is evaluated at $n = 8{,}000$ and nominal levels $1 - \delta \in \{0.80, 0.90, 0.95\}$. Report-or-abstain evaluation uses per-probe $\delta = 0.1$, tolerances $\varepsilon \in \{0.1, 0.2\}$, and reference sizes $n \in \{2{,}000, 8{,}000, 32{,}000\}$. Eligibility and macro-averaging rules appear in Appendix E.2.

Landscape visualization. Figure 12 uses 64 directions, 40 distances in $[0.25, 2.5]$, $(h, \kappa) = (0.25, 24)$, and $n = 8{,}000$. Fixed-level reach uses $\alpha = 0.02$, a 30° cone half-angle, and a posterior-weighted reference pool of size $200{,}000$.

D.3 RQ3: Complementary Properties of Direction and Reach

Endpoint evaluation. We use the five mixtures in Appendix D.2, with 32 directions, 40 distances in $[0.25, 2.5]$, $(h, \kappa) = (0.25, 24)$, a 30° cone half-angle, $\rho = 1$, and $\alpha = 0.02$. For a direction attaining the level, the endpoint is $\mathbf{x} + \widehat{d}^*(\mathbf{x}, \boldsymbol{\theta})\boldsymbol{\theta}$. Table 4 reports attainment, the fraction of query–seed evaluations with a finite crossing; endpoint posterior η; and normalized distance $\widehat{d}^*/\rho$. Crossings are estimated from $8{,}000$ labeled observations for each of 20 queries and five seeds per scenario. Posterior and distance are averaged over available endpoints. Fixed-level attainment is undefined for pointwise counterfactual methods.

Baselines. All methods use the same reference samples and queries, and classifier-based methods receive the Bayes rule $\eta \geq 1/2$. The implementations below are adaptations for this evaluation rather than the released implementations.

Conventional counterfactual methods are evaluated at their returned points:

- **Nearest retrieval** (Keane & Smyth, 2020) returns the closest observed different-outcome instance.
- **NICE** (Brughmans et al., 2024) substitutes coordinates of the nearest unlike neighbor one at a time, retains the sparsest valid substitution, and breaks ties by 10-NN density. It returns the neighbor if no single substitution is valid.
- **DiCE** (Mothilal et al., 2020) greedily selects $k = 5$ cases from the 200 nearest different-outcome instances, maximizing proximity with weight 0.5 plus log-determinant diversity under the kernel $1/(1 + d/\rho)$ with weight 1. Posterior and distance are averaged over the selected cases.
- **DACE** (Kanamori et al., 2020) selects the different-outcome instance minimizing distance in units of ρ plus the 10-neighbor local outlier factor with weight 1.
- **FACE** (Poyiadzi et al., 2020) builds a 10-NN graph over the reference sample and query, weighting each edge by its length times the negative log 10-NN density at its midpoint. It returns the different-outcome instance with the lowest shortest-path cost among those above the tenth density percentile.
- **MCCE** (Redelmeier et al., 2024) replaces its tree-based conditional model with a five-component Gaussian mixture fitted to different-outcome instances, draws $1{,}000$ samples, and returns the nearest valid one.
- **PPCEF** (Wielopolski et al., 2024) replaces its normalizing flow with the same mixture and minimizes squared distance in units of ρ minus 0.5 times log-density over valid points on a 121×121 lattice of half-width 2.5 around the query.
- **Robust region** (Maragno et al., 2024) returns the nearest valid lattice point whose radius-0.25ρ circle is valid at all 16 checked angles.

Direction-selecting methods choose among the 32 grid directions and use the same fixed-level endpoint construction:

- **Random direction** (Marsaglia, 1972) provides an isotropic angular control.
- **Local support** (Gemou et al., 2026) selects the direction whose probe point $\mathbf{x} + \rho\boldsymbol{\theta}_t$ is closest to its 20th nearest different-outcome instance.
- **Path geometry** (Sokol et al., 2025) finds shortest paths from the query to different-outcome instances on the 10-NN graph. It selects the direction whose 30° cone contains the greatest sum of $\exp(-\text{cost}/\rho)$ over paths grouped by their first hop.
- **Responsiveness** (Cheon et al., 2025) scores the four signed coordinate directions by the fraction of displacements in $[0.05, 2.5]$ at which the Bayes rule flips, then selects the grid direction closest to the highest-scoring axis.
- **Posterior-gradient oracle** (Baehrens et al., 2010) uses a finite-difference gradient of the known posterior and therefore has oracle access to the synthetic distribution.
- **SACA direction** selects the mode of $\widehat{q}_T$; **SACA reach** selects the earliest finite crossing.

Table 18 compares direction-selecting methods on identical scenario–query–seed blocks. Endpoint posterior is differenced against SACA direction where both methods attain the level, and attainment against SACA reach over all blocks. Differences are summarized by a percentile bootstrap over blocks with $5{,}000$ resamples and a fixed seed.

Crossing diagnostics. The experiments evaluate two crossing targets. Table 16 measures error in the first radial crossing of $p_0 = 0.5$ on a 150-point grid, using $h = 1.5n^{-1/5}$, $\kappa = 4n^{2/5}$, and 10 seeds. Error is evaluated only where both estimated and reference crossings are finite. Table 5 evaluates fixed-level crossings with 16 directions, 300 distances in $[0.05, 3]$, a 30° cone half-angle, eight queries, and five seeds per scenario. Reference cumulative mass uses a posterior-weighted Monte Carlo pool of size $200{,}000$. The reported local slope is the nonnegative least-squares slope of reference F within 0.25 of its crossing; the theoretical lower crossing margin in Theorem 4 is denoted by λ.

D.4 RQ4: Coverage and Efficiency of Support-Guided Acquisition

Controlled-pool replay. Following pool-based active learning (Lewis & Gale, 1994; Sener & Savarese, 2018), each run starts with $1{,}500$ labeled observations and a disjoint pool of $30{,}000$ can-

didates. The 6×12 probe grid spans distances $[0.4, 2.4]$, with radial half-width 0.20 and cone half-angle $30°$. At $\varepsilon = 0.3$ and per-probe $\delta = 0.1$, numerical inversion of the worst-case empirical-Bernstein width gives an operational count threshold $n_\varepsilon^{\mathrm{op}} = 71$. The conservative closed-form threshold from Theorem 5 is $n_\varepsilon^{\mathrm{suff}} = 171$; simultaneous certification instead uses $\delta/|\mathcal{G}|$. Batched count-deficit greedy selection acquires 25 candidates per gain computation. We report the fraction $C(S)$ of probes meeting the operational threshold over five scenarios, six queries, and three seeds. Table 21 evaluates budgets $B \in \{0, 200, 500, 1{,}000, 2{,}000, 4{,}000\}$, with $B = 2{,}000$ used for the summary comparison.

Real-pool replay. On the real benchmarks, held-out training labels are revealed only upon acquisition (Bae et al., 2025). We use the datasets, classifiers, and training-based scaling of Appendix D.1. The candidate pool contains at most $5{,}000$ cases, and each run acquires 300 labels from an initial reference of size $n_0 = \min\{1{,}500, \max\{100, \lfloor 0.4n \rfloor\}\}$. For seed s, a permutation seeded by $500+s$ determines the reference and pool split; four negative test queries are then sampled without replacement. The 6×12 grid spans $[0.5\rho, 6\rho]$, with radial half-width 0.75ρ and cone half-angle $90°$. With $\varepsilon = 0.5$ and total error probability $\delta = 0.1$ allocated across 72 probes, the operational threshold is $n_\varepsilon^{\mathrm{op}} = 82$.

Acquisition baselines and variants. We compare SACA with eight real-pool baselines. A ninth baseline, controlled-pool uncertainty, is used only in the controlled-pool replay. All methods use the adaptations specified below.

Classifier-guided and uncertainty-based selection.

- **Uncertainty Herding** (Bae et al., 2025) greedily maximizes $\sum_j w_j \max_{\mathbf{s}\in S} K(\mathbf{z}_j, \mathbf{s})$ using a Gaussian kernel of width ρ and weights $w_j = (1-\beta) + \beta u_j$, where u_j is classifier-margin uncertainty and $\beta = 0.5$. The original method varies β with budget; here it is fixed.
- **BADGE** (Ash et al., 2020) embeds each candidate by its last-layer cross-entropy gradient, the outer product of predicted probabilities minus their hard label with penultimate activations. Selection starts from the largest embedding and proceeds in k-means++ order.
- **Controlled-pool uncertainty** (Lewis & Gale, 1994) prioritizes candidates intersecting the widest initial regional intervals.
- **Classifier margin** (Lewis & Gale, 1994) ranks candidates by $1 - 2|\widehat{\pi}_1(\mathbf{z}) - 1/2|$, where $\hat{p}_1$ is the supplied classifier's positive-class probability, and acquires the least confident first.

Feature-space coverage.

- **TypiClust** (Hacohen et al., 2022) partitions the pool into $\min\{B, 500\}$ k-means clusters and, largest first, selects each cluster's member with the smallest mean distance to its 20 nearest pool neighbors.
- **ProbCover** (Yehuda et al., 2022) greedily selects the candidate whose radius-ρ ball covers the most uncovered pool points in scaled feature space, then samples randomly once all points are covered.
- **Facility** (Gonzalez, 1985; Sener & Savarese, 2018) performs farthest-first selection in scaled input space, starting from a randomly seeded candidate.

Width reduction and random acquisition.

- **Expected Width Reduction (EWR)** (Roy & McCallum, 2001; Maurer & Pontil, 2009) sums the expected one-step decrease in empirical-Bernstein width over incident probes using their current outcome proportions, and updates the score after each revealed label.
- **Random** (Lewis & Gale, 1994) samples uniformly without replacement.

Except for EWR and SACA, each method computes one candidate order and replays it sequentially. The controlled-pool replay compares Random, Controlled-pool uncertainty, Facility, ProbCover with radius 0.2, TypiClust, and batched SACA. The real-pool replay compares all eight baselines with sequential and batched SACA over three seeds; the SACA variants follow Appendix C.5, with batch size 25.

Metrics and timing. Seeded jitter of magnitude 10^{-9} breaks gain ties. Let $w_\gamma(S)$ be the observed empirical-Bernstein width after acquiring S. Width-qualified coverage is $C_{\mathrm{width}}(S) = |\mathcal{G}|^{-1} \sum_{\gamma \in \mathcal{G}} \mathbb{I}[w_\gamma(S) \le \varepsilon]$. For budgets $0 = B_0 < \cdots < B_J = B$ and acquired sets S_j, its normalized area under the coverage curve is

$$\frac{1}{B} \sum_{j=1}^{J} (B_j - B_{j-1}) \frac{C_{\mathrm{width}}(S_{j-1}) + C_{\mathrm{width}}(S_j)}{2}.$$

Selection time is the wall-clock duration of a complete $B = 300$ run, averaged over datasets, queries, and seeds. Batched SACA refreshes its scores $\lceil B/25 \rceil$ times, whereas replayed orders take B sequential bookkeeping steps.

D.5 Compute Environment and Scaling Protocol

Compute environment. Runtime experiments use an Apple M2 Pro (10 CPU cores, 16 GB RAM), macOS 15.5, Python 3.12.9, NumPy 2.3.2, SciPy 1.17.0, and scikit-learn 1.8.0.

Scaling protocol. The scalability experiment uses unit-covariance Gaussians centered at $\mp \mathbf{e}_1$ in $\mathbb{R}^{10}$. Scaling curves use $T = 32$, eight distances in $[0.3, 1.4]$, $(h, \kappa) = (0.3, 8)$, cutoff radius 1.5, and 20 queries per sample size from 10^2 to 10^6, with sample seed 42 and direction seed 3. The component benchmark averages three repetitions using $T = 32$, ten distances in $[0.25\rho, 2\rho]$, lookup cutoff 2ρ, radial half-width 0.15ρ, cone half-angle 60°, $\varepsilon = 0.3$, and total $\delta = 0.1$ allocated across the grid.

E Additional Experimental Results

E.1 Real-Data Evidence Profiles and Robustness

Evidence fields and query profiles. Figures 5–10 show full-model labels on two-feature projections in the top panels. The bottom panels retain full-space radii; German uses top-10 PCA directions. The projected features are `education-num` and `hours-per-week` for Adult, `priors_count` and `length_of_stay` for COMPAS, `duration` and `amount` for German, `sysBP` and `BMI` for Heart, and `Glucose` and `BMI` for Pima. Bottom-panel scatter plots include only queries with computable prevalence. Gray crosses indicate undefined directional entropy, and highlighted profiles require both quantities to be defined.

- **Adult** (Figure 5). High prevalence appears mainly at high `education-num` and long working hours. Evidence for low-income queries points toward this region and is directionally concentrated ($H = 0.71$, the lowest among the six), while prevalence near these queries is low.
- **COMPAS** (Figure 6). Low-risk cases cluster around fewer priors and shorter stays. Evidence for high-risk queries points toward that region, with shorter reach in its direction than elsewhere.
- **German** (Figure 7). High prevalence and short reach occur throughout the projection. The nearly flat directional profiles in panel (f) indicate little directional preference ($H = 1.00$).
- **HELOC** (Figure 8). Prevalence changes gradually across credit-bureau summaries, producing a wide range of prevalence at similar $d_{\min}/\rho$ in panel (d).
- **Heart** (Figure 9). Different-outcome cases are patients flagged by the classifier. They make up 13% of the sample and cluster at high systolic pressure. Prevalence is concentrated near the edge of the populated region, and mean directional attainment is the lowest among the six (0.53).
- **Pima** (Figure 10). About a third of patients are flagged, along a glucose–BMI gradient. Prevalence is intermediate (0.15), evidence directions align, and the highlighted queries have different prevalence despite similar $d_{\min}$.

Sorted directional profiles in panel (f) decline more sharply for Adult and Heart, whose first-rank shares are 0.6 and 0.2, than for HELOC and German, whose shares are 0.15 and 0.035. This ordering is consistent with the directional entropies in Table 1.

Dataset profiles. Tables 9–11 report valid-query counts, inverse prevalence, ranking stability, and classical data-complexity measures under Appendix D.1. Table 9 gives the number of queries for

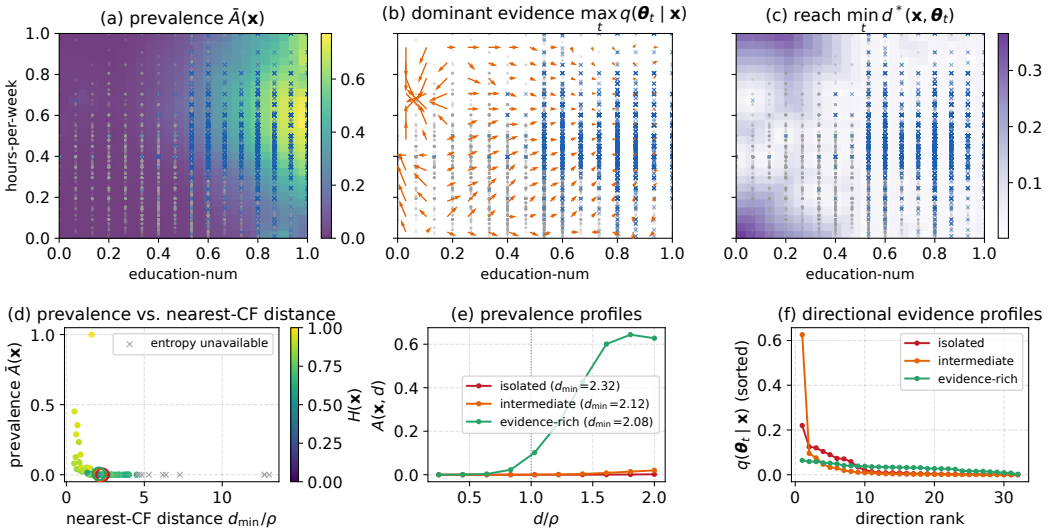


Figure 5: Adult evidence field (top) and matched-distance query profiles (bottom).

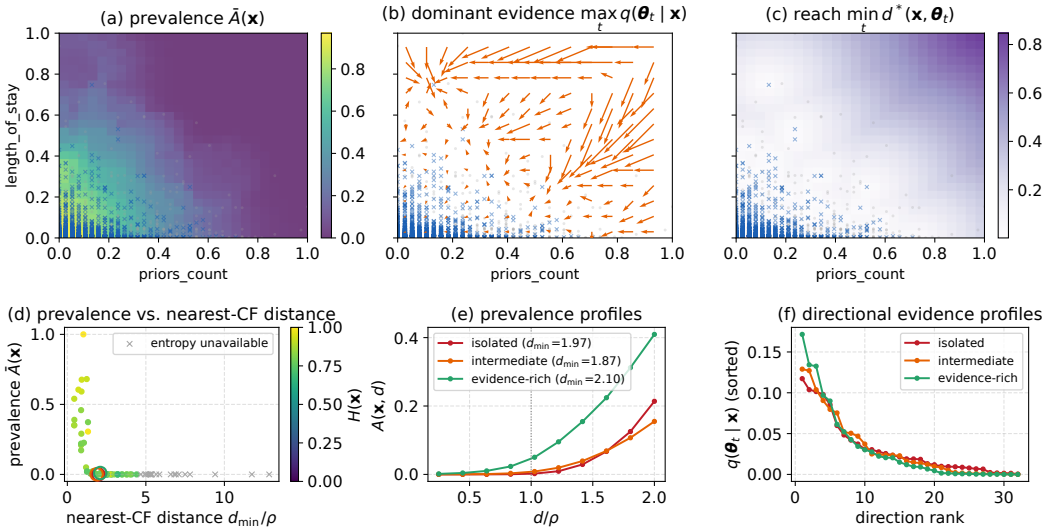


Figure 6: COMPAS evidence field (top) and matched-distance query profiles (bottom).

which each perspective is computable and the median capped inverse prevalence IA $= 1/\bar{A}$, interpretable as the inverse local frequency of different-outcome evidence. All queries are estimable on German (IA $= 1.4$) and Pima (12.5). On German, approved and rejected cases overlap substantially across its 60 largely categorical attributes. Adult and COMPAS reach the cap of 999, indicating little different-outcome evidence over the probed range. Of their 100 queries, 4 and 24, respectively, have no computable prevalence, while 13 and 37 have undefined directional entropy, shown as gray crosses in Figures 5 and 6. HELOC (17.7) and Heart (394) lie between these extremes, with almost all queries estimable.

Rankings by uncapped inverse prevalence remain stable across bandwidth and neighborhood choices: mean pairwise Spearman correlations are 0.97 and 0.99, respectively (Table 10). German has the lowest bandwidth stability (0.88); on the other datasets, the query rankings change little over the tested settings.

Table 11 gives a different ordering because its measures summarize entire datasets rather than query neighborhoods. German has the highest complexity by the reported measures ($F_1 = 0.10$,

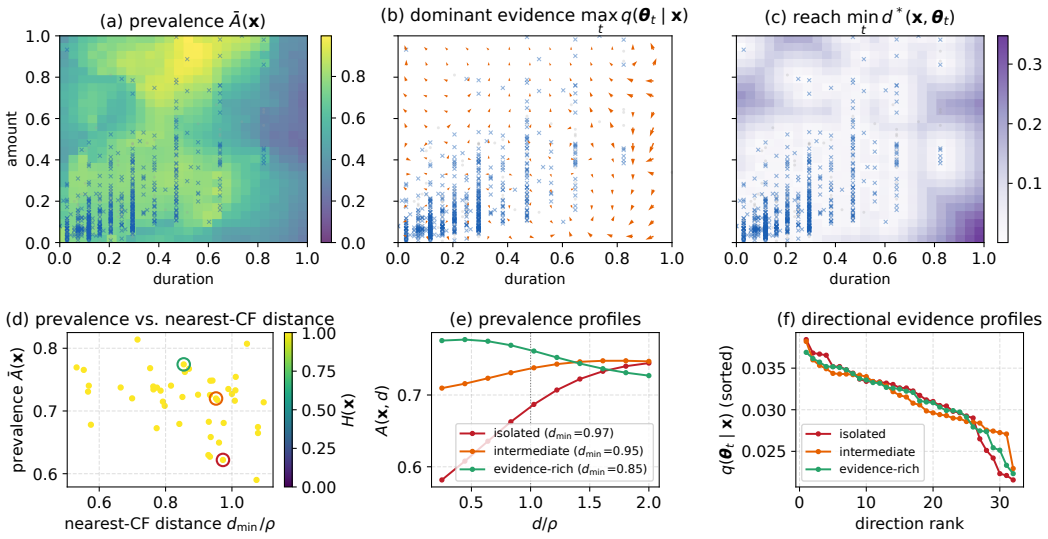


Figure 7: German evidence field (top) and matched-distance query profiles (bottom).

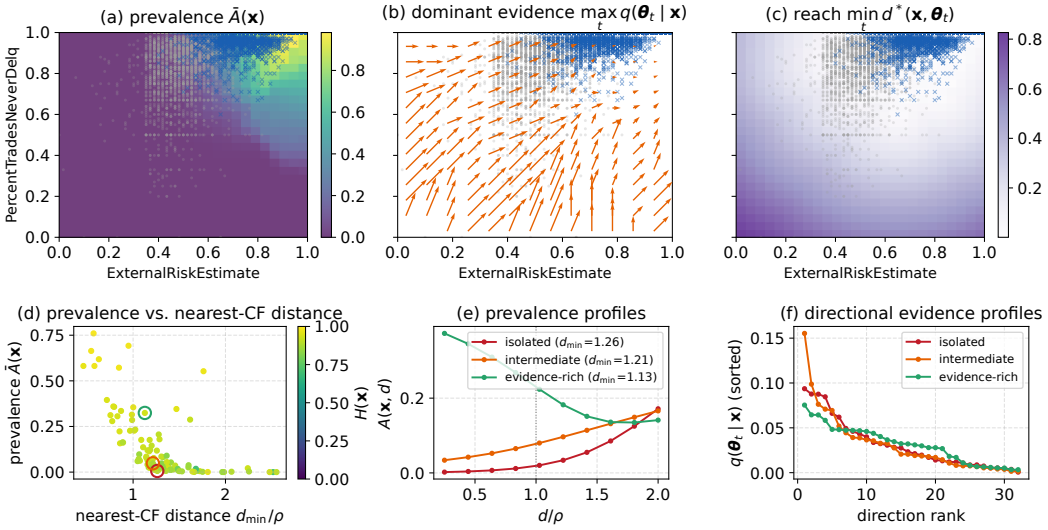


Figure 8: HELOC evidence field (top) and matched-distance query profiles (bottom).

$N_1 = N_3 = 0.29$, kDN $= 0.32$) and high local prevalence, consistent with substantial class overlap. COMPAS is the most separable ($F_1 = 2.20$, $N_1 = 0.02$) but has the fewest queries with computable prevalence. Thus, global separability does not determine the evidence available around a given query. Hubness is highest on German (1.14) and HELOC (0.93), the two highest-dimensional benchmarks.

Case study: counterfactual points and evidence landscapes. Figure 11 overlays cases returned by conventional counterfactual methods on SACA landscapes and reach curves, using the two-feature projections of Figure 3(a–c). All methods use the same reference sample and query. We include sample-based nearest retrieval (Keane & Smyth, 2020), DiCE (Mothilal et al., 2020), DACE (Kanamori et al., 2020), and FACE (Poyiadzi et al., 2020). The generator-based methods in Appendix D.3 require a validity oracle for unobserved candidates, which these projections do not provide. For each benchmark, the left and right panels show the candidate queries with the highest and lowest prevalence $\bar{A}$, respectively.

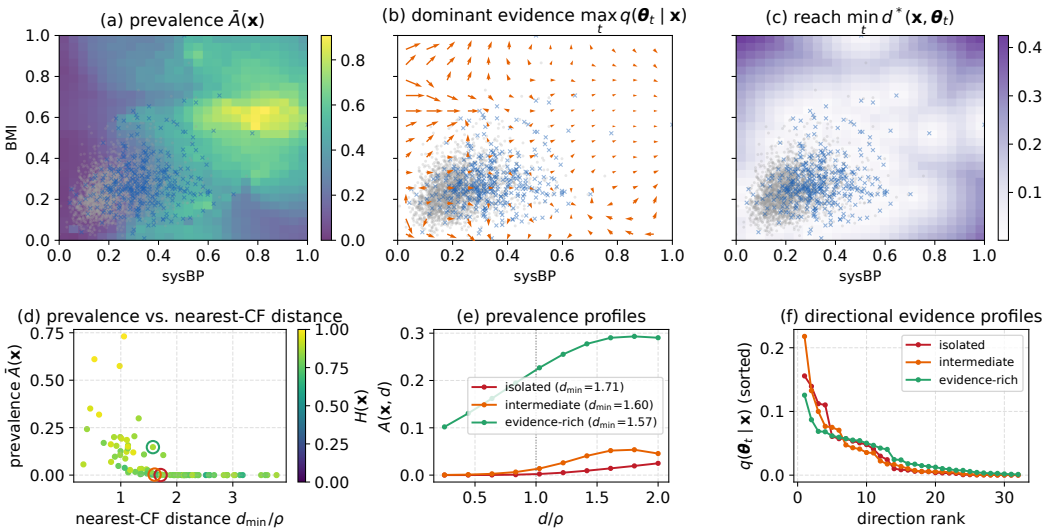


Figure 9: Heart evidence field (top) and matched-distance query profiles (bottom).

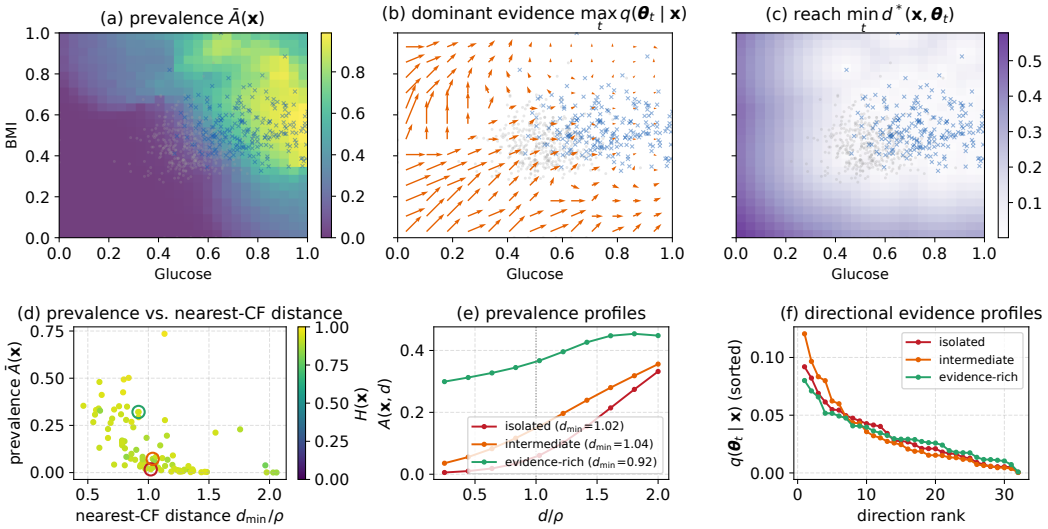


Figure 10: Pima evidence field (top) and matched-distance query profiles (bottom).

Stars mark queries, shading shows the estimated landscape $\widehat{p}$ from Eq. (5), and purple curves show existence-level reach $\widehat{d}^*$ from Eq. (22) where finite. The unshaded disc lies inside the smallest probe radius. Open markers denote returned cases: one each from nearest retrieval, DACE, and FACE, and five from DiCE. Axes show the projection features scaled to $[0, 1]$. Geometry is computed in the projection, while outcomes come from the full-space classifier.

Prevalence falls from 0.54 to 0.00 across the Adult pair, from 0.81 to 0.12 across COMPAS, and from 0.66 to 0.00 across Pima. The landscapes show broad different-outcome evidence around the first query of each pair and weaker evidence around the second. Reach is short where evidence accumulates nearby and moves outward where more distance is needed. It is absent in the second Pima panel because no direction attains the specified level within the probed range.

The point methods return cases for both queries in each pair, often at similar locations near the query. In the second Pima panel, they return cases even though no probed direction attains the existence level. The returned cases identify alternatives; the landscape and reach additionally show how much different-outcome evidence surrounds each query and where it occurs.

Table 9: Valid-query counts and median capped IA.

Dataset	valid queries ($\bar{A}$ / H/total)	median capped IA
Adult	96/87/100	999.0
COMPAS	76/63/100	999.0
German	45/45/45	1.4
HELOC	100/100/100	17.7
Heart	99/100/100	394.4
Pima	90/90/90	12.5

Table 10: Rank stability of inverse prevalence.

Dataset	bandwidth ↑	scale ↑
Adult	0.99	1.00
COMPAS	0.99	1.00
German	0.88	0.98
HELOC	0.97	0.99
Heart	0.99	1.00
Pima	0.98	1.00
Mean	0.97	0.99

Table 11: Classical data-complexity measures.

Dataset	F_1	N_1	N_3	kDN	hub.
Adult	1.29	0.063	0.027	0.050	0.30
COMPAS	2.20	0.020	0.015	0.032	-0.09
German	0.10	0.294	0.294	0.322	1.14
HELOC	2.80	0.095	0.086	0.108	0.93
Heart	1.18	0.096	0.095	0.121	0.58
Pima	1.67	0.144	0.143	0.198	0.42

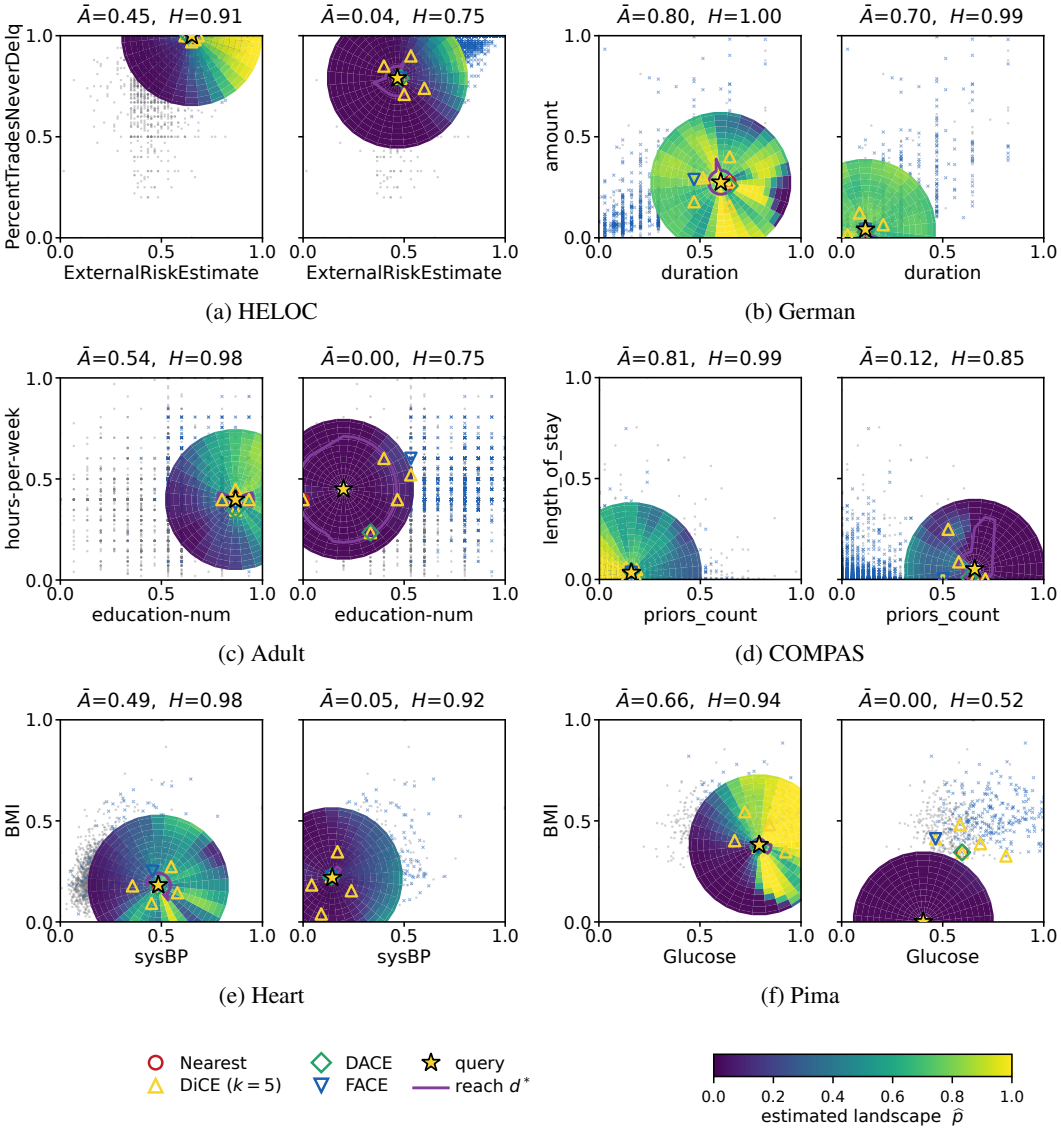


Figure 11: What conventional counterfactual methods return (open markers), drawn on the evidence SACA returns (shading: landscape $\widehat{p}$; purple: reach $\widehat{d}^*$), for the query with the largest (left) and smallest (right) prevalence on each benchmark.

E.2 Synthetic Recovery and Certificate Coverage

Landscape recovery. Figure 12 shows one column per scenario. The top row displays the reference sample, query, different-outcome cases, and estimated reach. The remaining rows show the

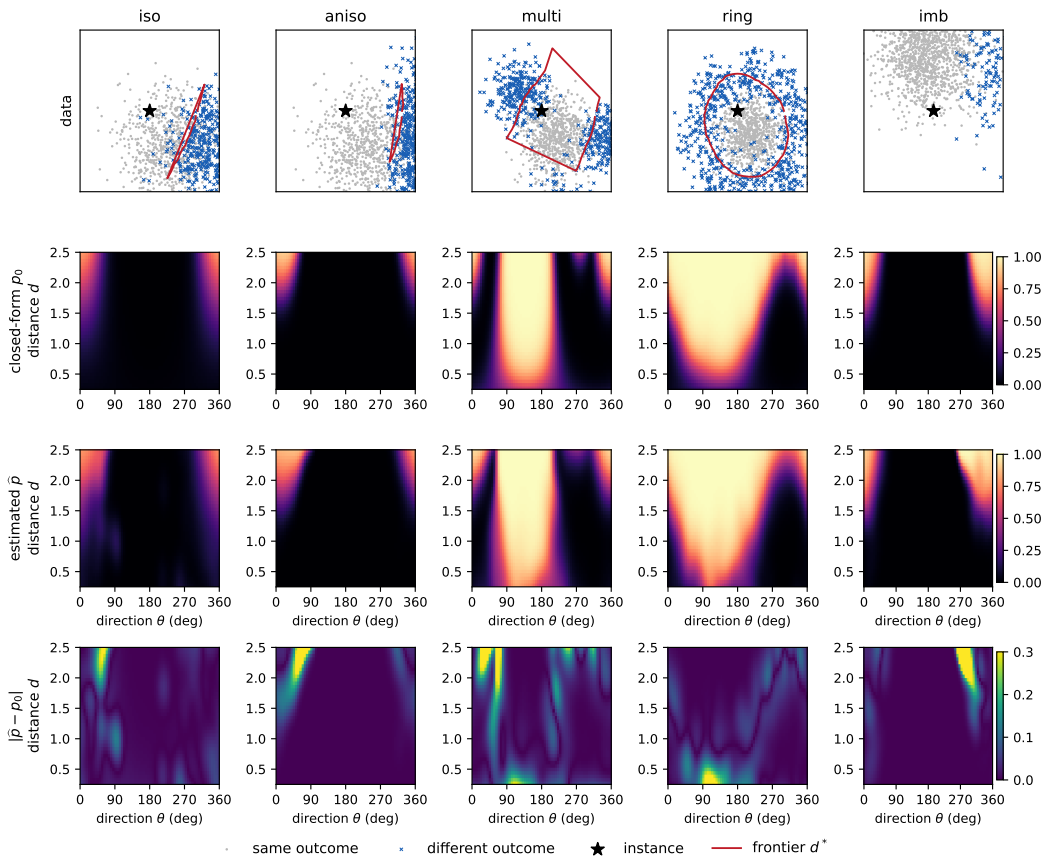


Figure 12: Synthetic evidence-landscape recovery. Columns represent mixture scenarios; rows show sampled observations, the closed-form ground truth p_0, the estimate $\widehat{p}$, and pointwise absolute error. Overlaid reach estimates mark the first radial crossing of a fixed positive cumulative-evidence level.

closed-form landscape p_0, its estimate $\widehat{p}$ from $8{,}000$ samples, and their absolute difference, with direction $\boldsymbol{\theta}$ horizontally and distance d vertically. The estimator recovers the main pattern in each scenario: one directional band in *iso*, a narrower band in *aniso*, three pockets in *multi*, evidence at a common radius in *ring*, and weak evidence in *imb*. Errors are greatest near sharp transitions, where smoothing blurs the landscape, and at outer radii, where fewer observations contribute to each kernel window. These patterns are consistent with the bias and effective-sample-size terms in Theorem 2.

Pointwise landscape error. Under the protocol of Appendix D.2, Table 12 adds three estimators to Table 2. The MLP posterior has the lowest mean MAE (0.020), followed by the Euclidean kernel (0.023), k-NN proportion (0.026), and SACA (0.029). The Euclidean kernel and k-NN estimate the posterior using cases near each probe center, while the MLP fits a global posterior. SACA instead smooths in query-centered radial and angular coordinates (Eq. (1)); at fixed angular width, its footprint grows with radius. Relative to the Euclidean kernel, its additional error is 0.019 in *multi*, where separate pockets occur at different radii, and at most 0.005 in the other scenarios. The MLP's pointwise advantage does not address the separate regional report-or-abstain task of Section 3.2.

Higher-dimensional recovery. Table 13 repeats the comparison on the ten-dimensional mixture of Appendix D.5, using directions in a three-dimensional subspace and each estimator's best tested setting. Local-estimator MAEs rise to 0.26–0.31, with the binned proxy at 0.312, while the MLP reaches 0.145. SACA is within 0.004 of the Euclidean kernel and outperforms the k-NN proportion. The deterioration of local estimates is consistent with the reduced effective sample size in higher dimensions described by Theorem 2.

Table 12: Landscape MAE (↓) for additional estimators under the protocol in Appendix D.2. Bold marks the lowest error among these estimators.

Method	*iso*	*aniso*	*multi*	*ring*	*imb*	Mean
k-NN proportion	0.032	0.022	0.041	0.027	0.010	0.026
Euclidean kernel (density ratio)	0.028	0.017	**0.040**	**0.020**	0.010	0.023
MLP posterior	**0.021**	**0.012**	0.041	0.021	**0.006**	**0.020**

Table 13: Landscape MAE (↓) on the ten-dimensional mixture.

Method	MAE ↓
SACA	0.263
Binned	0.312
Euclidean kernel	0.259
k-NN proportion	0.281
MLP posterior	**0.145**
Logistic posterior	0.241

Table 14: Perspective-level error of the landscape estimators.

Estimator	$A(d)$ MAE	$\bar{A}$ error	q L_1	H error
SACA	0.019	0.013	0.133	0.026
SACA, distance-adaptive κ	0.013	0.011	**0.120**	**0.024**
Binned	0.023	**0.007**	0.206	0.046
Euclidean kernel	0.013	0.008	**0.120**	**0.024**
k-NN proportion	**0.012**	**0.007**	0.142	0.029

Perspective-level error. Table 14 evaluates the resulting perspectives against the closed-form landscape: the MAE of the prevalence curve $A(d)$, absolute error of scalar prevalence $\bar{A}$, L_1 error of the direction distribution q, and absolute error of normalized entropy H. Setting $\kappa = (d/h)^2$ keeps the angular window's arc length constant across radii. Relative to fixed κ, this distance-adaptive variant reduces prevalence-curve MAE from 0.019 to 0.013, direction error from 0.133 to 0.120, and entropy error from 0.026 to 0.024, matching the Euclidean kernel on these three measures.

The perspective-level ranking differs from the pointwise ranking. The binned proxy ties for the lowest scalar-prevalence error (0.007), as averaging regional proportions can attenuate local errors, but has the highest direction and entropy errors (0.206 and 0.046), for which coarse cells lose directional detail. As Proposition 3 shows, error in q also depends on total directional evidence μ. Pointwise MAE alone therefore does not determine which estimator best recovers each perspective.

Certificate coverage. Table 15 evaluates pointwise empirical-Bernstein coverage under Appendix D.2. Each scenario has approximately $15{,}000$ eligible query–probe–seed triples, each with at least two reference cases and 200 independent Monte Carlo observations in the corresponding region. Empirical coverage is at least 99.97% at every evaluated nominal level. Its limited variation across levels is consistent with the conservative width in Eq. (10), including the union-bound allocation across $|\mathcal{G}|$ probes.

Posterior crossing. Table 16 measures error in the first radial crossing of $p_0 = 0.5$, using the adaptive bandwidths of Appendix D.3. Error falls in every scenario as n increases from $2{,}000$ to $32{,}000$. It is lowest for *imb* (0.157 to 0.064), where the posterior rises steeply through 0.5, and highest for *multi* (0.227 to 0.112), where crossings between evidence pockets can be shallow. A shallower crossing converts a given posterior error into a larger radial error. Across scenarios, error falls by 1.7–$2.5\times$. This posterior-level crossing differs from the cumulative-mass reach evaluated in Table 5.

Reporting and abstention. Table 17 reports mean run-level maximum reported error and violation rate across five scenarios, eight queries, and five seeds. A target is included only if its Monte Carlo region contains at least 50 observations. Metrics are macro-averaged over query–seed runs; reported-side metrics exclude runs without reported probes, and abstained MAE excludes runs without abstained probes. No reported estimate exceeds its tolerance in the evaluated runs. The reporting fractions at different sample sizes and tolerances appear in Table 3.

E.3 Reach Stability and Endpoint Quality

Paired direction comparisons. Table 18 compares direction-selecting methods on identical scenario–query–seed blocks under Appendix D.3. Endpoint posterior is compared with SACA direction where both methods attain $\alpha = 0.02$. Every competitor has a lower mean posterior: differences

Table 15: Pointwise coverage of the regional empirical-Bernstein certificate.

	Nominal level $1-\delta$		
Scenario	0.95	0.90	0.80
iso	1.0000	1.0000	1.0000
aniso	1.0000	0.9999	0.9998
multi	1.0000	1.0000	1.0000
ring	1.0000	0.9999	0.9997
imb	1.0000	1.0000	0.9999

Table 16: First-crossing MAE for the pointwise posterior threshold $p_0 = 0.5$.

Scenario	$n = 2{,}000$	$n = 32{,}000$
iso	0.190	0.112
aniso	0.194	0.085
multi	0.227	0.112
ring	0.199	0.110
imb	0.157	0.064

Table 17: Maximum reported error and tolerance violations.

ε	n	mean run-level max. reported $\lvert\texttt{error}\rvert$	reported violation rate
0.10	2,000	0.007	0.000
0.10	8,000	0.033	0.000
0.10	32,000	0.041	0.000
0.20	2,000	0.034	0.000
0.20	8,000	0.067	0.000
0.20	32,000	0.060	0.000

Table 18: Mean paired differences over 500 blocks; negative values favor SACA.

	Paired runs	Endpoint posterior vs. SACA direction	Attainment vs. SACA reach
Random direction	242	-0.113	-0.372
Local support	420	-0.028	$+0.000$
Path geometry	393	-0.080	-0.066
Responsiveness	420	-0.012	-0.008
Posterior-gradient oracle	405	-0.015	-0.046
SACA direction	500	–	-0.030
SACA reach	420	-0.032	–

are -0.113 for random direction over 242 blocks, -0.080 for path geometry, -0.028 for local support, -0.015 for the posterior-gradient oracle, and -0.012 for responsiveness. The oracle follows the true posterior gradient at the query but does not account for evidence farther away. In *multi*, it reaches 0.84 versus SACA direction's 0.90, as the steepest local ascent need not lead toward the strongest accumulated evidence.

Attainment is compared with SACA reach over all blocks. SACA reach attains the level whenever any grid direction does, so no method exceeds its attainment; local support ties it ($+0.000$), while random direction is 0.372 lower. In Table 4, SACA reach also has the shortest normalized distance (1.45ρ) and highest attainment (0.87, tied with local support), at a posterior 0.032 below SACA direction. This reflects their different selection criteria: SACA direction maximizes accumulated directional evidence, whereas SACA reach selects the earliest finite crossing.

Conventional endpoints. Table 19 evaluates conventional counterfactual methods at their returned points. Nearest retrieval, NICE, DACE, and FACE yield similar mean posteriors and displacements (0.11 at 0.47ρ). DiCE's diverse set reaches 0.183 at 0.96ρ; MCCE and PPCEF reach about 0.55; and the robust-region control reaches 0.749. These posteriors describe endpoints chosen under different objectives and are not directly comparable with the fixed-level endpoints. In particular, the generator-based methods and robust-region control select points valid under the Bayes rule, at displacements of at least 1.26ρ.

Scenario-level endpoint quality. Table 20 breaks down posterior at the first $\alpha = 0.02$ crossing by scenario. SACA direction is best or tied for best in *iso*, *multi*, and *ring*, while random direction has a higher conditional average in *aniso* (0.75 versus 0.72) and *imb* (0.84 versus 0.81). These averages include only blocks where each method attains the level. Random direction attains it in just 0.50 of blocks (Table 4), so the methods are averaged over different subsets. The paired results in Table 18 provide the comparison on common blocks.

E.4 Support-Guided Acquisition

Table 21 reports the fraction of probes meeting the operational count threshold as acquisition budget increases. All methods start at 0.10. At $B = 200$, batched SACA reaches 0.15, tied with the best baseline, but its advantage grows as acquisitions accumulate: at $B = 500$ it reaches 0.24 versus at most 0.19, and at $B = 1{,}000$ it reaches 0.48 versus at most 0.28. By $B = 2{,}000$, SACA qualifies 0.86 of probes, compared with 0.34–0.42 for the baselines. Its coverage remains 0.86 at $B = 4{,}000$,

Table 19: Conventional counterfactual methods: endpoint posterior, normalized distance, and per-scenario posteriors.

	Overall		Endpoint posterior by scenario				
Method	Posterior↑	Dist./ρ ↓	*iso*	*aniso*	*multi*	*ring*	*imb*
Nearest retrieval	0.108	**0.47**	0.20	0.12	0.10	0.09	0.02
NICE	0.108	**0.47**	0.20	0.12	0.10	0.09	0.02
DiCE	0.183	0.96	0.22	0.20	0.18	0.22	0.10
DACE	0.109	0.48	0.20	0.12	0.10	0.09	0.02
FACE	0.107	0.48	0.20	0.12	0.10	0.09	0.02
MCCE	0.552	1.30	0.55	0.57	0.58	0.55	0.52
PPCEF	0.551	1.26	0.55	0.56	0.57	0.56	0.51
Robust region	**0.749**	1.44	**0.66**	**0.75**	**0.81**	**0.80**	**0.72**

Table 20: Direction-selecting methods: per-scenario posteriors at the first crossing of $\alpha = 0.02$.

Method	*iso*	*aniso*	*multi*	*ring*	*imb*
Random direction	**0.58**	**0.75**	0.74	0.83	**0.84**
Local support	0.53	0.71	0.86	0.94	0.79
Path geometry	0.52	0.68	0.74	0.89	0.80
Responsiveness	**0.58**	0.72	0.86	**0.95**	0.80
Posterior-gradient oracle	**0.58**	0.71	0.84	0.94	0.80
SACA direction	**0.58**	0.72	**0.90**	**0.95**	0.81
SACA reach	0.53	0.71	0.85	0.93	0.79

Table 21: Controlled-pool fraction meeting the per-probe count threshold. Bold marks the best result at each positive budget.

Budget B	Random	Uncertainty	Facility	ProbCover	TypiClust	SACA (batched)
0	0.10	0.10	0.10	0.10	0.10	0.10
200	0.13	0.13	0.11	0.13	**0.15**	**0.15**
500	0.19	0.18	0.13	0.14	0.16	**0.24**
1,000	0.26	0.28	0.19	0.23	0.24	**0.48**
2,000	0.40	0.42	0.34	0.38	0.38	**0.86**
4,000	0.54	0.57	0.58	0.54	0.56	**0.86**

while the baselines rise to 0.54–0.58. The growing gap reflects the objective used for selection: a candidate can reduce support deficits in several overlapping probe regions, and coverage increases once those regions cross the threshold. The plateau indicates that additional acquisitions in this replay do not qualify further probes. This experiment measures *count-qualified* coverage; the real-pool experiment instead measures empirical-Bernstein *width-qualified* coverage, which depends on both regional case counts and observed outcome variance (Eq. (10)).

Real-pool results. Table 22 reports width-qualified coverage by dataset. Mean initial coverage is 0.195, and sequential and batched SACA have identical final coverage at the reported precision on all six datasets. HELOC and Heart begin highest (0.49 and 0.41) and reach 0.70 and 0.66 after 300 acquisitions. Their high initial coverage is consistent with the broadly computable query profiles in Table 9. Adult and COMPAS begin near zero (0.002 and 0.000). SACA raises their coverage to 0.35 and 0.12, versus 0.28 and 0.06 for EWR and at most 0.01 for any other baseline. These low initial values are consistent with the sparse evidence around some queries reported in Appendix E.1. German and Pima have smaller initial samples (320 and 245) and pools (480 and 369). SACA reaches 0.22 and 0.39, versus 0.13–0.16 and 0.31–0.33 for the baselines. The seven baselines that do not use probe regions finish within 0.02 of one another on every dataset. EWR and SACA target regional support deficits, with SACA achieving higher coverage by crediting a candidate toward every deficient region it covers before its label is revealed. Lazy evaluation reduces exact sequential greedy selection time by $9.4\times$ to 0.011 seconds per run, about $10\times$ faster than EWR.

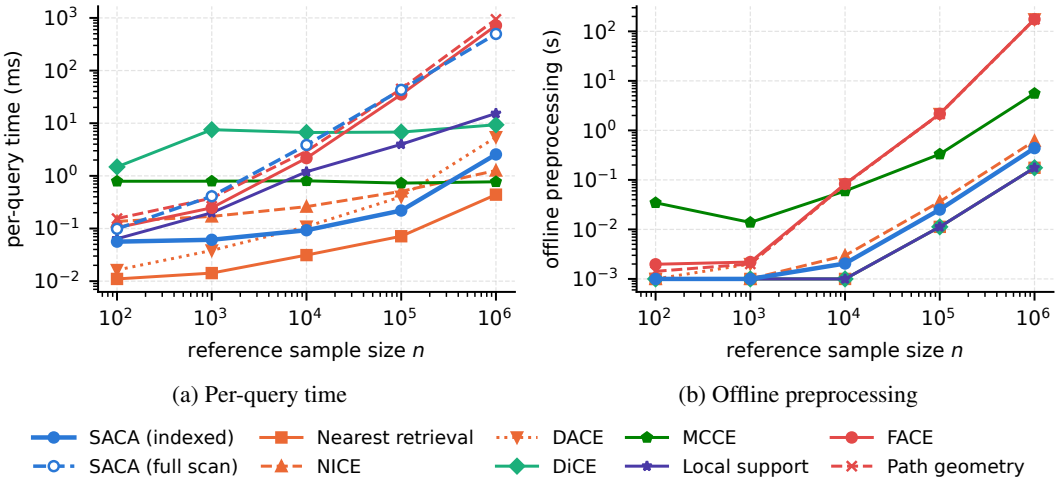
E.5 Computational Scalability

Indexed evaluation. Figure 13 and Table 23 benchmark the radius-truncated indexed implementation under the protocol in Appendix D.5. At 10^6 reference cases, landscape and perspective evaluation takes 2.56 ms per query. The spatial index is built once; acquisition examines at most $30{,}000$ candidates per step, and peak RSS denotes maximum resident-set size. The measured range, from 10^2 to 10^6 cases, contains the sample sizes of all real benchmarks. All statistical experiments and the full-scan baseline instead use the exhaustive estimator in Eq. (5). Because radius truncation changes that estimator, the indexed results establish computational cost only.

Comparison methods. Figure 13 evaluates the data-grounded methods of Appendix D.3 under the same protocol, including their required offline structures and a shared k-d tree where applicable.

Table 22: Final width-qualified coverage by dataset at $B = 300$ for every acquisition strategy: sequential and batched count-deficit greedy, the latter being the SACA variant reported in Table 6, followed by EWR, Uncertainty Herding, BADGE, ProbCover, TypiClust, classifier margin, farthest-first facility, and random.

Dataset	Seq.	Batch	EWR	Herd.	BADGE	ProbC.	TypiC.	Margin	Facil.	Rand.
Adult	**0.35**	**0.35**	0.28	0.01	0.00	0.01	0.00	0.01	0.00	0.01
COMPAS	**0.12**	**0.12**	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
German	**0.22**	**0.22**	0.16	0.14	0.14	0.14	0.14	0.14	0.13	0.14
HELOC	**0.70**	**0.70**	0.63	0.51	0.51	0.51	0.51	0.50	0.52	0.51
Heart	**0.66**	**0.66**	0.59	0.46	0.46	0.46	0.46	0.46	0.46	0.46
Pima	**0.39**	**0.39**	0.33	0.31	0.31	0.31	0.31	0.31	0.31	0.31
Mean	**0.41**	**0.41**	0.34	0.24	0.24	0.24	0.24	0.24	0.24	0.24



(a) Per-query time (b) Offline preprocessing

Figure 13: Runtime as the reference sample grows, for SACA and the data-grounded comparison methods of Appendix D.3. (a) Wall-clock time to answer one query, which for SACA is the full landscape and its three perspectives and for each comparison method its returned instance or direction. (b) Time to build the structure each method needs once per reference sample. Both axes are logarithmic; methods of one family share a color.

Lookup-based methods (nearest retrieval, NICE, DiCE, MCCE, and local support) cost 0.44–15 ms per query at $n = 10^6$. DiCE spends 6.66–9.34 ms on greedy selection from 200 candidates. Indexed SACA costs 0.056 ms at $n = 10^2$ and 2.56 ms at 10^6, whereas exhaustive full-scan evaluation rises from 0.099 ms to 496 ms.

FACE and path geometry search a sample-wide 10-NN graph for each query, costing 718 ms and 941 ms at $n = 10^6$. Building that graph takes 175 s, compared with 0.44 s for SACA's spatial index; DACE also requires offline neighbor-based outlier scores. Thus, indexed landscape evaluation has per-query cost comparable to the lookup-based methods and is over two orders of magnitude faster than the graph-search methods at the largest measured sample size.

E.6 Parameter Sensitivity

Tables 24–27 vary one parameter at a time from the defaults $T = 32$, $n_d = 10$, $[0.25\rho, 2\rho]$, and $\psi = 60°$. Each parameter has four settings evaluated on 30 queries from each of six benchmarks. Entries average per-dataset medians of relative change in prevalence and absolute change in entropy and reach attainment. Prevalence averages the landscape over directions and radii up to ρ; entropy summarizes the direction distribution. Existence-level reach uses all reference radii, so its attainment is unaffected by the radial grid but can change with the sampled directions or cone half-angle.

Table 23: Runtime and memory of the radius-truncated implementation.

	Runtime (ms)					Memory (MB)	
n	index	lookup	landscape	cert.	acquisition	data	peak RSS
10^2	0.0	0.02	0.16	0.07	1.3	0.0	89
10^3	0.1	0.03	0.28	0.20	1.8	0.1	91
10^4	2.1	0.17	2.32	0.64	9.9	0.8	116
10^5	26.1	0.98	1.88	1.48	25.1	8.4	198
10^6	458.8	7.31	1.17	1.26	28.4	83.9	576

Table 24: Number of directions T.

Setting	Prevalence	Entropy	Attainment
$T = 8$	0.132	0.104	0.078
$T = 16$	0.075	0.031	0.034
$T = 64$	0.060	0.016	0.034
$T = 128$	0.071	0.025	0.049

Table 25: Number of radii n_d.

Setting	Prevalence	Entropy	Attainment
$n_d = 5$	0.110	0.001	0.000
$n_d = 20$	0.169	0.001	0.000
$n_d = 30$	0.144	0.001	0.000
$n_d = 40$	0.132	0.001	0.000

Table 26: Outer radius $\overline{d}$.

Setting	Prevalence	Entropy	Attainment
$\overline{d} = 1.5\rho$	0.124	0.011	0.000
$\overline{d} = 2.5\rho$	0.225	0.007	0.000
$\overline{d} = 3\rho$	0.055	0.013	0.000
$\overline{d} = 4\rho$	0.131	0.028	0.000

Table 27: Cone half-angle ψ.

Setting	Prevalence	Entropy	Attainment
$\psi = 30°$	0.000	0.000	0.432
$\psi = 45°$	0.000	0.000	0.292
$\psi = 75°$	0.000	0.000	0.370
$\psi = 90°$	0.000	0.000	0.518

Number of directions T. Table 24 shows the largest entropy change at $T = 8$ (0.104), where the coarse angular grid and normalization by $\log T$ both affect the summary. From $T = 16$ upward, changes are 0.06–0.075 for prevalence and 0.016–0.031 for entropy, with no consistent decline at $T = 128$. These values also reflect resampling the random direction grid at each setting. Attainment changes by 0.034–0.078 because different sampled directions can contain different-outcome cases within their $60°$ cones.

Number of radii n_d. Changing n_d moves prevalence by 0.110–0.169 in Table 25, while entropy changes by at most 0.001 and attainment is unchanged. The radii serve as quadrature points for $\bar{A}$ on $[\underline{d}, \rho]$; at $n_d = 5$, only two lie below ρ. Changing their placement therefore affects this average. The distance-weighted direction distribution in Eq. (7) is less sensitive to the tested placements, and reach is evaluated over all reference radii.

Outer radius $\overline{d}$. With n_d fixed, changing $\overline{d}$ also changes the spacing of radii below ρ. In Table 26, prevalence changes by 0.055–0.225, with the largest change at $\overline{d} = 2.5\rho$. Entropy changes by 0.007–0.028, with the largest change at 4ρ, where more grid points lie far from the query. Attainment remains unchanged.

Cone half-angle ψ. Prevalence and entropy do not use ψ and are unchanged in Table 27. Relative to the default $60°$, attainment changes by 0.432 at $30°$, 0.292 at $45°$, 0.370 at $75°$, and 0.518 at $90°$. The cone angle changes which cases contribute to reach, so attainment should be compared at a common ψ. Dataset profiles use $60°$, whereas the RQ3 endpoint experiments use $30°$.

Overall, prevalence is sensitive to radial discretization, entropy is comparatively stable under the tested radial changes, and reach attainment is insensitive to that discretization but changes with cone angle and direction sampling.

F Extended Related Work

This section expands Section 6 across counterfactual construction, model guarantees, local support, and data acquisition. Table 28 compares representative methods by their outputs, use of observed data, and formal guarantees.

Table 28: Families of prior work and what they select and guarantee. No family estimates the different-outcome rate itself, and the guarantees listed concern model predictions (their stability over models or perturbations, or their conformal coverage) rather than the accuracy of a regional evidence estimate.

Family	Representative work (earlier → recent)	Returns	Role of data support	Formal guarantee
Model-based optimization	(Wachter et al., 2018; Lucic et al., 2022)	one point	none	none
Case retrieval	(Keane & Smyth, 2020; Brughmans et al., 2024)	one observed point	nearest unlike neighbor	none
Plausibility: connectivity or density	(Laugel et al., 2019; Poyiadzi et al., 2020; Artelt & Hammer, 2020; Kanamori et al., 2020; Joshi et al., 2019; Pawelczyk et al., 2020a; Van Looveren & Klaise, 2021; Kenny & Keane, 2021; Wielopolski et al., 2024; Nemecek et al., 2025; Redelmeier et al., 2024; Zhang et al., 2026; Tan et al., 2026)	one point	ϵ-chain, graph, learned density, LOF, or prototype	none
Diversity among alternatives	(Russell, 2019; Mothilal et al., 2020; Vo et al., 2023; Nguyen et al., 2023; Furman et al., 2025; Bobek et al., 2025; Arefeen et al., 2025)	a set of points	prototypes or a density term	none
Paths to the query	(Laugel et al., 2023; Sokol et al., 2025; Xuan et al., 2025)	a path	a graph over observed cases	none
Actionable changes	(Ustun et al., 2019; Karimi et al., 2021; Kothari et al., 2024; Kanamori et al., 2024; Cheon et al., 2025; Budde et al., 2026)	point or per-feature scores	none; reachability is defined on the model	flags reasons without recourse
Shared recourse rules	(Rawal & Lakkaraju, 2020; Kanamori et al., 2022; Ley et al., 2023; Bewley et al., 2024; Kavouras et al., 2026; Furman et al., 2026; Chatzis et al., 2026; König et al., 2025; Ehyaei et al., 2025)	rules, translations, or population effects	none	none
Guarantees about the model	(Maragno et al., 2024; Hamman et al., 2023; Jiang et al., 2024a;b; Turbal et al., 2025; Noorani et al., 2025; Kostrzewa et al., 2026; Schut et al., 2021; Altmeyer et al., 2024; Löfström et al., 2024; Bilkhoo et al., 2025; Zhu et al., 2027)	a region, a robust point, or a prediction set	none; uses the model's own energy or uncertainty	prediction or coverage stable over models, perturbations, or a conformal set
Data support as a descriptor	(Pawelczyk et al., 2020b; Gemou et al., 2026)	a scalar per query	k-NN distance to the target class	cost bound under multiplicity; none on the support itself
SACA (ours)	—	landscape over distance and direction; prevalence, direction, reach	regional counts enter both the estimate and its width	simultaneous $\pm\varepsilon$ for all reported regions; abstains otherwise and acquires to close deficits

Counterfactual generation and plausibility. Counterfactual methods optimize inputs against a model (Wachter et al., 2018; Lucic et al., 2022) or retrieve and adapt observed cases (Keane & Smyth, 2020; Brughmans et al., 2024). Plausibility can be encouraged through connectivity to data (Laugel et al., 2019; Poyiadzi et al., 2020), estimated density (Artelt & Hammer, 2020;

Kanamori et al., 2020; Wielopolski et al., 2024; Tan et al., 2026), autoencoders and prototypes (Joshi et al., 2019; Pawelczyk et al., 2020a; Van Looveren & Klaise, 2021; Kenny & Keane, 2021), or learned data distributions (Nemecek et al., 2025; Redelmeier et al., 2024; Zhang et al., 2026). These mechanisms assess whether an alternative is compatible with the feature distribution, whereas SACA measures how frequently different outcomes occur around the query. A dense region can still contain mostly original-outcome cases.

Sets and paths. Counterfactual-set methods return multiple alternatives (Russell, 2019; Mothilal et al., 2020), often balancing diversity, plausibility, and actionability (Vo et al., 2023; Nguyen et al., 2023; Furman et al., 2025; Bobek et al., 2025; Arefeen et al., 2025). Path-based methods examine routes through observed cases (Laugel et al., 2023; Sokol et al., 2025), including execution under noisy actions (Xuan et al., 2025). These methods identify endpoints or routes; SACA summarizes the prevalence and statistical support of different-outcome evidence across the surrounding distance–direction landscape.

Actionable recourse. Actionable recourse constrains permissible changes (Ustun et al., 2019; Cheon et al., 2025; Budde et al., 2026), while verification-oriented methods determine or score whether a prediction can change under specified constraints (Kothari et al., 2024; Kanamori et al., 2024). Appendix C.7 incorporates feasibility constraints into SACA's regional analysis. Its evidence summaries do not guarantee that an intervention will cause the desired outcome.

Population-level rules. Population-level methods produce explanation trees (Kanamori et al., 2022), metarules (Bewley et al., 2024), or shared recourse rules and actions (Rawal & Lakkaraju, 2020; Ley et al., 2023; Kavouras et al., 2026; Furman et al., 2026; Chatzis et al., 2026). Other work examines what happens when many individuals act on recommendations (König et al., 2025; Ehyaei et al., 2025). SACA describes evidence around an individual query without prescribing a shared action or modeling population-wide responses.

Model guarantees. Maragno et al. (2024) certify regions where a target prediction remains valid. Related work studies predictions across model or perturbation sets (Hamman et al., 2023; Jiang et al., 2024a;b; Turbal et al., 2025; Noorani et al., 2025; Kostrzewa et al., 2026). Conformal methods provide finite-sample prediction-set coverage guarantees (Altmeyer et al., 2024; Löfström et al., 2024; Bilkhoo et al., 2025), and other methods account for classifier uncertainty when generating or assessing explanations (Schut et al., 2021; Zhu et al., 2027). SACA's guarantee instead concerns the accuracy of a regional evidence estimate: its report-or-abstain rule distinguishes reliably low different-outcome prevalence from insufficient observations.

Local support. Pawelczyk et al. (2020b) relate local data support to counterfactual cost under model multiplicity. Gemou et al. (2026) characterize support using nearby target-class observations; their k-th nearest target-class distance can be viewed as crossing an empirical mass level $\alpha = k/n$ in a direction-agnostic neighborhood. SACA also resolves how evidence varies by direction and whether observations support each regional estimate. For a loan applicant, this distinguishes a direction with reliably few observed approvals from one where approval prevalence remains uncertain because the reference sample is sparse.

Coverage-based acquisition. Pool-based active learning acquires labels to improve prediction (Lewis & Gale, 1994; Roy & McCallum, 2001; Deng et al., 2023). Selection criteria include representative coverage (Sener & Savarese, 2018; Yehuda et al., 2022; Hacohen et al., 2022), submodular informativeness (Wei et al., 2015), and predictive uncertainty combined with coverage (Ash et al., 2020; Bae et al., 2025). Core-set and ProbCover methods select representative points or cover the pool with fixed-radius balls (Sener & Savarese, 2018; Yehuda et al., 2022), while TypiClust selects cluster-typical points (Hacohen et al., 2022). SACA instead reduces support deficits in prespecified query-centered regions relative to their sufficient reporting counts. We compare these coverage methods in Section 5.

Uncertainty-based selection and boundary queries. Uncertainty-based selection depends on a predictive model (Ash et al., 2020; Bae et al., 2025). SACA scores candidate locations without using their unknown outcomes: either label adds an observation to every region containing the candidate. Once revealed, that label affects estimated prevalence and empirical-Bernstein width.

Black-box boundary queries (Guerrero & Rojas, 2026) also allocate a limited budget, but obtain model evaluations near a decision boundary rather than labels for pooled cases used in regional evidence reporting.