A Canonical Background Distribution for Shapley Attribution

Ludger Hentschel

July 13, 2026

Abstract

Shapley-value explanations depend on a background distribution that defines the reference population against which feature contributions are measured. In common practice, this background is often chosen heuristically—as the full training sample, a random subsample, or representative cluster centers—on grounds of representativeness or computational convenience rather than correspondence to the explanatory question.

We argue that the attribution question should determine the background distribution. For the canonical question of why a model produces a given output rather than its unconditional prediction, the natural reference is the observed-data distribution localized around the prediction-neutral manifold. This background uses genuine observations whose fitted predictions are close to the neutral level, thereby combining a meaningful reference prediction with realistic reference inputs. Because the background is constructed directly from the observed data rather than by random sampling or simulation, it is deterministic conditional on the fitted model and reference sample. Under this construction, Shapley efficiency implies that feature attributions sum to the difference between the prediction of interest and the neutral prediction—the quantity the explanation is intended to decompose.

The method applies to scalar outputs and to multi-class classification using the full centered-logit vector. In the multi-class case, a single vector-neutral background yields coherent class attributions and preserves the natural cross-class adding-up identities. The construction changes neither the Shapley algorithm nor its axioms. It alters only the reference distribution and can therefore be supplied directly to existing SHAP implementations.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

1 Introduction

Shapley-value attribution methods explain a prediction by allocating the difference between the prediction of interest and a reference prediction across the input features. In machine-learning applications this reference is determined by a background distribution: the population used to define what it means for features to be absent or unknown. The widely used SHAP implementation (Lundberg and Lee, 2017) makes this dependence explicit by requiring the user to supply a background sample from which the reference distribution is constructed. The resulting feature attributions depend materially on this choice. When the background itself is obtained by random sampling or simulation, the explanations also inherit an additional source of Monte Carlo variability.

Every feature attribution answers a counterfactual question, and for Shapley (Shapley, 1953) methods that counterfactual is fixed by the background distribution (Merrick and Taly, 2020). Changing the background changes the comparison being made even though the attribution algorithm is unchanged. The central question is therefore how to choose a background distribution that matches the explanatory question while remaining supported by the observed data and reproducible from one analysis to the next. In common practice this choice is often made on other grounds: the background is taken as the full training sample, a random subsample, or representative cluster centers, selected for representativeness or computational convenience rather than for correspondence to any particular explanatory question.[1]

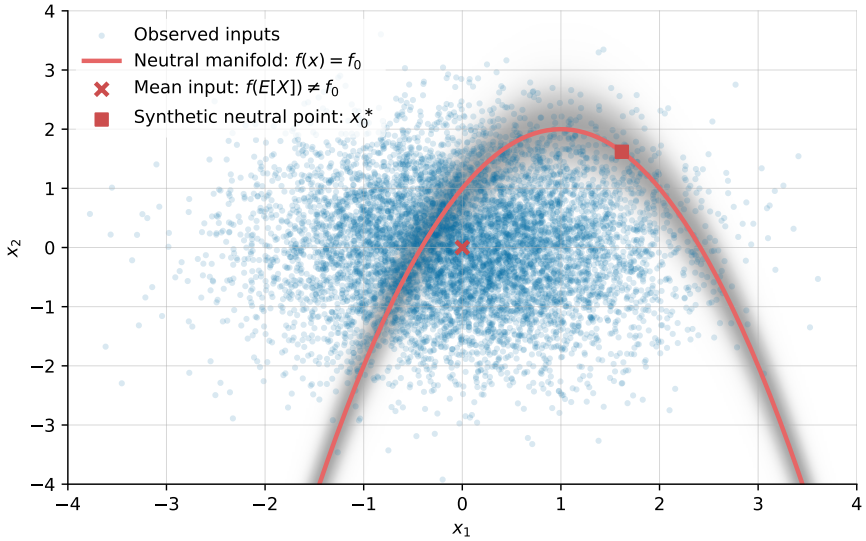
Two ideas in the literature move away from this default, and our proposal combines them in a constructive manner.

The first idea is that the reference should be a baseline whose prediction has semantic meaning rather than an arbitrary input such as the origin. Izzo, Lipani, Okhrati, and Medda (2021) anchor the baseline on the decision boundary for classification problems, so that Shapley values are read directly against a decision-relevant reference rather than an arbitrary one. Using monotonicity in each feature, their construction collapses the neutral set to a single synthetic point, which need not lie in the data support.[2] Our proposal adopts the same general philosophy but a different reference prediction. We use the unconditional prediction, so that feature attributions explain departures from the model's typical prediction rather than departures from the decision boundary. More generally, we argue that the appropriate

[1] Any background distribution implicitly answers *some* contrastive question; the issue is whether it is the question the analyst intends.

[2] Izzo, Lipani, Okhrati, and Medda (2021) recognize that their neutral point belongs to a larger neutral set but leave the full neutral set and regression models to future work.

Figure 1: Neutral Prediction Manifold



Simulated inputs (blue markers) and the neutral prediction manifold $\mathcal{M}_0 = \{x : f(x) = f_0\}$ (red curve) for a synthetic regression The shaded band shows observations with high kernel weight $K(f(x) - f_0)$; its Euclidean width varies across the input space and across input axes because the kernel acts on prediction distance; where the prediction changes rapidly a small move off the manifold produces a large prediction gap.

The figure marks the three reference choices the paper contrasts. The mean input $\mathbb{E}[X] = (0, 0)$ (cross) predicts $f(\mathbb{E}[X]) \neq f_0$ and so does not sit on $\mathcal{M}_0$: a mean-input baseline decomposes the wrong gap. The synthetic neutral point x_0^* (square), solves $f(x_0^*) = f_0$ in the manner of Izzo, Lipani, Okhrati, and Medda (2021): it is on the manifold in output, but in a sparsely populated tail far from the bulk of the data. The prediction-neutral background distribution instead draws on the observed inputs within the shaded band: simultaneously neutral in output and supported by data. Regions of $\mathcal{M}_0$ far from the data are represented sparsely precisely because they contain few observations.

background distribution is determined by the prediction level relevant to the attribution question.

The second idea is that the reference should be a *distribution* rather than a point. Moreover it should be a distribution anchored at a meaningful output level directly associated with the attribution question. Counterfactual SHAP (Albini, Long, Dervovic, and Magazzeni, 2022) takes the background to be a set of generated counterfactuals near the decision boundary and observes that, as those points approach the boundary in output, the Shapley values sum to the gap between the prediction and the boundary. A material challenge for this approach is generating realistic observations in high dimensions. In general, this requires an accurate high-dimensional model for the feature distribution. The prediction-specific anchor and the distributional background are therefore each established in the literature.

What has not been proposed is to realize a prediction-neutral background distribution directly from the observed data. We propose exactly this. If the

explanatory question is why the model produces output $f(x)$ rather than a neutral output f_0, the natural background is the empirical data distribution restricted to observations the fitted model already treats as approximately neutral. Every reference is therefore a genuine observation, no monotonicity assumption or feature-distribution model is required, and the resulting background is a deterministic function of the fitted model and the reference sample. Because the reference population is neutral by construction, the resulting feature attributions explain the intended prediction difference, $f(x) - f_0$. This is the distinguishing contribution of the paper.

Figure 1 illustrates the idea. The figure shows a sample of two-dimensional observations as blue markers, the prediction-neutral manifold as a red line, and a neighborhood around the manifold as a shaded region. The observations most relevant to the attribution question are the blue markers inside the shaded region. The resulting background distribution is associated simultaneously with approximately neutral predictions and realistic observations. Areas of the neutral manifold far from the data are represented sparsely precisely because they contain few observations. Because the construction depends only on the observed sample and the fitted model, repeated analyses produce the same background without introducing additional sampling variability.

In contrast to Izzo, Lipani, Okhrati, and Medda (2021), we treat the neutral set as a distribution populated with data rather than collapse it to a constructed point. In contrast to the counterfactual background of Albini, Long, Dervovic, and Magazzeni (2022), we use observed inputs drawn from a single fixed population rather than instance-specific points generated to flip a decision. The reference distribution is simultaneously neutral in output, supported by observed data, and deterministic conditional on the fitted model and reference sample. The same prediction-neutral principle underlies a canonical baseline for integrated gradients (Hentschel, 2026); the present paper develops its counterpart for Shapley attribution, where the neutral reference enters through the background distribution rather than an integration path. The construction requires no change to the Shapley algorithm or its axioms; it alters only the reference distribution and can be supplied directly to existing SHAP implementations.

The remainder of the paper proceeds in two main steps. First, we formalize the prediction-neutral background distribution and explain its additive payoff. We then present two examples: a synthetic regression illustrating the distinction between conventional and prediction-neutral backgrounds, and a multi-class handwritten-digit classification problem illustrating both the

practical construction and the elimination of background-induced sampling variability. Appendix A describes the deterministic background-construction algorithm implemented in the accompanying software package.

2 Prediction-Neutral Background Distributions

2.1 Defining the Background Distribution

We consider the canonical attribution question: *why does the model produce output $f(x)$ rather than a neutral output f_0?* Throughout, f denotes the attribution target—the prediction in regression and the centered-logit vector in classification—and f_0 denotes its neutral value. For regression, the attribution target is the scalar prediction and the neutral value is

$$f_0 = \mathbb{E}[f(X)]. \tag{1}$$

For multi-class classification, the attribution target is the centered-logit vector

$$\widetilde{z}(x) = z(x) - \overline{z}(x)\mathbf{1}, \tag{2}$$

where $z(x)$ is the vector of model logits and $\overline{z}(x)$ is their average. The neutral centered-logit vector is

$$\widetilde{z}_0 = \mathbb{E}[\widetilde{z}(X)]. \tag{3}$$

The natural background distribution is the data distribution restricted to inputs the fitted model treats as neutral,

$$Q_N = P_{X|f(X)=f_0}. \tag{4}$$

For continuously distributed inputs, conditioning on the equality $f(X) = f_0$ is understood as a limiting localization. Let

$$Q_{N,h}(A) = \frac{\mathbb{E}\big[\mathbf{1}_{\{X\in A\}} K_h\big(f(X) - f_0\big)\big]}{\mathbb{E}\big[K_h\big(f(X) - f_0\big)\big]}, \tag{5}$$

where K_h is a kernel in prediction space. The prediction-neutral background is the limiting distribution as $h \downarrow 0$, when this limit is well defined. The notation $P_{X|f(X)=f_0}$ is shorthand for this localization.

The observed data may provide little support near the neutral prediction. In that case, a finite-sample implementation must use a wider neighborhood

to obtain a background of useful size. This does not change the target distribution: it introduces an approximation whose quality is summarized by the distance between the resulting background prediction and f_0. Thus the asymptotic construction uses a shrinking bandwidth, while finite samples may require a wider bandwidth when the neutral region is sparsely populated.

Equivalently, the reference population is supported on the prediction-neutral manifold

$$\mathcal{M}_0 = \left\{x : f(x) = f_0\right\}. \tag{6}$$

The centered-logit neutral vector $\widetilde{z}_0$ corresponds uniquely to the probability vector

$$p_0 = \mathrm{softmax}(\widetilde{z}_0). \tag{7}$$

The two representations therefore define the same neutral prediction set. We work on the centered-logit scale because it removes the irrelevant common-logit location and preserves additivity.

The proposed explanation is the ordinary Shapley attribution computed using Q_N as the background,

$$\phi_N(x) = \phi(x; Q_N), \tag{8}$$

where $\phi(x; Q)$ denotes the Shapley attribution induced by background distribution Q.

Using a background *distribution* rather than a single reference point is well defined for Shapley attribution: the Shapley value of the game averaged over references equals the average of the single-reference Shapley values (Chen, Lundberg, and Lee, 2022). Averaging single-reference attributions over draws from Q_N is therefore exactly the attribution under Q_N, and the efficiency identity below applies to the averaged game with $\mathbb{E}_{Q_N}[f]$ as its reference level.

Treating the neutral set as a distribution rather than a point is a direct consequence of nonlinearity. For nonlinear models, neutrality and representativeness are in direct competition since the mean input need not produce the neutral output

$$f(\mathbb{E}[X]) \neq \mathbb{E}[f(X)]. \tag{9}$$

More fundamentally, there is typically no single neutral input: many

distinct inputs can produce the same neutral output. Neutrality is therefore naturally a property of a reference distribution, and the prediction-neutral background replaces the search for a canonical baseline point with a canonical background population. While we take f_0 to be the marginal mean output, the same construction accommodates an externally meaningful neutral level, such as a decision threshold, by conditioning on that value instead.

The empty-coalition value depends on a background Q only through its mean prediction,

$$v_Q(\varnothing) = \mathbb{E}_Q[f]. \tag{10}$$

Different backgrounds can therefore share the same reference prediction while inducing different Shapley games. For nonempty coalitions,

$$v_Q(S) = \mathbb{E}_{R\sim Q}\left[f(x_S, R_{\bar{S}})\right], \tag{11}$$

which depends on the full distribution of the reference inputs, not only on $\mathbb{E}_Q[f]$.

The unconditional data distribution and the prediction-neutral distribution both have reference prediction f_0 at the population level. They nevertheless answer different counterfactual questions. The unconditional distribution averages over references with predictions potentially far above and far below f_0, whereas the prediction-neutral distribution uses references that the fitted model treats as neutral individually. Conditioning directly on neutrality therefore fixes both the desired reference prediction and the reference population used to define missing features.

2.2 The Additive Payoff

Conditioning on neutrality makes the additive bookkeeping exact. For any background Q, efficiency gives

$$\sum_j \phi_j(x; Q) = f(x) - \mathbb{E}_Q[f]. \tag{12}$$

Under the prediction-neutral background, $\mathbb{E}_{Q_N}[f] = f_0$, so

$$\sum_j \phi_{N,j}(x) = f(x) - f_0. \tag{13}$$

Thus the attribution decomposes precisely the gap from the neutral output, which is the explanatory quantity of interest.

For multi-class classification, the attribution is vector-valued. Let

$$\Phi_j(x) \in \mathbb{R}^K \tag{14}$$

denote the attribution vector assigned to feature j. Using one common vector-neutral background, efficiency gives

$$\sum_j \Phi_j(x) = \widetilde{z}(x) - \widetilde{z}_0. \tag{15}$$

A class-specific attribution is the corresponding coordinate projection,

$$\phi_{j,c}(x) = e_c' \Phi_j(x). \tag{16}$$

It therefore satisfies

$$\sum_j \phi_{j,c}(x) = \widetilde{z}_c(x) - \widetilde{z}_{0,c}. \tag{17}$$

Because centered logits obey

$$\mathbf{1}'\widetilde{z}(x) = 0, \tag{18}$$

linearity of the Shapley operator also implies the feature-level cross-class identity

$$\mathbf{1}'\Phi_j(x) = 0 \tag{19}$$

for every feature j. Thus positive attribution to some classes must be offset by negative attribution to others. These identities require a single common background distribution; separate class-specific backgrounds would define different games and need not be mutually consistent.

The proposal modifies neither the Shapley algorithm nor its axioms. Efficiency, symmetry, dummy, and additivity continue to hold for the game induced by the chosen background. What changes is the reference distribution, and therefore the contrast being explained.

The scope of this guarantee is important. Neutrality fixes the empty-coalition value,

$$v(\varnothing) = \mathbb{E}_{Q_N}[f] = f_0. \tag{20}$$

For a nonempty coalition S, standard interventional SHAP uses values of the form

$$v(S) = \mathbb{E}_{R\sim Q_N}\left[f(x_S, R_{\overline{S}})\right], \tag{21}$$

which paste the observed features x_S onto a neutral reference draw. The resulting composite input need not itself lie on $\mathcal{M}_0$ or even on the data manifold. Prediction-neutral backgrounds therefore control the reference anchor and its interpretation, not the realism of every intermediate coalition. The choice between interventional and conditional value functions is an orthogonal issue.

Identity equation (13) is exact for the population background Q_N, where $\mathbb{E}_{Q_N}[f] = f_0$ holds by definition of conditioning on $\mathcal{M}_0$. A finite background sample attains it up to a neutrality tolerance

$$\varepsilon = \left\|\overline{f}_S - f_0\right\|, \qquad \overline{f}_S = \frac{1}{|S|}\sum_{i\in S} f(x_i), \tag{22}$$

the gap between the sample's average prediction and the neutral output. The construction of Section 2.3 controls ε and reports it as a diagnostic, so the decomposition $\sum_j \phi_{N,j}(x) = f(x) - f_0$ holds to that tolerance in practice and exactly in the limit of an exactly neutral background.

2.3 Constructing a SHAP Background Sample

In practice, we approximate Q_N from the observed training data. Let x_i denote the training observations and define prediction gaps

$$r_i = f(x_i) - f_0, \tag{23}$$

which are scalar in regression and binary classification. For multi-class classification, the prediction gaps are vector-valued. A kernel on these gaps assigns larger relevance scores to observations whose predictions lie closer to the neutral output. Gaussian, Epanechnikov, and uniform kernels (Epanechnikov, 1969; Bierens, 1994) are natural choices. The resulting scores identify a neighborhood of the prediction-neutral manifold populated by observed inputs.

For SHAP, the practical object is a background sample. Standard SHAP implementations consume an equally weighted collection of background rows. The goal is therefore to return a compact set of observed inputs whose

average prediction is neutral, or very close to neutral:

$$\frac{1}{|S|}\sum_{i\in S} f(x_i) \approx f_0. \tag{24}$$

We can pass this background directly to SHAP without modifying the attribution algorithm.

Absent weights, we cannot guarantee exact neutrality. But in practical samples of useful size we can generally select backgrounds that are very close to neutral. Although we can simulate weighted backgrounds by resampling observations, this often materially increases computational effort in the attribution stage.

Our accompanying software constructs an equal-weight background by translating a local neighborhood in prediction space. This preserves localization: every selected background is a compact neighborhood rather than an arbitrary collection of observations chosen only to satisfy a moment condition.

For a scalar attribution target, let

$$r_i = f(x_i) - f_0. \tag{25}$$

For a fixed background size m, the initial background contains the m observations nearest zero in prediction-gap space. The center of this fixed-size interval is then shifted along the real line, and the shift is chosen to minimize the absolute mean prediction gap of the selected observations.

For a vector attribution target, let

$$u_i = W^{1/2}\{f(x_i) - f_0\}, \tag{26}$$

where W is a regularized inverse covariance matrix that standardizes distances in prediction space. The initial background contains the m observations nearest the origin. If their mean is $\overline{u}_0 \neq 0$, the neighborhood center is translated along the one-dimensional direction

$$d = -\frac{\overline{u}_0}{\|\overline{u}_0\|}. \tag{27}$$

For each scalar shift $a \geq 0$, let $S_m(a)$ contain the m observations nearest the

translated center ad. The selected background uses

$$\widehat{a} = \arg\min_{a \geq 0} \left\| \frac{1}{m} \sum_{i \in S_m(a)} u_i \right\|. \tag{28}$$

This is the vector analogue of sliding a fixed-size interval in the scalar case.

A fixed-width alternative selects all observations inside a translated kernel neighborhood. That formulation is natural for asymptotic analysis because its bandwidth can shrink with sample size. We use the fixed-size version in the empirical examples to hold the SHAP background budget constant across methods.

This construction is intentionally modular. The contribution of the paper is the prediction-neutral background distribution, not a new Shapley algorithm. The software supplies one practical way to build the corresponding background sample. Appendix A describes the algorithmic details.

2.4 Practical Considerations

Prediction-neutral selection provides a principled alternative to *ad hoc* background reduction. Observations far from f_0 are discarded not because of random thinning or clustering, but because they correspond less directly to the explanatory benchmark. The retained references are actual observations and remain tied to the empirical data distribution near the neutral manifold.

This concentration is a relevance gain rather than a universal variance guarantee. If very few observations lie near neutrality, the selected background may be small or may require a wider neutral band. The retained count, the effective number of references, and the average background prediction are therefore useful diagnostics. They report how well the data support the neutral comparison.

Due to computational costs, conventional background distributions often sample from the data. Especially when the sample is relatively small, this can introduce material Monte Carlo variability in the attribution results. Our construction is fully deterministic and avoids this sampling error. For a chosen bandwidth or sample size, we construct a fixed sample of the data that is closest to prediction-neutral.

The proposal is deliberately minimal. Its contribution is not a new attribution rule but a background distribution motivated directly by the explanatory question, realized with observed inputs, and usable within existing SHAP workflows.

2.5 Scope and Limitations

The prediction-neutral construction determines the reference population and therefore the empty-coalition value of the Shapley game. It does not ensure that every hybrid input formed by combining target and background features lies on the data manifold. That issue depends on the choice between interventional and conditional Shapley value functions and is separate from background selection.

The amount of empirical support near the neutral prediction can vary. When the neutral region is well populated, a narrow neighborhood provides a direct approximation to the prediction-neutral distribution. When support is sparse, a wider neighborhood may be required to obtain a background of useful size. The resulting neutrality gap should then be reported rather than treated as zero.

For vector-valued predictions, localization takes place in prediction space, whose effective dimension is $K - 1$ for K centered class logits. The method therefore avoids dependence on the potentially much larger input dimension, but problems with very many classes may still require additional regularization or dimension reduction in prediction space.

Finally, the observed background is deterministic only conditional on the training sample, fitted model, and chosen construction. It eliminates background-draw variability, but not uncertainty arising from resampling the training data, refitting the model, or changing the explained observations.

3 Examples

We now illustrate the canonical background distribution in two examples. The first is a synthetic regression, where we know the analytical attribution results. The second uses digit images in a multi-class classification setting.

3.1 Synthetic Regression

A simple nonlinear regression gives us control over both the data and the neutral manifold. Let

$$f(x) = a + (x_1 - c)^2 + b\,x_2, \qquad X_1, X_2 \overset{iid}{\sim} N(0,1), \tag{29}$$

with $a = 0$, $c = 1$, $b = 1$. The neutral prediction is

$$f_0 = \mathbb{E}[f(X)] = a + 1 + c^2 = 2, \tag{30}$$

whereas the mean input predicts

$$f(\mathbb{E}[X]) = a + c^2 = 1. \tag{31}$$

The gap $f_0 - f(\mathbb{E}[X]) = 1$ is entirely a Jensen effect from the squared term: the mean input sits one unit below the neutral level because the model is nonlinear. A single mean-input baseline therefore poses the wrong question, decomposing $f(x) - 1$ rather than $f(x) - f_0 = f(x) - 2$. This gap is a property of the baseline point, not of any Shapley computation built on it, and it does not vanish as the sample grows.

The neutral manifold $\mathcal{M}_0 = \{x : f(x) = 2\}$ is the curve $(x_1 - 1)^2 + x_2 = 2$. Typical observations sit near $(0, 0)$, while reaching the neutral level with x_1 near zero requires $x_2 \approx 1$, so points on or near $\mathcal{M}_0$ carry systematically elevated x_2. This displacement of the reference moments away from the unconditional means is what a neutral background encodes and what a full-sample background discards. Figure 1 illustrates the data and the neutral manifold for this example.

Because f is additive across coordinates with no interaction, the coordinate attributions take a transparent form for any background Q,

$$\phi_1(x; Q) = (x_1 - 1)^2 - \mathbb{E}_Q[(X_1 - 1)^2], \tag{32}$$

$$\phi_2(x; Q) = x_2 - \mathbb{E}_Q[X_2]. \tag{33}$$

Each coordinate is the target's feature contribution minus its expectation under the background, so the background enters only through these two moments. This is exact here—not a kernel or sampling approximation—and lets us read the effect of each background directly from its reference moments.

Table 1 reports SHAP attributions for five background constructions, averaged over one million target observations. The two panels separate the two axes along which backgrounds differ: the *prediction level* the background represents, and, among neutral backgrounds, *how* the neutral set is represented.

Panel A contrasts two single points. The mean input $\overline{x} = (0, 0)$ predicts $f(\overline{x}) = 1 \neq f_0$, so it answers the wrong question: its mean total attribution is approximately one—the Jensen gap—rather than zero. This is not a property of the Shapley algorithm but of the comparison point. The synthetic neutral point x_0^* is on the manifold in output ($f(x_0^*) = 2$) but lies in a sparsely populated tail; it assigns almost the entire positive contribution to the nonlinear feature x_1 ($\phi_1 \approx 1.68$) and a large offsetting negative contribution to x_2

Table 1: Average attributions for synthetic regression example

Background	n_B	$\overline{f}_B$	$\overline{f}_B - f_0$	$\overline{\phi}_1$	$\overline{\phi}_2$	$\overline{\lvert\phi_1\rvert}$	$\overline{\lvert\phi_2\rvert}$
Panel A: Conventional single-point backgrounds							
$\overline{x}$	1	1.0000	-1.0000	1.0657	-0.0040	1.7018	0.7997
x_0^*	1	2.0000	0.0000	1.6837	-1.6220	1.8335	1.6658
Panel B: Distributional backgrounds							
Sample	100	2.1745	0.1745	-0.1529	0.0401	1.8746	0.7999
Sim. neutral	100	2.0000	0.0000	0.5959	-0.5342	1.7130	0.9118
Canonical	101	2.0000	0.0000	0.6211	-0.5594	1.7103	0.9219

The table reports SHAP attributions for the synthetic regression model $f(x) = (x_1 - 1)^2 + x_2$, where (x_1, x_2) are independent standard normal variables and the prediction-neutral level is $f_0 = E[f(X)] = 2$. Panel A compares single-point backgrounds: the mean input $\overline{x} = E[X]$ and the prediction-neutral point x_0^* satisfying $f(x_0^*) = f_0$. Panel B compares distributional backgrounds with approximately 100 observations: an ordinary empirical sample, a generated prediction-neutral sample, and the proposed canonical background obtained from observed points in a prediction-neutral band.

The columns $\overline{f}_B$ and $\overline{f}_B - f_0$ summarize the prediction level of each background. The empirical sample is not exactly prediction neutral because it is a finite random sample, whereas the generated and canonical backgrounds are constructed to satisfy $\overline{f}_B \approx f_0$. The signed and absolute SHAP attributions are averaged over one million target observations drawn from the population. Because SHAP satisfies efficiency, the mean signed attributions sum to the average prediction difference between the target population and the background prediction, while the mean absolute attributions summarize the typical attribution magnitude for each feature.

($\phi_2 \approx -1.62$). A single neutral point is neutral in output yet unrepresentative in input, and the resulting allocation reflects the arbitrary location of that point rather than the data.

Panel B compares three distributional backgrounds of roughly one hundred observations. The random empirical background is not exactly neutral in this realization: its mean prediction is $\overline{f}_B \approx 2.17$, differing from f_0 by about 0.17. So, like the mean input it partly answers a non-neutral question, and its signed attributions ($\phi_1 \approx -0.15$, $\phi_2 \approx 0.04$) reflect that displacement. The simulated-neutral and canonical backgrounds are both constructed to satisfy $\overline{f}_B \approx f_0$, and they produce nearly identical attributions ($\phi_1 \approx 0.60$ and 0.62, $\phi_2 \approx -0.53$ and -0.56). This agreement is expected and is the point of the synthetic case: because the generated counterfactuals here are drawn from the true data-generating process, the observed prediction-neutral background reproduces what a correctly specified generative model would give, using only observed inputs. The synthetic example therefore validates the construction and its decomposition; the digit example, where the data-generating process is unknown, is where drawing on observed data rather than a fitted model matters.

Table 2: Attribution variability for synthetic regression example

Background	n_B	Reps.	CV($\overline{f}_B - f_0$)	CV($\overline{\phi}_1$)	CV($\overline{\phi}_2$)
Sample	101	1000	0.2647	0.1345	0.1257
Sim. neutral	101	1000	0.0000	0.0561	0.1087
Canonical	101	–	0.0000	0.0000	0.0000

The table reports the sampling variability of distributional SHAP backgrounds for the synthetic regression model $f(x) = (x_1 - 1)^2 + x_2$. All reported quantities are conditional on the same reference sample and the same target sample.

For the empirical background, each repetition draws a new random sample of 101 observations from the reference distribution. For the simulated neutral background, each repetition draws a new sample of 101 observations from the generated prediction-neutral distribution. The canonical background is constructed deterministically from the observed reference sample using the prediction-neutral uniform-band procedure and therefore does not vary across repetitions.

The attribution-variability columns report coefficients of variation: the standard deviation across repeated background draws of the mean signed attribution, divided by the corresponding mean absolute attribution.

The signed attributions also make the allocation concrete. Under a neutral background, $\mathbb{E}_{Q_N}[X_2] > 0$ (the elevated-x_2 displacement above), so a typical target with x_2 near zero receives a negative x_2 attribution; symmetrically $\mathbb{E}_{Q_N}[(X_1 - 1)^2]$ falls below its unconditional value and x_1 receives a positive attribution. The unconditional background distribution, whose reference moments equal the unconditional moments, produces zero population-average signed attributions that average away this structure—not because the features are unimportant, but because a full-sample background answers a different question.

Table 2 turns to sampling variability, holding the target sample, the model, and the background size fixed at roughly one hundred. The variability usually associated with SHAP has two sources. A random empirical background varies in both the prediction level it represents and the particular reference observations drawn; its coefficients of variation are substantial (about 0.13 for each signed attribution) and it additionally carries variation in $\overline{f}_B - f_0$. A simulated-neutral background fixes the prediction level ($\overline{f}_B = f_0$ by construction, zero variability in that column) but still resamples reference observations, so attribution variability remains ($CV(\phi_2) \approx 0.11$). The canonical background removes both sources: conditional on the reference sample it is a deterministic function of the data, so its attribution has no sampling variability at all. At a common background size, the proposed background sits at zero where the sampling backgrounds are strictly positive; that zero is structural, not a matter of using a larger background.

3.2 Digit Recognition

We illustrate the proposed background construction on the handwritten-digit data set of Alpaydin and Alimoglu (1996)[3] . The data contain 1,797 grayscale 8×8 images of the digits 0–9. Pixel intensities are scaled to the unit interval and standardized before model fitting. We train a feed-forward neural network with two hidden layers (128 and 64 units) using cross-entropy loss on 70% of the observations and evaluate on the remaining 30%. The resulting classifier achieves a test accuracy of 98.15% and a test log loss of 0.0618.

Our attribution target is the complete vector of centered logits,

$$\widetilde{z}(x) = z(x) - \overline{z}(x)\mathbf{1}, \tag{34}$$

where $z(x)$ denotes the vector of model logits and $\overline{z}(x)$ their average. Centering removes the redundant location degree of freedom, leaving a $(K-1)$-dimensional prediction vector. The unconditional reference prediction is therefore

$$\widetilde{z}_0 = E[\widetilde{z}(X)]. \tag{35}$$

Rather than constructing separate backgrounds for individual classes, we construct a single background distribution near the vector-neutral manifold

$$\mathcal{M}_0 = \{x : \widetilde{z}(x) = \widetilde{z}_0\}\,. \tag{36}$$

The same background is then supplied to SHAP for every output coordinate, so the resulting explanation decomposes the complete centered-logit vector. Projecting this common vector attribution onto any class yields the corresponding class-specific heatmap. This construction avoids the inconsistencies that arise when each class is explained relative to a different background.

Observed vector-neutral backgrounds are constructed by the fixed-size sliding procedure of appendix A. Starting from the 100 training images nearest the neutral score vector, the neighborhood is translated in whitened prediction space to minimize the neutrality error while keeping the background size fixed. For comparison, we also consider four commonly used alternatives: a single black image, the mean training image, a random empirical sample of 100 training images, and a Gaussian-generated background obtained by simulating images from a fitted Gaussian pixel model and retaining the 100 simulated images nearest the neutral score vector. Unlike the proposed

[3] Available through `sklearn.datasets.load_digits`.

Table 3: Average SHAP attributions for the digit classification example

Background	n_B	$\|\bar{z}_B - z_0\|$	Compl. err	$\overline{\|\phi\|}$	$RMS(\phi)$
Panel A: Common single-point backgrounds					
Black image	1	2.1287	0.0000	0.9306	0.4837
Mean image	1	0.8300	0.0000	0.6588	0.2938
Panel B: Distributional backgrounds					
Sample	100	0.2702	0.0000	0.6042	0.2719
Sim. neutral	100	0.2788	0.0000	0.6156	0.2619
Canonical	100	0.1619	0.0000	0.6133	0.2828

The table summarizes SHAP explanations for the handwritten-digit classification example. The neural network predicts ten classes, and SHAP is computed jointly for the centered-logit vector. Reported attribution quantities are averaged over all 540 held-out test images.

Panel A reports common single-point backgrounds consisting of the all-black image and the mean training image. Panel B reports three distributional backgrounds of equal size ($n_B = 100$): a random empirical sample from the training data, a Gaussian-generated vector-neutral sample, and the proposed observed vector-neutral background.

The column $\|\bar{z}_B - z_0\|$ reports the Mahalanobis norm of the difference between the SHAP background prediction and the unconditional neutral centered-logit vector. Smaller values indicate backgrounds that more closely satisfy vector neutrality. The completeness error reports the average reconstruction error of the vector SHAP decomposition and is numerically zero for all backgrounds. The remaining columns summarize attribution magnitude: $\overline{\|\phi\|}$ is the average Euclidean norm of the per-pixel attribution vector, and $RMS(\phi)$ is the root-mean-square attribution over test images, pixels, and classes.

method, the simulated background is used exactly as generated, without further calibration or refinement.

Several conclusions emerge from Table 3. First, the choice of background has a substantial effect on the reference prediction being explained. The black-image baseline is far from neutral, with a centered-logit gap of 2.13, while the mean image remains substantially displaced (0.83).[4] Random empirical sampling reduces the neutrality gap through averaging, but the resulting background remains noticeably displaced from the unconditional reference prediction. The Gaussian-generated background performs similarly. The proposed canonical background reduces the neutrality residual to 0.16, approximately 40% below that of the random empirical sample and

[4] Izzo, Lipani, Okhrati, and Medda (2021) construct a neutral point by searching along a one-dimensional family of candidate baselines indexed by a scalar quantile parameter. For binary or scalar prediction the neutrality condition is a single equation, so a one-dimensional search is natural. In multi-class prediction, however, vector neutrality requires satisfying $K - 1$ independent score conditions simultaneously. A single scalar parameter cannot, in general, satisfy all of these constraints, so there is no obvious multi-class analogue of the Izzo, Lipani, Okhrati, and Medda (2021) construction.

Table 4: Variability of SHAP Explanations

Background	n_B	Reps.	Relative RMS	Spearman	Top-10 overlap
Panel A: Common single-point backgrounds					
Black image	1	–	0.0000	1.0000	1.0000
Mean image	1	–	0.0000	1.0000	1.0000
Panel B: Distributional backgrounds					
Sample	100	1000	0.0982	0.9581	0.9153
Sim. neutral	100	1000	0.0715	0.9748	0.9325
Canonical	100	–	0.0000	1.0000	1.0000

The table reports the conditional variability of SHAP explanations when only the background construction is repeated. The training sample, fitted neural network, explained images, SHAP coalition seed, and background size remain fixed. Variability therefore reflects only randomness in the background.

Relative RMS reports the root-mean-square change in the attribution vector across repeated backgrounds, scaled by the RMS attribution itself. Mean Spearman is the average rank correlation between all sampled pairs of repetitions. Top-10 overlap is the average fraction of the ten largest-magnitude pixel attributions shared between two repetitions.

The black-image, mean-image, and observed vector-neutral backgrounds are deterministic functions of the data and therefore produce identical SHAP explanations in every repetition. Random empirical and Gaussian backgrounds vary because a new background is generated for every run.

substantially below the Gaussian-generated benchmark. Simply selecting observations that are individually close to the neutral prediction is therefore not sufficient to construct a neutral background distribution.

Second, all backgrounds satisfy SHAP's vector completeness identity to numerical precision once the default sparse regression penalty is disabled.[5] Consequently, every background exactly decomposes the prediction difference relative to its own reference prediction. The distinction between backgrounds is therefore not mathematical but interpretational: different backgrounds answer different attribution questions. The proposed construction explains deviations from the unconditional mean prediction, whereas non-neutral backgrounds explain deviations from displaced reference predictions.

The proposed background also differs conceptually from the Gaussian-generated alternative. The latter depends on a fitted model for the feature distribution and therefore inherits any misspecification of that model. The canonical background is constructed entirely from observed images and

[5] KernelSHAP's default sparse regression (`l1_reg`) breaks the multi-class adding-up identity because it fits each output independently with different sparse supports. Disabling the sparsity penalty restores the exact vector identity.

requires no feature generator. In this example it is simultaneously more neutral and supported entirely by the empirical distribution.

Table 4 examines the conditional variability of SHAP explanations while holding the training sample, fitted neural network, explained images, SHAP coalition sampling, and background size fixed. The only source of randomness is the background construction itself. The black-image, mean-image, and canonical backgrounds are deterministic functions of the data and therefore produce identical SHAP explanations in every repetition. In contrast, the sample and Gaussian-generated backgrounds must be reconstructed for each run and consequently introduce additional Monte Carlo variability into the explanations.

The magnitude of this variability is not negligible. With a common background size of $n_B = 100$, independently generated empirical backgrounds differ by approximately 10% of the typical attribution magnitude, have an average rank correlation of 0.96, and agree on only about 92% of the ten most important pixels. Gaussian-generated backgrounds are more stable because the simulated observations are concentrated near the neutral manifold, but they still exhibit measurable variability. The canonical background eliminates this source of randomness entirely while simultaneously achieving the smallest neutrality error among the distributional backgrounds.

4 Summary

Shapley feature attributions explain a prediction relative to a reference prediction, and the background distribution determines that reference. We point out that background selection is not merely a computational detail but a key part of the definition of the attribution problem itself. Once the explanatory question is specified, the appropriate background follows from the corresponding reference prediction.

For the canonical question of why a prediction differs from the model's unconditional prediction, the natural background is the observed data distribution localized around the prediction-neutral manifold. This construction combines two desirable properties that are often treated separately. The reference prediction is meaningful because it corresponds to the chosen attribution question, and the reference inputs are realistic because they are drawn from the observed data rather than generated synthetically. The resulting background can be supplied directly to existing SHAP implementations without modifying the attribution algorithm or its axioms.

The empirical examples illustrate three consequences. First, nonlinear models generally require prediction-neutral backgrounds that differ substantially

from conventional choices such as the mean input or a random background sample. Second, observed prediction-neutral backgrounds closely reproduce the behavior of simulated neutral backgrounds when simulation is reliable, while avoiding the need to specify and estimate a high-dimensional generative model. Third, conditional on the observed sample and fitted model, the proposed construction is deterministic, eliminating the Monte Carlo variability associated with repeatedly sampling background observations.

The prediction-neutral principle extends naturally beyond the unconditional mean considered here. Decision thresholds, pairwise classification comparisons, or other externally meaningful prediction levels each define their own attribution question and therefore their own background distribution. The broader contribution of this paper is to make this dependence explicit and to show that, for a wide class of attribution questions, the corresponding background can be constructed directly from observed data while remaining fully compatible with standard Shapley attribution methods.

Existing SHAP implementations require equally weighted background observations. Allowing weighted background distributions would provide a more direct empirical approximation to prediction-neutral reference distributions while reducing the computational overhead associated with resampling or equal-weight approximations.

5 References

Albini, Emanuele, Jason Long, Danial Dervovic, and Daniele Magazzeni, 2022, Counterfactual Shapley additive explanations, in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 1054–1070 (Association for Computing Machinery, New York, NY).

Alpaydin, Ethem, and Fevzi Alimoglu, 1996, Pen-based recognition of handwritten digits, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5MG6K.

Bierens, Herman J., 1994, The Nadaraya-Watson kernel regression function estimator, in *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*, chapter 10, 212–247 (Cambridge University Press, Cambridge, England).

Chen, Hugh, Scott M. Lundberg, and Su-In Lee, 2022, Explaining a series of models by propagating Shapley values, *Nature Communications* 13 (1), 4512.

Epanechnikov, Viktor A., 1969, Non-parametric estimation of a multivariate probability density, *Theory of Probability & Its Applications* 14 (1), 153–158.

Hentschel, Ludger, 2026, Canonical Integrated Gradients: Expectations over neutral prediction baselines, Working paper, Versor Investments, New York, NY.

Izzo, Cosimo, Aldo Lipani, Ramin Okhrati, and Francesca Medda, 2021, A baseline for Shapley values in MLPs: from missingness to neutrality, in *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 605–610 (Ciaco – i6doc.com, Bruges, Belgium).

Lundberg, Scott M., and Su-In Lee, 2017, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777 (Curran Associates Inc., Red Hook, NY).

Merrick, Luke, and Ankur Taly, 2020, The explanation game: Explaining machine learning models using Shapley values, in Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, eds., *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, 17–38 (Springer, New York, NY).

Shapley, Lloyd S., 1953, A value for n-person games, *Contributions to the Theory of Games* 2, 307–317.

A Constructing Prediction-Neutral Background Samples

This appendix describes the background-sample construction used in the accompanying software. The main text requires only that the background sample consist of observed inputs whose predictions are close to neutral and whose average prediction is neutral to the desired tolerance. Many algorithms can produce such a sample. The implementation described here is designed to return a compact, unweighted background suitable for current SHAP implementations, while also describing the weighted construction that naturally approximates the prediction-neutral distribution.

A.1 Inputs and Outputs

The algorithm takes as inputs a reference feature sample $\{x_i\}_{i=1}^N$, fitted-model outputs $\{f(x_i)\}_{i=1}^N$, and a neutral output f_0. For regression, $f(x_i)$ and f_0 are scalar. For multi-class classification they are centered-logit vectors.

The algorithm returns either

1. a weighted empirical background

$$\{(x_i, w_i)\}_{i=1}^N, \tag{37}$$

   or

2. an equal-weight subset

$$S \subset \{1, \ldots, N\}, \tag{38}$$

   for use with existing SHAP implementations.

The main diagnostics are the effective background size, the average background prediction

$$\overline{f} = \sum_i w_i f(x_i) \tag{39}$$

(or the ordinary average for equal weights), together with the neutrality error

$$\|\overline{f} - f_0\|. \tag{40}$$

A.2 Localization

Define the prediction gap

$$r_i = f(x_i) - f_0. \tag{41}$$

For scalar outputs this is a scalar. For multi-class classification it is the centered-logit difference vector.

Localization is performed in prediction space using a kernel.

For scalar outputs with bandwidth h, examples include

$$K_G(r) = \exp\left(-\frac{1}{2}(r/h)^2\right), \tag{42}$$

$$K_E(r) = \left(1 - (r/h)^2\right)_+, \tag{43}$$

$$K_U(r) = \mathbf{1}\{|r| \le h\}. \tag{44}$$

For vector outputs let

$$\widetilde{z}_i = \widetilde{z}(x_i), \tag{45}$$

and define the prediction-space distance

$$d_i^2 = (\widetilde{z}_i - \widetilde{z}_0)' W (\widetilde{z}_i - \widetilde{z}_0), \tag{46}$$

where W is a positive-semidefinite weighting matrix. A natural choice is a ridge-regularized inverse covariance matrix estimated from the centered logits, so that distances are measured in standardized prediction units.

Kernel weights are then

$$w_i = \frac{K(d_i/h)}{\sum_{\ell=1}^{N} K(d_\ell/h)}. \tag{47}$$

These weights provide a direct nonparametric approximation to the localized conditional distribution

$$Q_N = P(X \mid f(X) = f_0). \tag{48}$$

If an attribution method supports weighted background distributions, no further construction is required. Current SHAP implementations instead require equally weighted background observations. The software therefore converts this weighted representation into an equal-weight approximation by selecting a compact local background whose average prediction is approximately neutral.

Bandwidth selection.
The operational parameter is the kernel bandwidth h (or equivalently the effective retained count).

For the weighted empirical distribution to converge to the localized conditional distribution Q_N, the bandwidth must shrink while the effective sample size diverges,

$$h_n \downarrow 0, \qquad n h_n^{d_\eta} \to \infty, \tag{49}$$

where d_η is the dimension of the prediction gap.

A fixed percentile does not satisfy this condition. Retaining the nearest αn observations yields a fixed-width band around the neutral manifold rather than the limiting conditional distribution. Although this approximation is often accurate in practice, it is not asymptotically exact.

One asymptotically valid choice is

$$h_n = c\, \hat{\sigma}_\eta\, n^{-1/(d_\eta+4)}, \tag{50}$$

which implies

$$m_n \asymp n^{4/(d_\eta+4)} \to \infty. \tag{51}$$

More generally any

$$h_n = c\, \hat{\sigma}_\eta n^{-\gamma}, \qquad 0 < \gamma < 1/d_\eta, \tag{52}$$

is consistent. Smaller values of γ retain more observations and trade localization for stability while preserving consistency.

For multi-class classification the prediction gap has dimension $K-1$ because centered logits satisfy one linear constraint. The higher dimension changes

the localization problem but not the attribution question. When empirical support near the neutral manifold is sparse, one may increase the bandwidth, regularize the prediction-space metric, or use adaptive bandwidths without changing the underlying prediction-neutral target distribution.

A.3 Equal-Weight Approximation

Current SHAP implementations require equally weighted background observations. Our software therefore constructs an equal-weight approximation to the weighted kernel distribution.

For scalar outputs, the procedure begins with the m observations nearest the neutral prediction and translates the fixed-size neighborhood until the average prediction is as close as possible to the neutral value.

For vector outputs, let

$$u_i = W^{1/2}(\widetilde{z}_i - \widetilde{z}_0). \tag{53}$$

Beginning with the m observations nearest the origin, the neighborhood center is translated along the direction opposite the current mean,

$$d = -\frac{\overline{u}}{\|\overline{u}\|}, \tag{54}$$

and the translation minimizing the norm of the average prediction gap is selected.

This translation preserves localization while producing an equal-weight background whose average prediction is nearly neutral.

Adaptive bandwidths

$$h_n(x) = h_n a(x) \tag{55}$$

provide a natural refinement, allowing narrower neighborhoods where empirical support is dense and wider neighborhoods where it is sparse. We do not pursue this extension here.

The algorithmic details are not part of the Shapley attribution rule. They are a practical means of approximating the prediction-neutral background distribution implied by the attribution question. Future SHAP implementations that accept weighted background distributions could use the kernel-weighted

representation directly, eliminating the equal-weight approximation and its associated computational overhead.

B Alternative Attribution Questions and Decision-Boundary Backgrounds

The prediction-neutral background developed in the main text is one instance of a more general principle. Every attribution question specifies a reference prediction, and that reference prediction determines a target manifold in prediction space. A localized background distribution is then obtained by conditioning the observed data near that manifold. The unconditional prediction used in the main text is the canonical choice, but other attribution questions lead naturally to different target manifolds.

B.1 Binary Classification

Suppose a binary classifier produces a scalar score $s(x)$, with positive predictions corresponding to $s(x) > 0$. If the attribution question is

> "Why was this observation classified as positive rather than negative?"

then the natural reference prediction is the decision boundary, and the target manifold is

$$s(x) = 0. \tag{56}$$

A kernel centered on this manifold is

$$K_h(x) = \exp\left(-\frac{s(x)^2}{2h^2}\right), \tag{57}$$

where $h > 0$ controls the width of the neighborhood. The corresponding localized empirical distribution assigns kernel weights

$$\omega_i = \frac{K_h(x_i)}{\sum_{\ell=1}^{n} K_h(x_\ell)}. \tag{58}$$

As $h \to 0$, the weighted empirical distribution concentrates on observations lying arbitrarily close to the decision boundary. When attribution software supports weighted backgrounds, these kernel weights provide the natural representation of the localized distribution. Equal-weight neighborhoods, such as those constructed in the accompanying software, provide a practical approximation for current SHAP implementations.

This background distribution is appropriate for explaining why a prediction crossed the classification threshold rather than why it differs from the unconditional mean prediction.

B.2 Multi-class Classification

For multi-class classification, there is generally no unique decision boundary. Instead, there are many pairwise decision interfaces. Let

$$\widetilde{z}(x) = \left(\widetilde{z}_1(x), \ldots, \widetilde{z}_K(x)\right) \tag{59}$$

denote the centered-logit vector, and let

$$k = \arg\max_j \widetilde{z}_j(x) \tag{60}$$

be the predicted class.

If the attribution question is

> "Why was this observation classified as class k rather than class ℓ?"

then the relevant target manifold is the interface separating the decision regions for classes k and ℓ,

$$\widetilde{z}_k(x) = \widetilde{z}_\ell(x), \tag{61}$$

$$\widetilde{z}_k(x), \widetilde{z}_\ell(x) \geq \widetilde{z}_m(x), \qquad m \neq k, \ell. \tag{62}$$

A corresponding kernel is

$$K_h(x) = \exp\left(-\frac{(\widetilde{z}_k(x) - \widetilde{z}_\ell(x))^2}{2h^2}\right) \mathbf{1}\{\widetilde{z}_k(x), \widetilde{z}_\ell(x) \geq \widetilde{z}_m(x),\ \forall m \neq k, \ell\}\,, \tag{63}$$

with associated kernel weights

$$\omega_i = \frac{K_h(x_i)}{\sum_{r=1}^n K_h(x_r)}. \tag{64}$$

As in the binary case, the localized empirical distribution concentrates on observations lying close to the relevant decision interface. Equal-weight backgrounds can again be obtained by selecting observations with the largest kernel weights, but the kernel weighting itself is the natural statistical representation of the localized background distribution.

Unlike prediction-neutral attribution, multi-class decision-boundary attribution is inherently question specific. Asking

"Why class 5 rather than class 9?"

targets a different decision interface from asking

"Why class 5 rather than class 4?"

and therefore leads to a different background distribution. There is no unique decision-boundary background for multi-class classification because there is no unique decision boundary. By contrast, the unconditional prediction defines a single reference prediction and therefore a single canonical background distribution.

The constructions in this appendix illustrate the broader principle developed throughout the paper: once the attribution question specifies the reference prediction, the corresponding background distribution follows by localizing the observed data around the associated prediction manifold.