DiscoCATE: Heterogeneous Treatment Effect Estimation with Unit-Abducted Neighborhoods

Anonymous Author(s)

Working paper, July 18, 2026
Paper version: DISC-CATE-WP-2026-07-17-V1

Abstract

Estimating heterogeneous treatment effects is fundamentally a localization problem: which observed units should inform a causal query at a focal covariate value? Causal forests answer this question by learning adaptive neighborhoods in observed covariate space. We study a complementary localization principle based on *Unit Abduction*. A frozen abductor maps pretreatment evidence X_i to a probability measure $Q_i = \mathcal{Q}(X_i)$, interpreted as learner uncertainty over response-relevant types of the same focal unit. DiscoCATE compares complete belief objects Q_i and Q_j, converts their proximity into weights on observed donor rows, and averages cross-fitted augmented inverse-propensity scores.

The distinction between the resulting target and ordinary CATE is explicit. For a fixed belief map, distance, kernel, and bandwidth, the method targets the soft-population effect

$$\theta_{h,\mathcal{Q}}(q) = \frac{\mathbb{E}[K_h\{d(q, \mathcal{Q}(X))\}\{Y(1) - Y(0)\}]}{\mathbb{E}[K_h\{d(q, \mathcal{Q}(X))\}]}.$$

Under standard consistency, conditional exchangeability, positivity, and integrability assumptions, this estimand is identified from the observed law without identifying the true latent unit or the latent decomposition. As the neighborhood shrinks, it approaches the effect conditional on the information retained by Q_X; equality to focal CATE requires an additional effect-sufficiency or fiber condition. This separation yields a conservative identification theory, an honest doubly robust estimator, and a falsification protocol for testing whether full belief geometry improves localization relative to raw covariates, hard subgroups, or posterior point embeddings.

1 Introduction

Conditional average treatment-effect estimation is usually introduced as a regression problem. Given pretreatment covariates $X = x$, binary treatment T, and potential outcomes $Y(0), Y(1)$, the target is

$$\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\}.$$

In practice, however, every nonparametric estimator of $\tau(x)$ must answer a more primitive question: *which observed units should be allowed to speak for the focal query x, and with what weight?* Nearest-neighbor methods answer with a prespecified metric. Kernel methods answer with a bandwidth. Causal trees and forests learn an adaptive partition, and modern generalized random forests express the result as a data-adaptive weighting measure over training observations [2, 14]. Heterogeneous treatment-effect estimation is therefore also a problem of constructing a causally admissible local population.

This localization view explains much of the appeal of causal forests. A forest does not estimate a focal effect by treating the entire sample as equally relevant. Across trees, co-occurrence in terminal leaves defines an adaptive observed-unit neighborhood. The neighborhood is shaped by covariates and by a splitting objective designed to expose treatment-effect heterogeneity; local moment estimation then turns those weights into an effect estimate. Honesty and sample splitting separate neighborhood construction from within-leaf estimation, supporting consistency and inference under stated regularity conditions [2, 14]. This is a substantive statistical theory, not merely a metaphor for similarity.

Yet observed covariate geometry and response-relevant geometry need not be the same object. The vector X may simultaneously contain confounding, prognostic, administrative, and weak proxy information. Euclidean proximity is rarely meaningful without learning; even an adaptive partition represents locality through observed rows and point-valued covariate tests. In settings with weak or ambiguous evidence about response type, a focal record may be compatible with several qualitatively different mechanisms. Compressing that ambiguity into a single point representation or hard subgroup can erase precisely the uncertainty that should govern borrowing across units. Representation-based and latent-variable treatment-effect methods address parts of this problem [1, 7, 9, 13], but they do not by themselves impose the unit-belief semantics studied here.

Unit Abduction as a localization object. Distribution-Consistency Structural Causal Models (DiscoSCMs) distinguish a stable unit selector from event-level randomness [4]. For the treatment-effect slice considered in this paper, write

$$Y(t) = f_0(X, t, E_t; U), \qquad t \in \{0, 1\}, \tag{1.1}$$

where U is the single model-level Unit Selection Variable. Record i corresponds to the event $U = u_i$; the row index does not create a new latent random variable. The learner does not observe u_i. From admissible pretreatment evidence $O_i := X_i$, a Unit Abductor instead returns a belief

$$Q_i(du) = \mathcal{A}_\phi(X_i)(du) \approx P_{\text{tar}}(U \in du \mid X = x_i). \tag{1.2}$$

The approximation is a semantic calibration target, not something guaranteed by predictive fit. Draws from Q_i are candidate hypotheses about the same fixed focal unit; they are neither multiple actual units nor repeated realizations of its identity.

Equation (1.2) suggests a different geometry for causal localization. If Q_i and Q_j place similar mass on response-relevant candidate types, row j may be a useful donor for focal row i, even when the raw covariates are not close in a generic metric. Conversely, covariate-near records whose beliefs concentrate on different response types may deserve less influence. The proposed computation is

$$X_i \longrightarrow \widehat{Q}_i \longrightarrow d(\widehat{Q}_i, \widehat{Q}_j) \longrightarrow \widehat{\alpha}_{ij} \longrightarrow \widehat{\tau}_i = \sum_j \widehat{\alpha}_{ij} \widehat{\Gamma}_j, \tag{1.3}$$

where d compares probability measures, $\widehat{\alpha}_{ij}$ weights *observed donor rows*, and $\widehat{\Gamma}_j$ is an orthogonal causal score. We call the resulting method DiscoCATE, and its first doubly robust instantiation DiscoCATE–DR.

Because Q_X is constructed from pretreatment X, it does not enlarge the individual information set available at prediction time. Any gain must come from structurally useful regularization, externally learned mechanism information, or a better finite-sample bias–variance tradeoff—not from a stronger nonparametric identification result. The purpose of a complete belief is to preserve response-relevant ambiguity while reshaping borrowing, not to manufacture information absent from the evidence.

The sample-space distinction in (1.3) is essential. Q_i is a measure over candidate values of the latent selector U; α_{ij} is a measure over observed training records. One cannot literally replace causal-forest weights by Q_i. The bridge requires a similarity rule that takes two belief objects as input and returns an application-time weight on a donor row. This paper studies that bridge.

What effect is being estimated? A new geometry does not automatically preserve the old estimand. Let $\mathcal{Q} : x \mapsto Q_x$ be fixed, externally trained, or honestly cross-fitted; let d be a measurable distance between beliefs; and let K_h be a nonnegative localization kernel. For focal belief q, the population target naturally induced by these choices is

$$\theta_{h,\mathcal{Q}}(q) = \frac{\mathbb{E}[K_h\{d(q, \mathcal{Q}(X))\}\{Y(1) - Y(0)\}]}{\mathbb{E}[K_h\{d(q, \mathcal{Q}(X))\}]}. \tag{1.4}$$

This is the average treatment effect in a pretreatment-defined soft subpopulation. It is not the realized individual treatment effect of the physical focal token. It is also not automatically $\tau(x_i)$.

This distinction makes the identification argument simpler and more honest. Once $\mathcal{Q}$, d, K_h, and q are frozen before donor outcomes are used, the weight in (1.4) is a measurable function of pretreatment X. Under the same consistency, conditional exchangeability, treatment positivity, and moment assumptions used for ordinary CATE, $\theta_{h,\mathcal{Q}}(q)$ is an observed-law functional. Unit Abduction chooses which soft population is emphasized; it does not replace the causal assumptions needed to identify effects inside that population.

As h shrinks under an approximate-identity condition,

$$\theta_{h,\mathcal{Q}}(Q_{x_i}) \longrightarrow \mathbb{E}\{\tau(X) \mid Q_X = Q_{x_i}\}. \tag{1.5}$$

The limit is the effect conditional on the information retained by the belief map. It equals focal CATE when, for example,

$$\tau(X) = g(Q_X) \quad \text{almost surely}, \tag{1.6}$$

or when a corresponding local injectivity or fiber-equality condition holds. Without such a condition, the method still targets an identified Q-localized effect, but that effect is a coarsening of CATE. We regard this two-level statement—identification of the localized target first, equality to focal CATE second—as the correct answer to the method's identifiability question.

An honest estimator. We estimate (1.4) by averaging cross-fitted augmented inverse-propensity weighted scores. Outcome regressions and the propensity score use the full X needed for confounding adjustment. The Unit Abductor is trained externally or in an outer fold, and application-time donor weights depend only on pretreatment information. Thus the proposed geometry is not asked to do the job of the propensity score, and double robustness is not misstated as robustness to a wrong causal model or an uncalibrated abductor.

This restriction also clarifies a tempting alternative. Given a focal Q_i, one could score donor j by the likelihood of its realized (T_j, Y_j) under candidate values $U \sim Q_i$. Such a score may be useful for an ex-post matching or posterior-predictive task. In the ex-ante CATE problem, however, using Y_j both to select donor j and as part of its causal-score contribution creates outcome-dependent selection. Cross-fitting the likelihood model does not by itself restore the original estimand. DiscoCATE–DR therefore uses pretreatment-only weights; outcome-conditioned similarities are outside the scope of the first algorithm.

Contributions and claim boundary. The paper makes four contributions.

1. We formulate *unit-abducted neighborhoods*: full evidence-indexed beliefs over response-relevant unit types are compared and then converted into localization weights on observed donor records.
2. We define the finite-bandwidth target $\theta_{h,\mathcal{Q}}(q)$, identify it under standard causal assumptions for a fixed pretreatment belief map, and isolate the additional conditions needed for equality to focal CATE.
3. We give an honest estimator, DiscoCATE–DR, that combines these weights with cross-fitted augmented inverse-propensity scores while keeping abduction, treatment assignment, and confounding adjustment distinct.
4. We specify a falsification-oriented evaluation design that separates oracle belief geometry, learned-abductor error, multimodal uncertainty, treatment overlap, and outcome leakage, with matched raw-X, point-belief, subgroup, DR-learner, and causal-forest baselines.

The proposal is deliberately narrower than a general latent causal model. We do not claim that one-row observational data identify the true u_i, the calibrated posterior $P(U \mid X)$, a latent type-specific effect surface, or the realized distribution of individual effects. We do not inherit causal forest asymptotic normality by changing the geometry, and we report no empirical superiority result in this version. The scientific hypothesis is instead testable and specific: when pretreatment evidence leaves response-relevant type uncertainty that is multimodal or otherwise poorly summarized by a point, does the *complete belief measure* define a more useful causal neighborhood than observed-covariate proximity or simpler belief compressions?

Section 2 fixes the causal and unit semantics. Section 3 gives DiscoCATE–DR. Section 4 separates localized-effect identification from focal-CATE recovery. Section 5 states the empirical falsification protocol. Appendix A develops the DiscoSCM and Unit Abduction background in full.

2 Setup and target

2.1 Observed data and causal assumptions

Let

$$\mathcal{D}_n = \{Z_i = (X_i, T_i, Y_i)\}_{i=1}^n$$

be independent draws from a declared target population, with binary treatment $T_i \in \{0, 1\}$. The vector X_i contains only pretreatment information. Potential outcomes are $Y_i(0)$ and $Y_i(1)$, and the observed outcome is $Y_i = Y_i(T_i)$.

Assumption 1 (Causal identification)**.** *For $t \in \{0, 1\}$:*

(i) consistency and no interference hold;
(ii) $\{Y(0), Y(1)\} \perp T \mid X$*;*
(iii) $0 < e_0(X) < 1$ *almost surely, where* $e_0(x) = \mathbb{P}(T = 1 \mid X = x)$*; and*
(iv) $\mathbb{E}|Y(t)| < \infty$*.*

Write

$$\mu_{0,t}(x) = \mathbb{E}(Y \mid T = t, X = x).$$

Under Assumption 1, ordinary CATE is identified as

$$\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\} = \mu_{0,1}(x) - \mu_{0,0}(x). \tag{2.1}$$

Unit Abduction does not weaken these assumptions. In particular, it does not repair unmeasured confounding or lack of treatment overlap.

2.2 One unit selector and one focal belief

The DiscoSCM response slice is

$$Y(t) = f_0(X, t, E_t; U), \tag{2.2}$$

where $U : \Omega \to \mathcal{U}$ is one model-level Unit Selection Variable and E_t is event-level randomness. For row i, the actual world-side selection is expressed by the event $U = u_i$. We reserve subscripts for observations, realized values, and belief objects; we do not posit a separate model-level unit variable for each row.

The application-time evidence is

$$O_i := X_i, \qquad T_i, Y_i \notin O_i.$$

An abductor is a measurable map

$$\mathcal{Q} : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{U}), \qquad x \longmapsto Q_x, \tag{2.3}$$

where $\mathcal{P}(\mathcal{U})$ is a space of probability measures. The ideal semantic target, relative to a declared target population, is

$$Q_x^\star(du) = P_{\text{tar}}(U \in du \mid X = x).$$

We distinguish this oracle belief from a learned map $\widehat{\mathcal{Q}}$. Predictive fit of a latent response model need not identify or calibrate $Q_x^\star$.

Assumption 2 (Admissible belief map)**.** *For the fixed-map analysis, $\mathcal{Q}$ is predetermined, trained on external data, or produced in an outer sample split and then frozen. At application time, Q_x is a function of pretreatment x only. When $\widehat{\mathcal{Q}}$ is learned using treatment or outcome supervision, the belief assigned to a row is computed by a model that did not train on that row's realized treatment or outcome.*

Because $Q_X = \mathcal{Q}(X)$ is a function of X, $\sigma(Q_X) \subseteq \sigma(X)$. Unit Abduction therefore cannot create new individual-level information at prediction time. Its possible value is as a mechanism-informed inductive bias: external training, structural restrictions, and a distribution-valued representation can reshape finite-sample localization while retaining multimodality and epistemic uncertainty.

2.3 Belief geometry and soft populations

Let $d : \mathcal{P}(\mathcal{U})^2 \to [0, \infty]$ be measurable and let $K_h(r) = K(r/h)$ for a nonnegative kernel K. If beliefs admit densities p, q with respect to a common dominating measure, two bounded candidates are the Hellinger distance

$$H^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\nu$$

and the square root of the Jensen–Shannon divergence. Symmetrized Kullback–Leibler divergence is a useful sensitivity analysis but can be infinite under support mismatch. The method does not require one distance to be ontologically privileged.

For focal $q \in \mathcal{P}(\mathcal{U})$, define

$$w_{q,h}(X) = K_h\{d(q, Q_X)\}, \qquad 0 < \mathbb{E}\{w_{q,h}(X)\} < \infty. \tag{2.4}$$

The normalized reweighting of the observed population induced by (2.4) is the *belief-localized soft population*. Its treatment effect is

$$\theta_{h,\mathcal{Q}}(q) = \frac{\mathbb{E}[w_{q,h}(X)\{Y(1) - Y(0)\}]}{\mathbb{E}\{w_{q,h}(X)\}}. \tag{2.5}$$

whenever the weighted potential outcomes are integrable. The target depends on the declared population, belief map, distance, kernel, and bandwidth. Those choices must therefore be reported as part of the estimand rather than treated as implementation details.

3 DiscoCATE–DR

3.1 Orthogonal donor signals

Let $\eta_0 = (e_0, \mu_{0,0}, \mu_{0,1})$ collect the propensity score and outcome regressions. For a generic nuisance tuple $\eta = (e, \mu_0, \mu_1)$, define the augmented inverse-propensity score

$$\Gamma(Z;\eta) = \mu_1(X) - \mu_0(X) + \frac{T}{e(X)}\{Y - \mu_1(X)\} - \frac{1-T}{1-e(X)}\{Y - \mu_0(X)\}. \tag{3.1}$$

This is the standard doubly robust pseudo-outcome construction [3, 6]. At the true nuisance functions,

$$\mathbb{E}\{\Gamma(Z;\eta_0) \mid X\} = \tau(X). \tag{3.2}$$

We estimate the nuisances by cross-fitting on the full covariate vector X. For donor row j, let $\widehat{\Gamma}_j = \Gamma(Z_j; \widehat{\eta}^{(-k(j))})$, where the nuisance model used for row j was trained without its fold.

The full X remains in the nuisance functions even when localization is performed through Q_X. This is deliberate: the belief geometry determines which rows are locally relevant, whereas the nuisance functions adjust for assignment and outcome structure. A coarsened Q_X is not assumed to be a balancing score.

3.2 From belief distance to observed-row weights

For focal row i, let $\mathcal{I}_i^{\text{donor}}$ be an honest donor set. Compute out-of-fold beliefs

$$\widehat{Q}_i = \widehat{\mathcal{Q}}^{(-a(i))}(X_i), \qquad \widehat{Q}_j = \widehat{\mathcal{Q}}^{(-a(j))}(X_j).$$

The unnormalized and normalized donor weights are

$$\widehat{w}_{ij}^{(h)} = K_h\{d(\widehat{Q}_i, \widehat{Q}_j)\}\mathbf{1}\{j \in \mathcal{I}_i^{\text{donor}}\}, \tag{3.3}$$

$$\widehat{\alpha}_{ij}^{(h)} = \frac{\widehat{w}_{ij}^{(h)}}{\sum_{\ell \in \mathcal{I}_i^{\text{donor}}} \widehat{w}_{i\ell}^{(h)}}. \tag{3.4}$$

The first estimator in the proposed family is

$$\boxed{\widehat{\tau}_{\text{DiscoCATE},h}(x_i) = \sum_{j \in \mathcal{I}_i^{\text{donor}}} \widehat{\alpha}_{ij}^{(h)} \widehat{\Gamma}_j}. \tag{3.5}$$

Despite the notation $\widehat{\tau}(x_i)$, the exact finite-bandwidth target of (3.5) is $\theta_{h,\mathcal{Q}}(Q_{x_i})$. We use the CATE notation only when the conditions in Section 4 justify that interpretation.

For an arbitrary query belief q, the same construction replaces $\widehat{Q}_i$ by q. The effective donor sample size

$$n_{\text{eff}}(q) = \frac{1}{\sum_j \widehat{\alpha}_j(q)^2}$$

and the maximum donor weight should be reported with every estimate. They make bandwidth collapse and distance concentration visible.

3.3 Algorithmic contract

The first implementation follows five steps.

1. Split rows into outer abductor folds and causal nuisance folds.
2. Train the Unit Abductor in the outer folds; at application time, map each row's pretreatment X to a frozen belief $\widehat{Q}$.
3. Cross-fit $e(X)$, $\mu_0(X)$, and $\mu_1(X)$, and construct $\widehat{\Gamma}_j$ for each donor.
4. For each focal query, compare its complete belief with donor beliefs, apply K_h, and normalize over an honest donor set.
5. Average the donor scores as in (3.5); report overlap, belief-distance, and effective-neighborhood diagnostics.

This protocol allows the abductor to learn response-relevant structure from other rows while preventing a donor's realized outcome from directly determining its own application-time weight. A strictly external abductor is the cleanest first analysis. A jointly learned abductor requires an additional stability theory, discussed in Section 4.

3.4 Similarity choices and the outcome-likelihood boundary

The default method compares the complete probability measures $\widehat{Q}_i$ and $\widehat{Q}_j$. Point-mass distance between posterior means, hard maximum-a-posteriori class membership, and raw-X kernels are important ablations rather than equivalent implementations.

An asymmetric compatibility score based only on pretreatment evidence can also be admissible. For example, if a fixed generative evidence model $p_\psi(x \mid u)$ is available,

$$s^{X}_{i\to j} = \int p_\psi(X_j \mid u)\, \widehat{Q}_i(du)$$

asks whether donor j's evidence is compatible with focal belief i. This score defines a different geometry and requires its own normalization and stability checks.

By contrast, the posterior-predictive score

$$s^{Y,T}_{i\to j} = \int p_\theta(Y_j, T_j \mid X_j, u)\, \widehat{Q}_i(du)$$

uses the realized treatment or outcome of donor j. If that same row then contributes $\widehat{\Gamma}_j$, the target population has been selected using the realized causal signal. Cross-fitting p_θ does not remove this selection. Such scores may define meaningful ex-post matching estimands, but they are excluded from the ex-ante DiscoCATE–DR estimator.

4 Identification and estimand boundaries

The central identification question has two levels. First, is the belief-localized soft-population effect an observed-law functional? Second, when does that effect coincide with the focal CATE? The answers require different assumptions.

4.1 Identification of the fixed-map localized effect

Theorem 1 (Fixed-$\mathcal{Q}$ identification)**.** *Suppose Assumptions 1 and 2 hold, and let $w_{q,h}(X)$ satisfy (2.4) and*

$$\mathbb{E}[w_{q,h}(X)\{|Y(0)| + |Y(1)|\}] < \infty.$$

Then

$$\theta_{h,\mathcal{Q}}(q) = \frac{\mathbb{E}[w_{q,h}(X)\{\mu_{0,1}(X) - \mu_{0,0}(X)\}]}{\mathbb{E}\{w_{q,h}(X)\}} \tag{4.1}$$

$$= \frac{\mathbb{E}[w_{q,h}(X)\Gamma(Z;\eta_0)]}{\mathbb{E}\{w_{q,h}(X)\}}. \tag{4.2}$$

Consequently, for a fixed pretreatment belief map and localization rule, $\theta_{h,\mathcal{Q}}(q)$ is identified by the observed distribution of (X, T, Y).

The proof is an application of conditional exchangeability and iterated expectation and is given in Appendix B.1. The theorem does not require that Q_x equal the true latent posterior. It identifies the effect in the soft population induced by the declared map. Calibration to $P_{\text{tar}}(U \mid X = x)$ is needed for the stronger response-relevant-type interpretation, not for the observed-law equality in (4.1).

Theorem 1 also clarifies the role of double robustness. With cross-fitting and suitable nuisance convergence, the weighted empirical average of $\widehat{\Gamma}$ can consistently estimate (4.2) if one of the propensity or outcome nuisance components is correctly learned under the usual conditions. This robustness concerns (e, μ_0, μ_1); it does not protect against a changing estimand caused by unstable or outcome-dependent belief weights.

4.2 The shrinking neighborhood is a coarsening of CATE

Let $R = Q_X$ be a random element of the metric space $(\mathcal{P}(\mathcal{U}), d)$, and let

$$m(q) = \mathbb{E}\{\tau(X) \mid R = q\}$$

denote a version of the conditional effect given the belief representation.

Assumption 3 (Approximate identity at q)**.** *For the focal q, local mass is positive for all sufficiently small h, and the normalized kernels satisfy*

$$\frac{\mathbb{E}[K_h\{d(q,R)\}m(R)]}{\mathbb{E}[K_h\{d(q,R)\}]} \longrightarrow m(q).$$

This condition holds under standard kernel, support, and continuity conditions adapted to the distribution of R.

Proposition 1 (Belief-fiber limit)**.** *Under the conditions of Theorem 1 and Assumption 3,*

$$\theta_{h,\mathcal{Q}}(q) \longrightarrow \mathbb{E}\{\tau(X) \mid Q_X = q\}.$$

Because Q_X is a function of X, the right-hand side averages $\tau(X)$ across the fiber

$$\{x : \mathcal{Q}(x) = q\}.$$

Thus a perfectly estimated Q-localized effect can still differ from the focal $\tau(x)$. This is not estimator bias; it is the estimand implied by a coarsened representation.

Corollary 1 (Recovery of focal CATE)**.** *Suppose there is a measurable function g such that*

$$\tau(X) = g(Q_X) \quad \textit{almost surely.}$$

Then the limit in Proposition 1 is $g(q)$. In particular, for any focal x with $Q_x = q$,

$$\lim_{h \to 0} \theta_{h,\mathcal{Q}}(Q_x) = \tau(x)$$

at regular points. The same conclusion holds under the weaker focal fiber-equality condition $\mathbb{E}\{\tau(X) \mid Q_X = Q_x\} = \tau(x)$.

Injectivity of $x \mapsto Q_x$ on the relevant support is another sufficient route, but is stronger than necessary. For a fixed nonzero bandwidth, exact equality to focal CATE holds if and only if

$$\mathbb{E}[w_{Q_x,h}(X)\{\tau(X) - \tau(x)\}] = 0. \tag{4.3}$$

Accordingly, empirical evaluation must score finite-bandwidth estimates against $\theta_{h,\mathcal{Q}}$, and should report focal-CATE error only in regimes where the conditions of Corollary 1 are known to hold.

4.3 What is not identified

The observed-law identification of $\theta_{h,\mathcal{Q}}$ must not be confused with identification of the latent decomposition. A predictive response model has the form

$$p(dy \mid x, t) = \int_{\mathcal{U}} K_\theta(dy \mid x, t, u)\, Q_x(du). \tag{4.4}$$

Without additional restrictions, the observable mixture in (4.4) does not separately identify Q_x, the response kernel, the actual u_i, or a preferred latent coordinate system. For example, every observable conditional law admits both a degenerate latent representation and a representation with an auxiliary latent variable that the kernel ignores. Bijective reparameterizations of U and corresponding pushforwards of Q_x leave the mixture unchanged. Appendix A.5 gives the formal statements.

The correct identification ladder is therefore:

1. Under standard causal assumptions, ordinary $\tau(x)$ is identified from the observed law.
2. Conditional on a fixed pretreatment map $\mathcal{Q}$, the finite-bandwidth soft-population effect $\theta_{h,\mathcal{Q}}(q)$ is identified.
3. Under localization regularity, shrinking neighborhoods recover $\mathbb{E}\{\tau(X) \mid Q_X = q\}$.
4. Effect-sufficiency, injectivity, or fiber equality upgrades that limit to focal CATE.
5. Calibration or identification of the true latent belief $Q_x^\star$ requires additional information beyond these causal identification assumptions.

4.4 Estimated beliefs

When $\mathcal{Q}$ is learned, outer-fold honesty makes each application-time weight pretreatment-measurable but does not by itself make weight error small. A consistency theorem for DiscoCATE–DR with learned beliefs additionally requires stability of

$$K_h\{d(\widehat{Q}_x, \widehat{Q}_{x'})\}$$

relative to its population limit, nonvanishing local mass, and a growing effective donor neighborhood. The rate generally worsens as h shrinks. Neyman orthogonality of the AIPW score with respect to outcome and propensity nuisances does not imply orthogonality with respect to $\widehat{Q}$.

5 Evaluation design and falsification criteria

This version states an evaluation contract rather than reporting benchmark results. The primary empirical question is not whether a new model can obtain a favorable PEHE number under unmatched tuning. It is whether the *distribution-valued belief* improves causal localization beyond raw covariates and simpler compressions when all methods target the same effect and use the same admissible information.

5.1 Required regimes

Regime	Purpose
E0	Null heterogeneity or a setting in which Q_X contains no useful effect geometry. The method must not manufacture an advantage.
E1	Oracle Q with known response types. This isolates the weighting principle from abductor estimation error.
E2	Weak, multimodal pretreatment evidence with overlapping raw-X regions and opposing type-specific effects. This is the central regime in which a complete belief may outperform its posterior mean or hard class.
E3	Controlled belief misspecification: overconfidence, excessive dispersion, missing modes, and distance distortion.
E4	Randomized assignment, observed confounding with adequate overlap, weak overlap, and hidden confounding. Hidden confounding is a negative control, not a success regime.
E5	Deliberate treatment or outcome leakage into donor similarity. This tests whether apparent gains disappear under the honest pretreatment-only contract.

The oracle data-generating process must expose both the finite-bandwidth target $\theta_{h,\mathcal{Q}}$ and focal $\tau(x)$. E2 should include at least one case in which two beliefs have equal posterior means but different modal mass, so that full-belief and point-summary methods are genuinely distinguishable.

5.2 Matched baselines

At minimum, the comparison includes:
1. causal forest or generalized random forest;
2. a standard DR-learner and an R-learner using X;
3. an X-space kernel with matched effective sample size;
4. hard latent subgroup assignment;
5. distance between posterior point or mean embeddings;
6. the full-Q DiscoCATE weight; and

7. oracle-Q weighting when simulation permits it.

All causal learners should share nuisance folds where appropriate and must use the same pretreatment evidence. Bandwidths should be compared both at their best validation choice and at matched effective neighborhood size. A method cannot claim an improvement by silently changing the target population.

5.3 Metrics and failure conditions

The primary metric is error for the finite-bandwidth target actually induced by each localization rule. Focal-CATE error is secondary and is used only when effect-sufficiency or the relevant fiber equality holds. Required diagnostics include treatment overlap, propensity extremes, belief calibration when oracle truth exists, distance concentration, maximum donor weight, and effective neighborhood size. Policy value and effect ranking can be reported as secondary decision metrics.

The central hypothesis is not supported if the full-Q method fails to improve over point-belief or raw-X localization under repeated seeds and matched tuning; if gains vanish after matching effective sample size; if gains require realized-outcome leakage; if benign latent reparameterizations destabilize the distance; or if success requires evaluating a coarsened Q-effect as though it were focal CATE. Interval coverage is not a valid metric until an inference theorem for learned-Q weights is available.

6 Related work and positioning

Adaptive causal neighborhoods. Causal forests use honest tree ensembles to estimate heterogeneous treatment effects and support pointwise inference under unconfoundedness and regularity conditions [14]. Generalized random forests cast forest weights as adaptive localization for conditional moment equations [2]. These methods are the closest computational analogue for our starting intuition. We do not treat causal forest as a special case of Unit Abduction, nor claim that its theory transfers. Generalized-random-forest weights index observed training records; Q_x indexes candidate latent response types for one focal token. DiscoCATE composes these levels by using belief distances to induce a new set of observed-row weights.

Orthogonal and meta-learners. Meta-learners combine flexible regression algorithms to estimate heterogeneous effects [8]. The R-learner residualizes treatment and outcomes and admits quasi-oracle guarantees in studied regimes [10]; doubly robust learners use orthogonal pseudo-outcomes to reduce sensitivity to nuisance estimation [6]. DiscoCATE–DR uses an AIPW signal from this tradition. Its proposed novelty is not a new orthogonal score, but the pretreatment belief geometry used to localize that score.

Representation learning and Bayesian HTE. Balanced representations seek predictive features that support counterfactual generalization [1, 13]. Deep latent-variable methods model hidden confounding or potential outcomes through learned latent structure [9], while Bayesian causal forests regularize prognostic and treatment-effect components and quantify posterior uncertainty [5]. These approaches demonstrate that learned structure, balancing, and uncertainty can matter for HTE, but their latent variables, posteriors, and weights carry different semantics from a query-invariant belief about a fixed unit selector.

Latent subgroups. Latent class and latent-variable models explicitly estimate subgroup-specific treatment effects [7, 15]. DiscoCATE differs from hard subgrouping by retaining a complete focal belief and by defining a continuous soft neighborhood between observed rows. Nevertheless, soft membership and posterior similarity have substantial prior art. We therefore do not claim the first latent, probabilistic, or soft-subgroup HTE method.

DiscoSCM and counterfactual abduction. Structural causal models use abduction, action, and prediction to answer counterfactual queries [12]; deep structural causal models implement this pipeline with flexible mechanisms [11]. DiscoSCM introduces distribution-consistency and a unit-selection perspective for heterogeneous counterfactual modeling [4]. The present work uses only a bounded interface from that theory: one fixed unit selector, an epistemic belief formed from pretreatment evidence, and a requirement that the belief remain fixed when the query changes. It does not claim to identify cross-world coupling or individual treatment-effect distributions.

The narrow contribution is thus a composition rule:

$$\text{query-invariant unit belief} \longrightarrow \text{pretreatment donor localization} \longrightarrow \text{orthogonal causal estimation.}$$

Whether this composition yields a practical advantage is an empirical hypothesis governed by the protocol in Section 5.

7 Discussion and conclusion

DiscoCATE begins from a simple shared intuition with causal forests: individualized causal estimation requires a focal population, not indiscriminate use of the full sample. The difference is where localization is represented. Forests learn a neighborhood among observed records through covariate splits. Unit Abduction represents uncertainty over response-relevant candidate types; distances between those beliefs are then pulled back to weights on observed donors.

The identification result is intentionally modest. For a frozen, pretreatment-measurable belief map, the resulting finite-bandwidth soft-population effect is identified under standard causal assumptions. The method does not need to recover the true latent unit to make that target well-defined. At the same time, the target equals focal CATE only if the belief representation preserves the effect-relevant distinctions in X. This effect-sufficiency requirement prevents a useful structural representation from being mistaken for additional causal information.

Several limitations are immediate. First, one-row observational prediction does not generally identify the calibrated latent belief or response kernel. External mechanism knowledge, repeated measurements, multi-environment data, or stronger identifiable component restrictions may be needed. Second, cross-fitting makes the learned weights honest but does not guarantee their stability; rates and inference require a separate learned-Q analysis. Third, the choice of probability-measure distance can dominate performance in high-dimensional or weak-overlap belief spaces. Fourth, mean CATE requires integrable potential outcomes. A nondegenerate Cauchy response specialization must target a location, median, quantile, or another well-defined functional rather than silently invoking an undefined mean.

The most important next test is therefore controlled rather than broad. In an oracle simulation with multimodal weak evidence, compare full-belief localization against raw X, posterior means, hard classes, causal forest, and standard orthogonal learners at the same estimand and effective neighborhood size. A positive result would justify investing in learned-belief stability theory and scalable estimators. A negative result would be equally informative: it would show that

Unit Abduction is valuable as a response-model interface without necessarily improving CATE localization.

The broader research route is clear but not yet claimed as complete. Causal forests show how adaptive observed-unit neighborhoods can support rigorous HTE estimation. Unit Abduction supplies a distribution-valued, epistemic geometry for the focal unit. DiscoCATE asks whether these ideas can be combined without blurring their sample spaces, causal assumptions, or uncertainty semantics. The present formulation makes that question mathematically specific and empirically falsifiable.

References

[1] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1972–1980. PMLR, 2021. URL https://proceedings.mlr.press/v130/assaad21a.html.

[2] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

[3] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

[4] Heyang Gong, Chaochao Lu, and Yu Zhang. Distribution-consistency structural causal models. *arXiv preprint arXiv:2401.15911*, 2024. URL https://arxiv.org/abs/2401.15911.

[5] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020. doi: 10.1214/19-BA1195.

[6] Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023. doi: 10.1214/23-EJS2157.

[7] Hang J. Kim, Bo Lu, Edward J. Nehus, and Mi-Ok Kim. Estimating heterogeneous treatment effects for latent subgroups in observational studies. *Statistics in Medicine*, 38(3):339–353, 2019. doi: 10.1002/sim.7970.

[8] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

[9] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/94b5bde6de888ddf9cde6748ad2523d1-Abstract.html.

[10] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. doi: 10.1093/biomet/asaa076.

[11] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*,

volume 33, pages 857–869, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

[12] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[13] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

[14] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

[15] Yanyan Yin, Lan Liu, and Zhi Geng. Assessing the treatment effect heterogeneity with a latent variable. *Statistica Sinica*, 28:115–135, 2018. doi: 10.5705/ss.202016.0150.

A DiscoSCM and Unit Abduction background

This appendix fixes the unit semantics used throughout the paper. The key distinction is between a world-side unit realization and a learner-side distribution over candidate unit types. DiscoSCM contains one model-level Unit Selection Variable U, not a separate latent random variable for each row.

A.1 World-side unit semantics

Let $(\Omega, \mathcal{F}, P_{\text{tar}})$ denote a target-population probability space, and let

$$U : \Omega \longrightarrow \mathcal{U}$$

be the model-level Unit Selection Variable. The space $\mathcal{U}$ contains possible response-relevant unit values. Record i corresponds to the realization u_i, represented by the event

$$U = u_i.$$

The index i labels a realized record; it does not create a new model-level random variable.

We consider binary treatment $T \in \{0, 1\}$, pretreatment covariates X, and potential outcomes generated by

$$Y(t) = f_0(X, t, E_t; U), \qquad t \in \{0, 1\}. \tag{A.1}$$

For row i, this equation is instantiated as

$$Y_i(t) = f_0(X_i, t, E_{i,t}; u_i), \qquad Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

The mean-effect targets in this paper depend on the marginal potential-outcome laws. Assuming a shared event-noise realization across treatment worlds would add a cross-world coupling that is not required for those targets.

The word *selection* must therefore be qualified. The following objects play distinct roles:

Object	Role
U	Selects the actual unit realization queried by the response mechanism. Its realization is fixed across same-token treatment queries.
P_U^{tar}	Describes the composition of unit realizations in the declared target population.
T	Records or generates treatment assignment.
S^{obs}	Indicates inclusion in the observed sample.

Neither treatment assignment nor sample inclusion is represented by U. Candidate-belief support, treatment overlap, and sample-inclusion positivity are consequently three different support conditions.

A.2 Reference-population beliefs

In the ex-ante CATE setting of this paper, the admissible factual evidence is

$$O_i := X_i, \qquad T_i, Y_i \notin O_i.$$

The target-population oracle belief for a focal record with $X_i = x_i$ is the probability measure

$$Q_{x_i}^{\star,\mathrm{tar}}(du) := P_{\mathrm{tar}}(U \in du \mid X = x_i). \tag{A.2}$$

A learned Unit Abductor returns

$$\widehat{Q}_i = \mathcal{A}_{\widehat{\phi}}(X_i) \approx Q_{x_i}^{\star,\mathrm{tar}}.$$

When the distinction between oracle and learned objects is immaterial, we write Q_i. The subscript belongs to the evidence-indexed belief, not to the model-level Unit Selection Variable.

The conditional distribution in (A.2) is epistemic: it represents uncertainty about the fixed realization u_i. A draw

$$\widetilde{u}_i^{(m)} \sim Q_i$$

is a candidate hypothesis about that realization, not another actual unit and not the unknown truth u_i. A nondegenerate Q_i therefore does not imply that the focal record is physically composed of multiple units.

Reference law before abduction. The target population must be declared before abduction. If the training sample is selected, the directly learned law may instead be

$$Q_i^{\mathrm{obs}}(du) \approx P_{\mathrm{obs}}\big(U \in du \mid X = x_i, S^{\mathrm{obs}} = 1\big),$$

which need not equal $Q_{x_i}^{\star,\mathrm{tar}}$. Transport to the target population requires separate assumptions about selection, stability, and inclusion support.

The phrase *soft candidate subpopulation* has a precise but limited meaning. If $Q_i^{\mathrm{tar}} \ll P_U^{\mathrm{tar}}$ and there is a nonnegative evidence-compatibility function r_i such that

$$Q_i^{\mathrm{tar}}(du) = \frac{r_i(u) P_U^{\mathrm{tar}}(du)}{\int_{\mathcal{U}} r_i(v) P_U^{\mathrm{tar}}(dv)}, \tag{A.3}$$

then Q_i^{tar} can be viewed as a soft reweighting of target-population candidate types. Equation (A.3) still represents uncertainty about one fixed focal unit; it is not an ontic population of multiple focal units.

Finally, conditioning on X does not assert a world-side causal arrow $X \to U$. The expression $P(U \mid X = x)$ specifies a conditional belief. Any causal relationship between X and U requires separate structural assumptions.

A.3 Abduction once, reduction across queries

Let $K_\theta(dy \mid x, t, u)$ denote a response kernel. Unit Abduction first forms a belief from factual evidence:

$$\widehat{Q}_i = \mathcal{A}_{\widehat{\phi}}(X_i).$$

Reduction then answers treatment queries by integrating the response kernel against this fixed belief:

$$\widehat{\mathcal{L}}_{i,t}(dy) = \int_{\mathcal{U}} K_{\widehat{\theta}}(dy \mid X_i, t, u)\, \widehat{Q}_i(du), \qquad t \in \{0, 1\}. \tag{A.4}$$

Changing treatment changes the response query, not the learner's judgment about which fixed unit is being queried. The same stored $\widehat{Q}_i$ must therefore be used for both $t = 0$ and $t = 1$.

If paired Monte Carlo responses are desired, a candidate $\widetilde{u}_i^{(m)}$ should be drawn once and reused across treatment queries. Drawing a new candidate for each treatment would compare different simulated unit hypotheses. Whether event-noise realizations are shared across treatments is a separate cross-world modeling decision.

Outcomes may supervise shared abductor parameters in training data, subject to outer-fold honesty. But the focal belief used by the CATE estimator must be computed from pretreatment evidence only. Supplying the focal T_i or Y_i at application time would define an ex-post, factual-event-conditioned problem rather than the ex-ante CATE problem.

A.4 CATE as a mixture over response-relevant types

Suppose the relevant conditional means exist, and define

$$m_t(x, u) = \mathbb{E}\{Y(t) \mid X = x, U = u\}, \qquad \Delta_0(x, u) = m_1(x, u) - m_0(x, u).$$

Proposition 2 (Mixture representation of CATE)**.** *At every regular x,*

$$\tau(x) = \int_{\mathcal{U}} \Delta_0(x, u)\, Q_x^{\star,\mathrm{tar}}(du).$$

Proof. By iterated expectation,

$$\mathbb{E}\{Y(t) \mid X = x\} = \int_{\mathcal{U}} \mathbb{E}\{Y(t) \mid X = x, U = u\}\, P_{\mathrm{tar}}(U \in du \mid X = x).$$

Subtracting the expressions for $t = 0$ and $t = 1$ gives the result. □

Proposition 2 is a population decomposition, not an observational identification theorem for either $Q_x^{\star,\mathrm{tar}}$ or $\Delta_0(x, u)$. Ordinary (X, T, Y) data may identify their mixture $\tau(x)$ without identifying the two latent components separately.

Moreover, because $O = X$, two records satisfying $X_i = X_j = x$ receive the same oracle conditional belief. Unit Abduction cannot create within-x personalization absent from the declared evidence. Its additional modeling object is the full belief geometry, not extra observed information.

A.5 Predictive equivalence and latent non-identification

A learned response model constrains the observable predictive mixture

$$p_{\phi,\theta}(dy \mid x, t) = \int_{\mathcal{U}} K_\theta(dy \mid x, t, u)\, q_\phi(du \mid x).$$

Good predictive fit constrains this mixture, not necessarily its latent factorization.

Proposition 3 (A basic observational equivalence)**.** *Let $P_0(dy \mid x, t)$ be any observable conditional outcome law. It admits both:*

(i) a one-point latent space with $Q_x = \delta_0$ and $K(dy \mid x, t, 0) = P_0(dy \mid x, t)$; and

(ii) a two-point latent space with $Q_x = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $K(dy \mid x, t, u) = P_0(dy \mid x, t)$ for both values of u.

The observable law is identical under the two latent specifications.

Proof. Integrating either response kernel against its stated belief returns $P_0(dy \mid x, t)$. □

This elementary construction is enough to show that the existence, dimension, and belief law of an unrestricted latent selector are not identified from the predictive mixture alone. Richer response-relevant mixture decompositions can also be nonunique.

Proposition 4 (Latent reparameterization invariance)**.** *Let $b : \mathcal{U} \to \mathcal{U}'$ be a measurable bijection with measurable inverse. Define $Q'_x = b_{\#}Q_x$ and*

$$K'(A \mid x, t, u') = K(A \mid x, t, b^{-1}(u'))$$

for measurable outcome sets A. Then

$$\int_{\mathcal{U}'} K'(A \mid x, t, u')\, Q'_x(du') = \int_{\mathcal{U}} K(A \mid x, t, u)\, Q_x(du).$$

Proof. The equality is the defining change-of-variables identity for a pushforward measure. □

Thus a calibrated statement such as $\widehat{Q}_i \approx P(U \in du \mid X = x_i)$ requires semantic anchoring and calibration assumptions. Possible sources include externally known mechanisms, repeated measurements, multi-environment restrictions, identifiable component families, or direct calibration data. The same-token rule is a coherence condition; it does not identify the actual u_i or a preferred coordinate system.

A.6 Causal identification is a separate layer

A transparent structural sufficient condition for conditional exchangeability in (A.1) is

$$T \perp (U, E_0, E_1) \mid X.$$

If U affects both assignment and outcomes after conditioning on X, an abducted belief over U does not automatically remove the hidden confounding. Likewise,

$$q_\phi(u \mid X = x) > 0 \quad \not\Rightarrow \quad P(T = t \mid X = x) > 0,$$

and neither condition implies positive probability of inclusion in an observed sample. The three support concepts belong to different inferential problems.

Under Assumption 1, $\tau(x)$ can be identified directly from the observed law. This does not identify the latent decomposition $(Q^\star_x, \Delta_0)$. Conversely, even an oracle-calibrated $Q^\star_x$ would not repair confounding, treatment nonoverlap, or selection bias.

A.7 Belief families and moment requirements

An abductor may return a point mass, Gaussian law, Cauchy law, finite mixture, or another probability measure:

$$Q_i = \delta_{\mu_i}, \qquad Q_i = \mathcal{N}(\mu_i, \Sigma_i), \qquad \text{or} \qquad Q_i = \text{Cauchy}(\mu_i, \gamma_i).$$

These families encode different epistemic geometries; none is automatically the true conditional law.

A Cauchy belief has no latent mean, so a posterior-mean embedding is undefined. This does not prevent Reduction when the response functional is bounded or integrable under Q_i, and it illustrates why a complete probability measure may carry usable structure that a mean cannot. The mean-CATE formulation in this paper nevertheless requires

$$\mathbb{E}|Y(t)| < \infty, \qquad t \in \{0, 1\}.$$

If a response specialization violates this condition, the causal target must change to a well-defined location, median, quantile, or distributional contrast.

Scope of the present paper. The estimator uses learned beliefs to define a geometry over observed records and smooths an orthogonal causal signal in that geometry. It does not claim that observational prediction alone identifies the true latent belief, the realized unit value, or the latent type-effect surface.

B Proof details and estimation notes

B.1 Proof of Theorem 1

Proof. Because $w_{q,h}(X)$ is pretreatment-measurable, iterated expectation gives

$$\mathbb{E}[w_{q,h}(X)\{Y(1) - Y(0)\}] = \mathbb{E}[w_{q,h}(X)\tau(X)].$$

Under consistency and conditional exchangeability,

$$\mathbb{E}\{Y(t) \mid X\} = \mathbb{E}(Y \mid T = t, X) = \mu_{0,t}(X).$$

Substitution proves (4.1).

For the second equality, condition (3.1) on X at the true nuisances. Positivity and consistency imply

$$\mathbb{E}\left[\frac{T}{e_0(X)}\{Y - \mu_{0,1}(X)\} \,\middle|\, X\right] = 0$$

and the analogous control-arm term is zero. Hence $\mathbb{E}\{\Gamma(Z; \eta_0) \mid X\} = \tau(X)$. Multiplying by $w_{q,h}(X)$, taking expectations, and normalizing proves (4.2). □

B.2 Conditional double robustness of the donor score

Let $\eta = (e, \mu_0, \mu_1)$ be arbitrary candidate nuisance functions bounded away from the propensity endpoints. Direct conditioning yields

$$\mathbb{E}\{\Gamma(Z; \eta) \mid X\} - \tau(X) = \{e(X) - e_0(X)\}\left[\frac{\mu_1(X) - \mu_{0,1}(X)}{e(X)} + \frac{\mu_0(X) - \mu_{0,0}(X)}{1 - e(X)}\right]. \tag{B.1}$$

Thus the conditional bias vanishes if $e = e_0$, or if both outcome regressions are correct. Because the localization weight is a function of X, multiplying (B.1) by $w_{q,h}(X)$ preserves this population double-robust property. Cross-fitting is used to control empirical process and overfitting effects; it is not an additional identification assumption.

B.3 Proof of Proposition 1

Proof. Using Theorem 1 and conditioning on $R = Q_X$,

$$\begin{aligned}\theta_{h,\mathcal{Q}}(q) &= \frac{\mathbb{E}[K_h\{d(q,R)\}\tau(X)]}{\mathbb{E}[K_h\{d(q,R)\}]} \\ &= \frac{\mathbb{E}[K_h\{d(q,R)\}m(R)]}{\mathbb{E}[K_h\{d(q,R)\}]}.\end{aligned}$$

Assumption 3 makes the final ratio converge to $m(q)$. □

B.4 An implementation-level cross-fitting template

For a first reproducible implementation, use mutually exclusive outer folds $\{\mathcal{A}_k\}_{k=1}^{K_A}$ for the abductor and causal folds $\{\mathcal{C}_\ell\}_{\ell=1}^{K_C}$ for nuisance scores.

1. For each outer fold k, fit the abductor on $\mathcal{A}_k^c$, and compute $\widehat{Q}_i = \widehat{\mathcal{Q}}^{(-k)}(X_i)$ for $i \in \mathcal{A}_k$. If abductor training uses T or Y, no row may receive a belief from a model trained on its own realized T_i, Y_i.
2. For each causal fold ℓ, fit propensity and outcome nuisances on $\mathcal{C}_\ell^c$, and compute $\widehat{\Gamma}_i$ on $\mathcal{C}_\ell$.
3. For every focal query, exclude the focal row from its donor set, or use an independent evaluation sample. Form $\widehat{\alpha}_{ij}$ only from the stored pretreatment beliefs.
4. Tune the bandwidth without access to oracle focal effects. A pseudo-outcome validation rule must itself be cross-fitted. Report the induced n_{eff}, local mass, and largest donor weight.
5. Compare all baselines under identical focal/donor splits and, where appropriate, identical nuisance estimates.

   For an arbitrary external query x, compute

$$q = \widehat{\mathcal{Q}}(x), \qquad \widehat{\alpha}_j(q) = \frac{K_h\{d(q,\widehat{Q}_j)\}}{\sum_\ell K_h\{d(q,\widehat{Q}_\ell)\}},$$

and return $\sum_j \widehat{\alpha}_j(q)\widehat{\Gamma}_j$. The abductor may be queried once for x; hypothetical treatment values do not enter this step.

B.5 Open proof obligations

The following claims are not established by the fixed-map result:

- a rate or central limit theorem for learned-Q weights;
- Neyman orthogonality with respect to belief-map error;
- consistency under data-adaptive distance or bandwidth learning;
- identification of $Q_x^\star$, u_i, or $\Delta_0(x,u)$; and
- valid confidence intervals for DiscoCATE–DR.

These are theorem targets for later versions rather than implicit consequences of causal-forest or AIPW theory.