Canonical Integrated Gradients: Expectations over Neutral Prediction Baselines

Ludger Hentschel

June 22, 2026

Abstract

Integrated gradients explain a prediction relative to a baseline. We argue that baseline selection naturally corresponds to the choice of a reference distribution over inputs. The relevant attribution question is: why does the model predict $f(x)$ rather than a neutral prediction f_0? For regression, the neutral prediction is the unconditional mean outcome; for classification, the unconditional class-probability vector. Because a classifier's additive output is its logits, we attribute on the logit scale, taking the neutral value to be the logit that yields this probability vector. All other baseline choices incorporate additional information.

Neutral predictions generally do not correspond to a unique input. Instead, neutral inputs form a level set of the prediction function. We define Canonical Integrated Gradients (CIG) as the expected integrated gradients over the data distribution on this neutral manifold. We show how to estimate the corresponding reference distribution nonparametrically by weighting observed inputs according to the proximity of their predictions to f_0 while enforcing exact neutrality. The estimator is consistent and does not require a model of the high-dimensional feature distribution.

This framework unifies fixed-baseline integrated gradients, mean-input baselines, Expected Gradients, and CIG as members of a common family of reference-distribution explanations. For calibrated linear models these approaches decompose the same total change but in general they produce different attributions.

www.ludgerhentschel.com ludger@ludgerhentschel.com

Contents

1 Introduction

Integrated gradients (IG), introduced by Sundararajan, Taly, and Yan (2017), explain a prediction by attributing the difference between the prediction at the observed input and the prediction at a baseline input across the input features. The attribution depends materially on the baseline. Yet baseline selection in practice remains largely heuristic. Common choices discussed by Sturmfels, Lundberg, and Lee (2020) – zero vectors, mean feature vectors, random noise, and all-black images – are convenient, but they are usually justified by properties of the *input* rather than by the explanatory question the attribution is intended to answer.

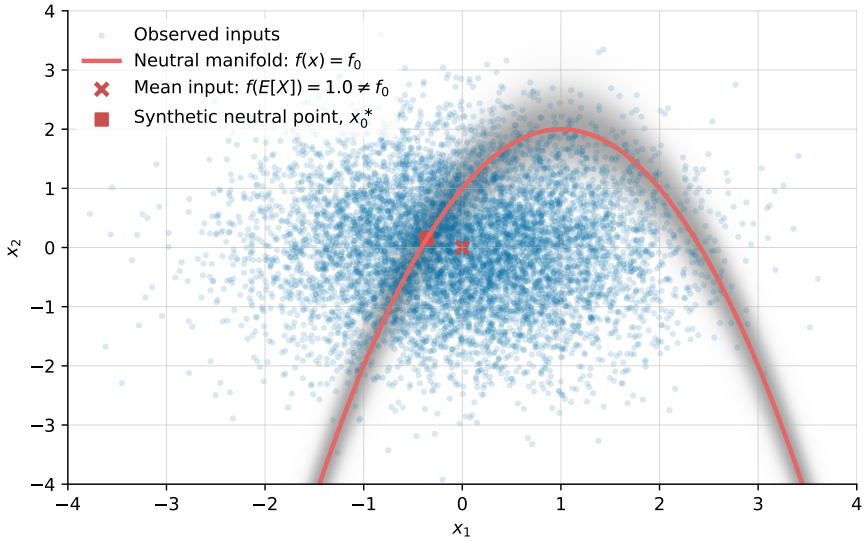
We begin from the explanatory question. A prediction explanation should answer the canonical attribution question: *why does the model predict this value rather than a neutral prediction?* For regression, the natural neutral prediction is the unconditional mean outcome, the prediction the model would make with no case-specific information. For classification, the analogous benchmark is the unconditional class distribution. More generally, the neutral prediction is the loss-optimal constant prediction for the task at hand: the benchmark prediction that carries no case-specific information. Attribution should then explain how the observed input moves the model away from this uninformed benchmark.

Once we phrase the problem this way, neutrality becomes the organizing constraint. A valid reference for integrated gradients should represent a neutral prediction. In a linear model this may correspond to a single input. In a nonlinear model it generally does not. Neutral predictions are typically produced by many different inputs, and those inputs form a level set of the prediction function rather than a single baseline point. The baseline problem is not fundamentally the choice of one baseline input; it is the choice of a *reference distribution* over inputs that the model treats as prediction-neutral.

This observation leads to a natural answer. We take the reference distribution to be the data distribution restricted to inputs that produce the neutral prediction. Attributing relative to this distribution compares the observed case to the population of realistic inputs the model itself treats as uninformative. We call the resulting method *Canonical Integrated Gradients* (CIG). In this view, the baseline problem is solved not by selecting a convenient synthetic point in input space, but by matching the reference distribution to the explanatory question.

Figure 1 illustrates the idea in a simple nonlinear regression. The red parabola is the set of inputs that produce the neutral prediction. In finite samples, observations typically lie near this set rather than exactly on it. CIG

Figure 1: Neutral Prediction Manifold



The figure illustrates simulated inputs using blue markers and a neutral prediction manifold using a red line in a synthetic regression example.

The shaded band shows points with high kernel weight $K_\tau(f(x) - f_0)$. Its Euclidean width varies because the kernel is based on prediction distance rather than geometric distance: where the model prediction changes rapidly, a small movement away from the manifold produces a large prediction gap.

The feature mean, marked with an **x** does not lie on the neutral manifold, so a mean-input baseline does not represent a neutral prediction.

assigns greater weight to observations whose predictions are closer to the neutral prediction and smaller weight to observations farther away.

A useful way to view the baseline problem is through reference distributions. Fixed-baseline IG, mean-input baselines, and Expected Gradients (Erion, Janizek, Sturmfels, Lundberg, and Lee, 2021) can all be interpreted as choosing a distribution over baseline inputs and averaging integrated gradients with respect to that distribution. Their differences arise from the choice of reference distribution, not from the attribution mechanism itself. The central question is not how to compute integrated gradients, but how to choose a reference distribution that corresponds to the explanation we seek.

We estimate this reference distribution nonparametrically. Each observed input receives a kernel weight based on how close its prediction is to the neutral prediction, and the weights are then adjusted minimally so that the weighted reference distribution is exactly neutral in finite samples. As the bandwidth shrinks with sample size, the weighted distribution converges to the target prediction-neutral reference distribution, so the resulting attribution is a consistent estimator of the canonical estimand.

This construction has several practical advantages. It applies without

modification to tabular, image, and multimodal inputs because every reference point is an observed data point. It applies to any prediction function for which integrated gradients can be computed.[1] It inherits the axiomatic properties of integrated gradients, including completeness. It concentrates computation near the neutral prediction set, which often allows accurate approximations using only a relatively small subset of the data. And because the reference distribution is determined by the explanatory question rather than chosen ad hoc by the analyst, it removes much of the discretionary baseline choice that drives the sensitivity of fixed-baseline IG.

This reference-distribution view also helps place CIG relative to the broader explanation literature. In the Shapley-value literature, a long-running debate concerns how explanations should handle realism and feature dependence. Lundberg and Lee (2017) introduced SHAP using conditional expectations, but practical implementations often replace these with marginal or interventional approximations. Janzing, Minorics, and Blöbaum (2020) defend the interventional approach, where one perturbs or scrambles features, on the grounds that explanations should be true to the model rather than to the empirical feature distribution. Aas, Jullum, and Løland (2021) and Frye, de Mijolla, Begley, Cowton, Stanley, and Feige (2021) defend conditional approaches, arguing that explanations should remain on the data manifold and respect feature dependence, even at the cost of estimating high-dimensional conditional distributions. Chen, Janizek, Lundberg, and Lee (2020) summarize the distinction as a choice between being "true to the model" and "true to the data," and Sundararajan and Najmi (2020) catalogue the different Shapley values produced by different value functions.

A similar tension appears in integrated gradients, although the IG construction always directly interrogates the fitted model through gradients along a path to the observed input. Fixed baselines and mean-input baselines use interventional point references, while Expected Gradients averages over the full empirical input distribution.

CIG highlights a third ingredient that is largely absent from this discussion: the attribution question itself. The question is not merely rhetorical; it constrains the relevant reference population. If the question is why the model predicts the observed value rather than a neutral prediction, then the relevant reference set is not an arbitrary point or an arbitrary sample from the data distribution, but the set of inputs that produce the neutral prediction. For the

[1] Hentschel (2026) show how to compute exact integrated gradients for tree-based, piecewise-constant models using distributional derivatives.

canonical question studied here, this set is the neutral-prediction manifold. We then restrict attention to the realistic part of that manifold by weighting observed inputs according to their proximity to it. In this way, the attribution question identifies the relevant prediction level set, while the data determine which points on that set are realistic and how much weight they receive. This preserves realism because every reference point is an observed input near the prediction-neutral set, but it avoids the main difficulty of conditional Shapley-style approaches: it does not require modeling the high-dimensional feature distribution. Instead, it conditions only on the prediction being near neutral, which is a low-dimensional object even when the feature space is large. Regions of higher data density along the neutral manifold are naturally represented more prominently than less realistic low-density regions.

Other attribution questions naturally imply other reference distributions. We discuss this briefly below, but our focus is the canonical neutral question.

The remainder of the paper formalizes the neutral-manifold estimand and places fixed-baseline IG, mean-input baselines, and Expected Gradients within a common family of reference-distribution explanations (section 2), develops the nonparametric estimator together with its consistency, feasibility, and bandwidth properties (section 3), and illustrates the method on synthetic, image, and tabular examples (section 4). Appendix A gives a precision-targeted algorithm whose cost is effectively independent of sample size.

2 The Neutral-Manifold Attribution

We now formalize the Canonical Integrated Gradient attribution.

2.1 Estimand

The explanatory question fixes the reference set. Attribution answers *why does the model predict* $f(x)$ *rather than* f_0*?*, so the natural reference is the distribution of inputs the model itself associates with the neutral prediction f_0. This is the data distribution conditional on predictive neutrality,

$$Q_N = P_X \mid f(X) = f_0, \tag{1}$$

the restriction of the data-generating distribution to the neutral manifold

$$\mathcal{M}_0 = \{x : f(x) = f_0\}. \tag{2}$$

This is general set and we require no assumptions about convexity, for example.

The central estimand of this note is the expected integrated gradient under this reference distribution,

$$A_{CIG}(x) = E\left[\, IG(X_0, x) \mid f(X_0) = f_0 \,\right], \tag{3}$$

where A is a k-element vector of attribution components and the expectation is over $X_0 \sim Q_N$. Equation (3) says that each feature's attribution is its average contribution to the gap $f(x) - f_0$, averaged over the population of inputs that are simultaneously realistic and neutral.

The development above is written for a scalar output f, as in regression. For a p-class classifier the model output is the vector of logits $z(x)$, defined only up to an additive constant; we work with the centered logits $\widetilde{z}(x) = z(x) - \overline{z}(x)\,\mathbf{1}$ and attribute them, since the logits are the classifier's additive output while the softmax that maps them to probabilities is a fixed, input-independent transform that adds no modeling content but saturates and couples the classes. The neutral output is $\widetilde{z}^*$, the centered-logit representative of the marginal class-probability vector p^*, and the neutral manifold and estimand carry over with $\widetilde{z}$ in place of f and $\widetilde{z}^*$ in place of f_0,

$$\mathcal{M}_0 = \{x : \widetilde{z}(x) = \widetilde{z}^*\}, \qquad A_{CIG}(x) = E[\, IG(X_0, x) \mid \widetilde{z}(X_0) = \widetilde{z}^* \,]. \tag{4}$$

Because the softmax is a bijection between centered logits and the probability simplex, $\{x : \widetilde{z}(x) = \widetilde{z}^*\} = \{x : p(x) = p^*\}$: the conditioning event, and hence the estimand, is identical whether neutrality is read in logit or in probability coordinates, and only the scale on which attributions are additive differs. Attributing the predicted-class logit, the completeness identity decomposes that logit relative to its neutral value, $\sum_i A_{CIG,i}(x) = \widetilde{z}_c(x) - \widetilde{z}_c^*$ for predicted class c, the direct analog of $f(x) - f_0$.

2.2 Completeness and the Neutral Family

We can write both the estimand (3) and the existing methods it generalizes as expected integrated gradients over a baseline distribution Q,

$$A_Q(x) = E_{X_0 \sim Q}\left[\, IG(X_0, x) \,\right]. \tag{5}$$

Fixed-baseline IG is the degenerate case $Q = \delta_{x_0}$, where the distribution collapses to a point mass at x_0; Expected Gradients use $Q = P_X$; CIG uses $Q = Q_N$.

Summing integrated gradients over features and applying the fundamental theorem of calculus gives

$$\sum_{i=1}^{k} IG_i(x_0, x) = f(x) - f(x_0) \tag{6}$$

for any starting input x_0 and ending input x.

Taking expectations over Q,

$$\sum_{i=1}^{k} A_{Q,i}(x) = f(x) - E_Q[f(X_0)]. \tag{7}$$

Define the neutral family of distributions as any distribution for which the average prediction is neutral,

$$\boldsymbol{Q}_0 = \{Q : E_Q[f(X_0)] = f_0\}. \tag{8}$$

For every $Q \in \boldsymbol{Q}_0$, equation (7) reduces to

$$\sum_{i=1}^{k} A_{Q,i}(x) = f(x) - f_0. \tag{9}$$

Every neutral reference distribution decomposes the same scalar gap, $f(x) - f_0$. The choice of Q determines how that gap is allocated across features, not what total is being decomposed. The neutral-manifold reference Q_N is a member of $\boldsymbol{Q}_0$ by construction.

A point-mass baseline $Q = \delta_{x_0}$ belongs to $\boldsymbol{Q}_0$ only if $f(x_0) = f_0$, that is, pointwise neutrality. A distribution can belong to $\boldsymbol{Q}_0$ while placing weight on observations above and below f_0, provided they average to f_0. This may be true for EG although it is not generally enforced. The neutral-manifold reference Q_N achieves the strongest form: it concentrates only on observations where $f(x_0) = f_0$ exactly.

2.3 A Single Inspectable Baseline

The family $\boldsymbol{Q}_0$ spans a spectrum. At one end is the conditional distribution Q_N, which spreads weight across all neutral inputs; at the other is a point mass $Q = \delta_{x_0}$, which collapses the reference to one input. The neutral-manifold estimand sits at the distributional end and is the default throughout. Occasionally, though, one must exhibit a single inspectable reference rather than a distribution (for an audit, a regulatory record, or a side-by-side

counterfactual) and the question is how to collapse the reference to one point without abandoning neutrality.

Two natural single points appear in Figure 1, and they behave differently. The first is the mean neutral input,

$$\mu_N = E_{Q_N}[X_0], \tag{10}$$

estimated in the figure by the weighted mean of the neutral references. Because f is nonlinear, μ_N is generally *not* neutral: averaging inputs that lie on the curved manifold lands off it, so $f(\mu_N) \neq f_0$—the same Jensen gap that displaces the mean-input baseline. An integrated gradient from μ_N decomposes $f(x) - f(\mu_N)$ rather than $f(x) - f_0$.

The second restores neutrality by projection. Among inputs that predict neutrally, take the most typical,

$$x_0^* = \arg\min_{x_0} \|x_0 - E[X]\|^2 \quad \text{subject to} \quad f(x_0) = f_0, \tag{11}$$

the closest synthetic neutral point. This trades typicality against neutrality, conceding only as much proximity to the data centroid as exact neutrality requires. Since $f(x_0^*) = f_0$ by construction, an integrated gradient from x_0^* is complete and sums to $f(x) - f_0$.

Both are single synthetic points, and both inherit the limitation of any point baseline. The neutral point x_0^* lies on $\mathcal{M}_0$ but possibly in a low-density region between observations, so the straight-line path from x_0^* to x may cross input regions the model never saw. This corresponds to the realism objection we raise against the mean-input and all-black baselines. This is the reason the estimand averages attributions rather than collapsing to a point: because $IG(\cdot, x)$ is nonlinear in the baseline,

$$A_{CIG}(x) = E_{Q_N}\big[IG(X_0, x)\big] \;\neq\; IG\big(\mu_N,\, x\big), \tag{12}$$

and only the left-hand side averages over *observed* neutral inputs and respects their joint distribution on $\mathcal{M}_0$. The single inspectable baseline x_0^* is therefore the reserved exception, appropriate when one reference must be shown and we are willing to accept the cost of low realism, while the neutral-manifold distribution remains the default.

2.4 Alternative Baselines as Estimators of CIG

Taking A_{CIG} as the estimand turns every other baseline distribution into a competing estimator of the same target, which we can assess by bias

and variance. Expected Gradients (Erion, Janizek, Sturmfels, Lundberg, and Lee, 2021) use $Q = P_X$, the full marginal distribution of inputs; the mean-input baseline uses the point mass $Q = \delta_{E[X]}$ and we write $A_0(x) = IG(E[X], x)$. These choices answer different questions – Expected Gradients explains how $f(x)$ differs from the prediction at a randomly drawn input, the mean-input baseline how it differs from the prediction at the average input – and neither is the canonical question.

As the number of baselines grows, the Expected-Gradients estimator converges to $A_{EG}(x) = E_{P_X}[IG(X_0, x)]$, not to $A_{CIG}(x)$, leaving the asymptotic bias

$$b_{EG}(x) = A_{EG}(x) - A_{CIG}(x) = E_{P_X}[IG(X_0, x)] - E_{Q_N}[IG(X_0, x)], \qquad (13)$$

which does not shrink with sample size. Expected Gradients is consistent for A_{EG} but inconsistent for A_{CIG}, except when P_X and Q_N induce the same expected integrated gradient.[2] The bias has a revealing structure. Summing over features and using completeness (7),

$$\sum_i^k b_{Q,i}(x) = \big(f(x) - E_Q[f(X_0)]\big) - \big(f(x) - f_0\big) = f_0 - E_Q[f(X_0)], \quad (14)$$

so the *total* bias of any baseline distribution relative to A_{CIG} equals its neutrality gap. For a neutral reference, such as Expected Gradients when $E_{P_X}[f(X)] = f_0$, the total bias is zero and the bias is a pure reallocation of a correctly sized gap across features. For the mean-input baseline the total bias is $f_0 - f(E[X])$, the Jensen gap, so it misstates the size of the gap, not only its allocation. Both baselines are inconsistent for A_{CIG}, but the mean-input error is qualitative (wrong gap) and the Expected-Gradients error is quantitative (right gap, wrong allocation).

When the bias is small—as it is in some of our examples—the remaining difference is efficiency. Each method estimates its target by averaging integrated gradients over sampled baselines, and an m-baseline estimate has variance $V_Q(x)/m$ with $V_Q(x) = \mathrm{Var}_{X_0 \sim Q}[IG(X_0, x)]$, so under approximate unbiasedness the ratio V_{EG}/V_{CIG} is the factor by which Expected Gradients needs more baselines to match the precision of CIG. The paths entering A_{EG} start from all training inputs, most far from the neutral manifold; those entering A_{CIG} start from near-neutral inputs and are more homogeneous, so V_{CIG} is smaller in our examples, where we quantify the gap (section 4). The

[2] For linear f this requires $E_{Q_N}[X_0] = E[X]$, which holds for Gaussian features but not in general.

mean-squared error of any baseline as an estimator of A_{CIG} thus decomposes as $\|b_Q(x)\|^2 + V_Q(x)/m$: the mean-input baseline can carry a wrong-gap bias; Expected Gradients can carry an allocation bias; CIG targets the neutral-manifold estimand directly and is the consistent reference with generally lower-variance.

2.5 Aggregation

In addition to attributing feature importance for individual predictions, we often aggregate attributions across predictions. For any fixed reference distribution Q, population summaries of the individual attributions are averages of those attributions,

$$\frac{1}{n}\sum_{i=1}^{n} A_Q(x_i) = E_{X_0\sim Q}\left[\frac{1}{n}\sum_{i=1}^{n} IG(X_0, x_i)\right]. \tag{15}$$

Thus, when we average signed attributions across observations, the aggregation commutes with the expectation over baselines: the population attribution is simply the average of the individual attributions computed using the same reference distribution Q.

Many empirical feature-importance summaries instead average absolute attributions, for example

$$\frac{1}{n}\sum_{i=1}^{n} |A_{Q,j}(x_i)|. \tag{16}$$

This quantity measures the average *magnitude* of feature j's contribution, rather than its average directional contribution. Because the absolute value is nonlinear, it does not generally commute with the baseline expectation

$$\frac{1}{n}\sum_{i=1}^{n} |A_{Q,j}(x_i)| = \frac{1}{n}\sum_{i=1}^{n} \left|E_{X_0\sim Q}[IG_j(X_0, x_i)]\right| \tag{17}$$

$$\neq E_{X_0\sim Q}\left[\frac{1}{n}\sum_{i=1}^{n} |IG_j(X_0, x_i)|\right]. \tag{18}$$

The first is the mean absolute value of the *CIG attribution itself*; the second would define a different estimand that averages absolute single-baseline attributions before averaging over baselines. Throughout the empirical examples we use the former, so all feature-importance summaries remain summaries of the same underlying attribution $A_Q(x)$ rather than of a different absolute-value estimand.

2.6 Contextual Attribution Questions

The neutral-manifold distribution Q_N is the reference distribution implied by the canonical attribution question of this note: why does the prediction differ from a neutral benchmark? Other attribution questions generally imply other reference distributions.

If the question is why a credit application was rejected rather than accepted, a natural reference population is the set of accepted applications. If the question is why a newborn's predicted adult height differs from the benchmark for girls, the natural reference distribution conditions on female observations and uses the corresponding subgroup-neutral prediction as the benchmark. More generally, one may ask why a prediction differs from a benchmark within a subpopulation defined by sex, age, geography, diagnosis, or any other information the analyst wishes to hold fixed. In each case the attribution takes the same form, $A_Q(x) = E_{X_0 \sim Q}[IG(X_0, x)]$, but the choice of Q changes because the explanatory question changes. We can view these alternative attribution questions as contextual rather than canonical because they incorporate additional information into the benchmark. The contribution of this note is not to enumerate all such questions, but to identify the canonical neutral question and its associated reference distribution.

3 Estimation on the Neutral Manifold

3.1 Kernel Weights

For continuous data, the event $f(X_0) = f_0$ has probability zero and we must estimate Q_N by local conditioning. Let $r_i = f(x_i) - f_0$ denote the prediction gap for each training observation; for a p-class classifier the gap is the vector $r_i = \widetilde{z}(x_i) - \widetilde{z}^*$ of centered-logit deviations from neutrality, of dimension $q = p - 1$. A kernel approximation to Q_N assigns preliminary weights

$$q_i = \frac{K_\tau(r_i)}{\sum_j K_\tau(r_j)}, \qquad K_\tau(r) = \exp\left\{-\tfrac{1}{2}\|r/\tau\|^2\right\}, \tag{19}$$

weighting observations by their proximity to the neutral manifold; for scalar output $\|r/\tau\|^2 = (r/\tau)^2$. Under standard bandwidth conditions ($\tau \to 0$ and $n\tau^q \to \infty$ as n grows) the weighted distribution converges to Q_N and the resulting weighted attribution is a consistent estimator of $A_{CIG}(x)$ in (3).

3.2 Exact Neutrality

The kernel weights q concentrate near $\mathcal{M}_0$ but their weighted prediction $\sum_i q_i f(x_i)$ need not equal f_0 exactly in finite samples. We project q to the

nearest exactly neutral distribution by minimizing the χ^2 divergence subject to the neutrality constraint,

$$w^* = \arg\min_w \sum_i \frac{(w_i - q_i)^2}{q_i} \qquad \text{subject to} \quad \sum_i w_i = 1, \quad \sum_i w_i r_i = 0, \quad w_i \geq 0. \tag{20}$$

Using $\sum_i (w_i - q_i)^2/q_i = \sum_i w_i^2/q_i - 1$, this is equivalent to minimizing $\sum_i w_i^2 \exp\{\frac{1}{2}(r_i/\tau)^2\}$ subject to the same constraints. The slightly relaxed problem is a convex quadratic program with two linear equality constraints and has an analytic solution. Let $\mu_q = \sum_i q_i r_i$ and $\sigma_q^2 = \sum_i q_i (r_i - \mu_q)^2$. The unconstrained (before imposing $w_i \geq 0$) solution is

$$w_i^* = q_i \left(1 - \frac{\mu_q (r_i - \mu_q)}{\sigma_q^2}\right). \tag{21}$$

Here, μ_q is the mean distance from prediction neutrality.

When q is already neutral ($\mu_q = 0$), the correction vanishes and $w^* = q$. The correction adjusts weight toward observations on the opposite side of f_0 from the weighted average deviation, and its magnitude shrinks as $\tau \to 0$ because q itself approaches neutrality in the limit.

For a vector output the projection generalizes without change of form. With $r_i \in \mathbb{R}^p$ the neutrality constraint $\sum_i w_i r_i = 0$ is p linear equalities; writing $\mu_q = \sum_i q_i r_i$ and $\Sigma_q = \sum_i q_i (r_i - \mu_q)(r_i - \mu_q)^\mathsf{T}$ for the weighted mean gap and its covariance, the analytic solution is

$$w_i^* = q_i \left(1 - \mu_q^\mathsf{T} \Sigma_q^{-1} (r_i - \mu_q)\right), \tag{22}$$

which recovers (**??**) when $p = 1$. It requires solving a single $p \times p$ linear system, negligible in the output dimension and independent of the number of features.

After applying the closed form (21), we clip the weights at zero and renormalize. Negative weights arise only when τ is very small and the near-neutral observations are insufficient to span f_0, which coincides with the infeasibility condition below.

The Canonical Integrated Gradients attribution estimate is

$$\widehat{A}_{CIG}(x) = \sum_i w_i^* \, IG(x_i, x). \tag{23}$$

3.3 Feasibility

A neutral empirical reference distribution exists if and only if f_0 lies in the convex hull of the observed predictions. For scalar regression this reduces to

$$\min_i f(x_i) \le f_0 \le \max_i f(x_i). \tag{24}$$

If (24) fails, the fitted model never produces the neutral prediction on the observed sample, which is a diagnostic of calibration or distribution shift rather than a baseline to be forced.

3.4 Bandwidth Selection

The kernel weights estimate the conditional distribution $P(X \mid f(X) = f_0)$ by smoothing in the prediction gap $r = f(X) - f_0$. This gap is scalar for scalar regression and, for p-class classification, the $(p-1)$-dimensional vector of centered-logit deviations from neutrality; we write q for its dimension. The neutral-manifold attribution $\widehat{A}_{CIG}(x)$ is a Nadaraya–Watson type estimator of the conditional mean $E[\, IG(X_0, x) \mid r(X_0) = 0\,]$. (See Nadaraya (1964), Watson (1964), or Bierens (1994) for details.) Because the smoothing variable is the q-dimensional output gap rather than the feature vector, the curse of dimensionality is governed by the number of outputs q, not by the feature dimension; q is generically far smaller than the feature dimension and equals one for scalar regression. Under conventional smoothness conditions on the conditional mean $r \mapsto E[\, IG(X_0, x) \mid r\,]$ and the density of r near zero, the leading bias is of order τ^2, the variance is of order $(n\tau^q)^{-1}$, and the mean-squared-error optimal bandwidth satisfies $\tau \asymp n^{-1/(q+4)}$. The consistency conditions $\tau \to 0$ and $n\tau^q \to \infty$ are satisfied along any sequence with this rate. For scalar regression ($q = 1$) this recovers the standard one-dimensional rate $\tau \asymp n^{-1/5}$; for higher-dimensional outputs the rate slows, so localization is correspondingly weaker at fixed n.

Under conventional smoothness conditions on the map $x_0 \mapsto IG(x_0, x)$ and the density of r near zero, the leading bias is of order τ^2, the variance is of order $(n\tau)^{-1}$, and the mean-squared-error optimal bandwidth satisfies $\tau \asymp n^{-1/5}$. The consistency conditions $\tau \to 0$ and $n\tau \to \infty$ are satisfied along any sequence with this rate.

The chi-square projection that restores exact neutrality in finite samples is not a source of asymptotic bias at this rate. It is a calibration step: when the kernel weights are only approximately neutral, as they inevitably are for any finite τ, the projection adjusts q to satisfy $\sum_i w_i^* r_i = 0$ exactly. This ensures exact completeness $\sum_i^k \widehat{A}_{N,i}(x) = f(x) - f_0$ without altering the asymptotic order of the estimator.

The standard rate $\tau \asymp n^{-1/5}$ guides the direction of shrinkage but does not determine a practical bandwidth. The multiplicative constant depends on unknown curvature and variance terms for the IG response surface, and may be large relative to the locality needed for stable attribution in a given application. We use the asymptotic conditions as a consistency requirement rather than a literal prescription.

In practice, we select the bandwidth by an empirical locality rule: We set τ to a low quantile of $\{|r_i|\}$, such as the fifth percentile. This makes the bandwidth adaptive to the scale of prediction gaps in the sample and requires no analyst input. A low quantile concentrates weight on the most nearly neutral observations while retaining support on both sides of f_0, which is necessary for the feasibility condition $\min_i f(x_i) \le f_0 \le \max_i f(x_i)$ to bind within the effective support. We also report the effective number of baselines $m^* = 1/\sum_i (w_i^*)^2$ alongside the attribution as a diagnostic of how local the resulting reference distribution is. A small m^* relative to the sample size confirms that weight is concentrated near the neutral manifold, as intended.

3.5 Computational Cost

By construction, the kernel weights decline as predictions move away from neutrality. In practice, this generally produces a weight distribution with a large right tail of small weights. Appendix A shows that we can clip this tail and materially reduce the number of reference points in a manner that sacrifices only a small degree of precision.[3] In the empirical examples that follow, we show that canonical integrated gradients often retain very high precision when we use only 10 percent, or fewer, of the sample observations as baselines. If we evaluate integrated gradients starting at 10% of the inputs, the computations are roughly 10 times faster than for Expected Gradients, since the baseline selection step is computationally cheap compared to integrated gradient computations.

4 Empirical Examples

We use three examples, chosen to span the regimes the framework predicts. First, we work with a synthetic regression that admits closed-form integrated gradients and isolates the concentrated case, in which the neutral manifold lies far from the data mode and baseline choice matters by construction. Second, we analyze digit images with a genuinely nonlinear classifier in which neutrality is broad, so reasonable distributional baselines largely agree

[3] An alternative approach is to use a compact-support kernel, such as the Epanechnikov (1969) kernel. We use a Gaussian kernel because it provides a smooth weighting scheme and yields a particularly simple comparison with Expected Gradients.

in aggregate while still differing on individual predictions. Here, the all-black image is an atypical point baseline and the associated attributions diverge sharply. Third, a regression model for the Ames housing data is a near-linear case in which all reasonable baselines broadly agree, but where the neutral reference still offers computational and diagnostic advantages.

4.1 Synthetic Nonlinear Regression

We use a simple nonlinear regression for which integrated gradients are available analytically, making the comparison exact rather than approximate. Let

$$f(x) = a + (x_1 - c)^2 + b\, x_2, \qquad X_1, X_2 \overset{iid}{\sim} N(0,1). \tag{25}$$

We set $a = 0$, $c = 1$, $b = 1$ throughout. The neutral prediction is

$$f_0 = E[f(X)] = a + 1 + c^2 = 2, \tag{26}$$

while the mean-input prediction is

$$f(E[X]) = a + c^2 = 1. \tag{27}$$

The gap $f_0 - f(E[X]) = 1$ arises entirely from Jensen's inequality applied to the squared term: the mean input predicts one unit below the neutral level because the model is nonlinear. The conventional mean-input baseline explains $f(x) - 1$ rather than $f(x) - f_0 = f(x) - 2$.

The neutral manifold $\mathcal{M}_0 = \{x : f(x) = 2\}$ is the one-dimensional curve $(x_1 - 1)^2 + x_2 = 2$ in the feature plane. Figure 1 illustrates this manifold. Typical data points sit near the mean $(0, 0)$, well away from the right half of the manifold. Observations on or near $\mathcal{M}_0$ tend to have elevated x_2 because the joint density concentrates where both features are moderate: reaching the neutral level with x_1 near zero requires $x_2 \approx 1$.

We compare three reference distributions. The *mean-input baseline* uses the single point $E[X] = (0, 0)$. *Expected Gradients* weights all training observations equally; it is approximately neutral because $E[f(X)] = f_0$. The *neutral-manifold attribution* applies the kernel weights and chi-square projection of Section 3, which concentrates mass on observations near $\mathcal{M}_0$ under the joint data density.

Table 1 reports attribution summaries over a large target sample. The mean-input baseline confirms the Jensen effect: its mean total attribution is approximately 1, not 0, reflecting that it explains a different gap than the

Table 1: Average attributions for synthetic regression example

Method	Mean signed x_1	Mean signed x_2	Mean absolute x_1	Mean absolute x_2	Eff. baselines $1/\sum_i (w_i^*)^2$
Mean-input baseline	1.00	0.00	1.64	0.80	1
EG	0.00	0.00	1.78	0.80	1,000,000
CIG	0.44	−0.44	1.68	0.87	87,756

The table reports simulation results for the synthetic regression described in the text, $f(x) = (x_1 - 1)^2 + x_2$, with standard normal features and neutral level $f_0 = 2$.

The mean signed and mean absolute attributions are averaged over 1 million target observations drawn from the same distribution. The mean-input baseline explains $f(x) - f(E[X]) = f(x) - 1$; the other two methods explain $f(x) - f_0 = f(x) - 2$.

The effective number of baselines equals the inverse sum of squared kernel weights. For the mean-input baseline it is 1 by construction; for Expected Gradients it is $n = 10^6$; for Canonical Integrated Gradients it is determined by the kernel bandwidth and the density of the data near the prediction-neutral manifold $\mathcal{M}_0$.

other two methods. Expected Gradients and the neutral-manifold method both explain $f(x) - f_0$ correctly, so their mean total attributions are both approximately zero.

The difference between them lies in how they split that total across features. Expected Gradients allocate nearly zero mean attribution to each feature individually, which reflects the symmetry of the marginal distribution: $E[X_2] = 0$ and $E[(X_1 - 1)^2]$ is the same above and below f_0. The neutral-manifold method attributes a mean of approximately +0.44 to x_1 and −0.44 to x_2. This reflects the level-set geometry: observations near $\mathcal{M}_0$ have elevated x_2 relative to typical targets, so a typical target has lower x_2 than the neutral reference requires, contributing negatively, and higher $(x_1 - 1)^2$ than the neutral reference, contributing positively. These signed attributions describe genuine structure in how the model departs from neutrality; they are not an artifact of the weighting.

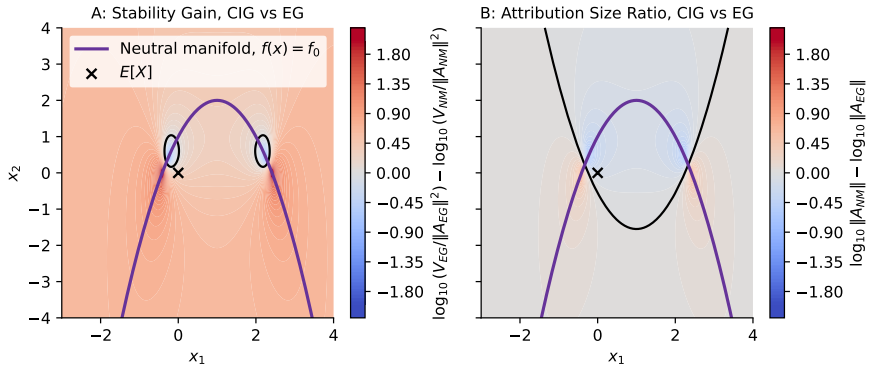
The table also reports the effective number of baselines for the neutral-manifold estimator; it is a small fraction of the full sample, confirming that the reference distribution concentrates.

Figure 2 plots two diagnostics across the feature plane for the synthetic regression in equation (25), with the neutral manifold $(x_1 - 1)^2 + x_2 = 2$ plotted as the purple curve and the mean input $E[X] = (0, 0)$ marked by a cross.

Figure 2 shows that expected gradients have roughly 2-5 times the baseline sensitivity of our approach. The exceptions occur in small areas of feature space where either approach produces attributions with vanishing magnitudes.

Panel A shows the normalized stability gain of the Canonical Integrated

Figure 2: Relative Precision



Panel A shows $\log_{10}(V_{EG}/\|A_{EG}\|^2) - \log_{10}(V_{NM}/\|A_{NM}\|^2)$, the relative precision of CIG and EG. CIG is materially more stable than EG over most of the input plane.

Panel B shows $\log_{10}\|A_{NM}\| - \log_{10}\|A_{EG}\|$, the relative attribution magnitudes. The attribution magnitudes are very similar over the entire input plane.

The purple line indicates the manifold of inputs corresponding to a neutral prediction f_0. The black **x** marks the average input, which does not produce a neutral prediction. Above the black parabola in panel B EG has slightly larger attributions than CIG; below the black parabola the reverse is true.

Gradients (CIG) attribution relative to Expected Gradients (EG), which we define as the ratio of baseline-selection variances $(V_{EG}/\|A_{EG}\|^2)/(V_{CIG}/\|A_{CIG}\|^2)$ displayed on a log scale. Here, V_{EG} is the variance of the EG attribution induced by the random sample baseline selection from the reference distribution, and $\|A_{EG}\|$ is the Euclidean norm of the corresponding mean EG attribution. Analogously, V_{CIG} is the variance induced by sampling from the neutral-manifold baseline distribution and $\|A_{CIG}\|$ is the norm of the corresponding mean CIG attribution. The numerator and denominator measure attribution variance per unit of attribution signal. The ratio measures the relative signal-to-noise efficiency of the two attribution schemes. Log values above zero indicate that CIG achieves a larger attribution signal for a given level of baseline-induced variability, while log values below zero indicate that EG is more stable after normalizing for attribution magnitude.

Most of the feature plane is associated with positive stability gains, indicating that CIG is substantially less sensitive to baseline variation than EG after accounting for attribution magnitude. This is shown in warm, reddish colors. Over much of the support of the data-generating distribution the gain is between approximately two and four, implying that EG exhibits roughly two to four times as much baseline-induced variation per unit of attribution signal.[4]

[4] The only exceptions are small neighborhoods where either attribution norm approaches zero. In these regions the normalization dominates the ratio, producing localized pockets in which the stability ordering reverses. Because these points correspond to nearly vanishing

Because f is quadratic in x_1 and linear in x_2, the straight-line integrated gradient is available in closed form. Writing $x_0 = (x_{0,1}, x_{0,2})$ for the baseline,

$$IG_1(x_0, x) = (x_1 - 1)^2 - (x_{0,1} - 1)^2, \qquad IG_2(x_0, x) = x_2 - x_{0,2}. \tag{28}$$

The baseline enters each coordinate only through a baseline-only term, so the baseline-selection variance is identical for every target and decomposes as $V_Q = \mathrm{Var}_Q[(X_{0,1} - 1)^2] + \mathrm{Var}_Q[X_{0,2}]$. For Expected Gradients, $\mathrm{Var}_{P_X}[(X_1 - 1)^2] = 6$ and $\mathrm{Var}_{P_X}[X_2] = 1$, so $V_{EG} = 7$. For the neutral reference, conditioning on $f = f_0$ forces $X_{0,2} = 2 - (X_{0,1} - 1)^2$, so the two coordinate variances coincide and $V_{CIG} = 2\,\mathrm{Var}_{Q_N}[(X_{0,1} - 1)^2] = 1.83$. The baseline-selection variance ratio is exact and target-independent,

$$\frac{V_{EG}}{V_{CIG}} = \frac{7}{2\,\mathrm{Var}_{Q_N}[(X_1 - 1)^2]} = 3.83. \tag{29}$$

Before any normalization by attribution magnitude, EG thus has nearly four times the baseline-induced variance of the neutral reference. The finite-bandwidth kernel estimator underlying Figure 2 returns 3.77; the small shortfall reflects the nonzero bandwidth, which admits slightly off-neutral baselines and inflates V_{CIG}, and the estimate converges to 3.83 as $\tau \to 0$.

Panel A combines this constant with the target-specific attribution magnitudes,

$$\log_{10}\left[\frac{V_{EG}/\|A_{EG}\|^2}{V_{CIG}/\|A_{CIG}\|^2}\right] = \log_{10}(3.77) + 2\log_{10}\left(\frac{\|A_{CIG}\|}{\|A_{EG}\|}\right). \tag{30}$$

The constant term $\log_{10}(3.77) \approx 0.58$, the figure's finite-bandwidth estimate of $\log_{10}(3.83)$, shifts the plane in favor of CIG throughout, while the second term explains the local departures from that average advantage. In more general nonlinear models both effects may vary across the feature space.

The same closed form makes the bias exact. Because IG depends on the baseline only through $E_Q[(X_{0,1} - 1)^2]$ and $E_Q[X_{0,2}]$, each method's attribution is a fixed offset from A_{CIG}, constant across targets. These offsets are exactly the differences in mean signed attribution in Table 1,

$$b_{EG} = A_{EG} - A_{CIG} = (-0.44,\ +0.44), \tag{31}$$

attributions, they have limited practical significance.

and

$$b_0 = A_0 - A_{CIG} = (+0.56,\ +0.44). \tag{32}$$

Expected Gradients is off by a vector summing to zero—it allocates the correct total differently—while the mean-input baseline is off by a vector summing to one, the Jensen gap, decomposing $f(x) - 1$ in place of $f(x) - 2$. Neither offset vanishes with sample size, so both are inconsistent estimators of A_{CIG}; the synthetic regression exhibits the inconsistency in closed form and shows its two forms, wrong gap and wrong allocation, side by side. In general nonlinear models these biases are target-dependent; the additive structure here makes them constant.

Panel B shows the attribution magnitude ratio $\|A_{CIG}\|/\|A_{EG}\|$ on the same log scale. The gray color for most of the surface indicates that the attribution magnitudes are generally similar.

The black zero-contour divides the feature plane into regions where CIG produces slightly larger attribution magnitudes than EG (warm colors in the lower part) and regions where EG produces slightly larger attribution magnitudes than CIG (cool colors in the upper part). The pattern reflects the shift in reference moments between the two methods. CIG concentrates weight on observations whose predictions are close to the neutral prediction f_0, which under the joint data distribution tend to have elevated values of x_2 and reduced values of $(x_1 - 1)^2$ relative to the unconditional means used by EG. Consequently, target observations with x_2 below the CIG reference level receive larger x_2 attributions under CIG than under EG, while those above receive smaller attributions. Similar logic applies to the nonlinear x_1 component. The black contour marks the locus where these effects balance exactly in norm.

4.2 Digit Multi-Class Image Classification

We next consider a multiclass image-classification problem using the handwritten digit data set from `sklearn.datasets.load_digits` (Alpaydin and Alimoglu, 1996). The data consist of 1,797 grayscale images representing the digits 0 through 9. Each image contains 8×8 pixels. Pixel intensities are scaled to the unit interval and standardized before model fitting. We train a feed-forward neural network with two hidden layers of sizes 128 and 64 using cross-entropy loss on 70% of the sample and then evaluate and attribute the model on the 540 hold-out images. The resulting classifier achieves a test accuracy of 98.1% and a test log loss of 0.062.

Unlike the regression examples, the model output is the vector of logits, and we attribute on the logit scale and define neutrality there (Section **??**). Since the logits are defined only up to an additive constant, let $z(x)$ denote the model logits and

$$\widetilde{z}(x) = z(x) - \overline{z}(x)\,\mathbf{1} \tag{33}$$

the centered logits. Let

$$p^* = E[p(X)] \tag{34}$$

be the empirical mean of the model's class-probability vector $p(X)$—the class-frequency vector—and $\widetilde{z}^*$ its centered-logit representative. We require the neutral reference distribution to satisfy

$$E_w[\widetilde{z}(X)] = \widetilde{z}^*, \tag{35}$$

so the kernel weights are based on Euclidean distances between $\widetilde{z}(x)$ and $\widetilde{z}^*$, and the chi-square projection of Section **??** restores this equality exactly in finite samples. Completeness then decomposes the predicted-class logit relative to its neutral value,

$$\sum_i A_{CIG,i}(x) = \widetilde{z}_c(x) - \widetilde{z}_c^* \tag{36}$$

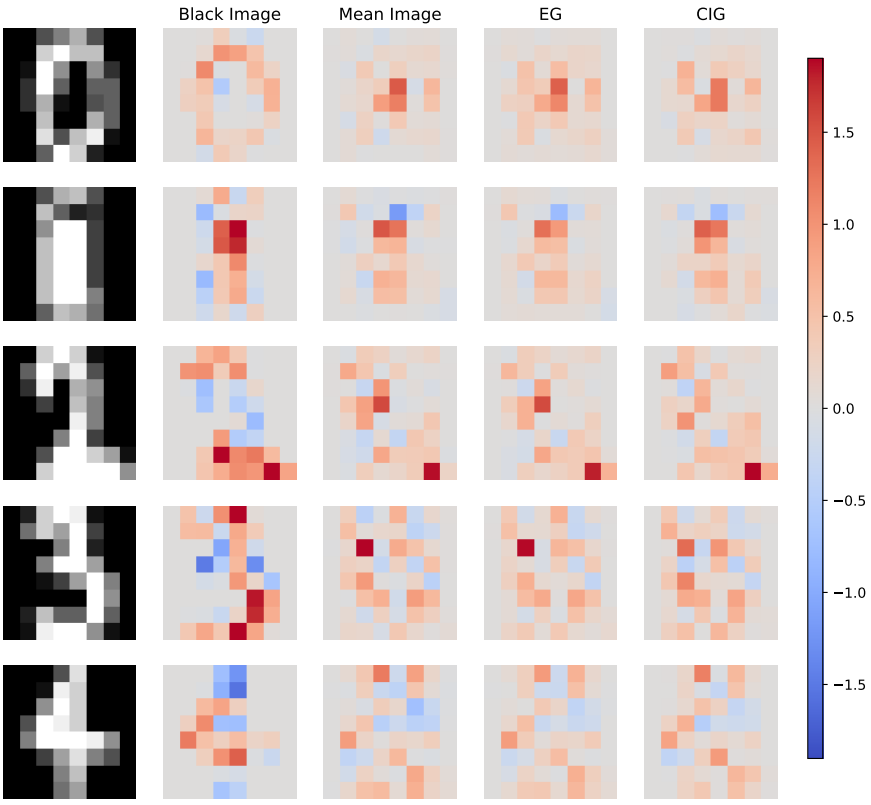
for predicted class c. Attributing logits rather than class probabilities preserves the additive structure of integrated gradients and avoids the saturation and cross-class coupling introduced by the softmax transform.

We compare four attribution methods:

1. Integrated gradients with an all-black image baseline.
2. Integrated gradients with the mean training image as the baseline.
3. Expected Gradients (EG).
4. Canonical Integrated Gradients (CIG) using neutral-manifold weighting.

The neutral weighting procedure identifies a prediction-neutral reference set that remains substantial but is no longer close to the full image distribution. Under the centered-logit neutrality criterion, the full neutral reference distribution has an effective sample size of approximately 341 observations out of 1,257 training images. Retaining only the dominant support points yields a sparse neutral reference with 23 baseline images and an effective

Figure 3: Digit Attribution Examples



The figure illustrates integrated gradient attributions for 5 sample digits using different baselines. The left column shows the input image used for attribution. The column labeled "black image" shows attributions using a black image baseline. The column labeled "mean image" shows attributions using an average image baseline. The column labeled "EG" shows Expected Gradient attributions that equally average attributions starting from all sample images. The column labeled "CIG" shows Canonical Integrated Gradients attributions based on weighted average attributions close to the image manifold associated with neutral predictions.

sample size of approximately 14 while preserving neutrality to numerical precision. Because CIG needs integrated gradients only from these retained baseline images, the sparse implementation reduces the baseline-attribution computations to approximately $23/1{,}257 \approx 2\%$ of those required by full EG.

Figure 3 illustrates the attribution results for 5 example digits using 4 different baselines: an all-black image, the mean image, Expected Gradients (EG), and Canonical Integrated Gradients (CIG). The examples suggest a clear qualitative pattern. The all-black image baseline often produces visibly different saliency maps, while the mean-image, EG, and CIG explanations are usually aligned on the main strokes of the digit, though they need not agree exactly on which pixels receive the largest attribution.

Table 2 summarizes the aggregate feature-importance results. At the level

Table 2: Feature Importance Correlations

Comparison	Pearson	Spearman
Black image vs. Mean image	0.765	0.840
Black image vs. EG	0.751	0.836
Black image vs. CIG	0.754	0.833
Mean image vs. EG	0.995	0.995
Mean image vs. CIG	0.990	0.986
EG vs. CIG	0.989	0.989

The table reports correlations between aggregate pixel-importance rankings under different integrated-gradient baseline choices for the `sklearn` digits multi-class classification problem. For each method, pixel importance is measured by mean absolute attribution across 540 test images, and the table reports Pearson and Spearman correlations between the resulting 64-dimensional pixel-importance vectors.

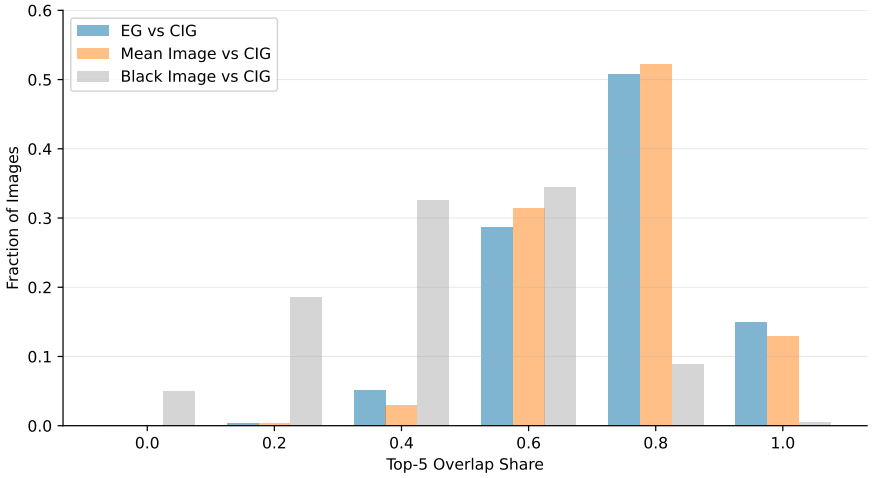
The black-image method uses an all-black image baseline. The mean-image method uses the average training image as the baseline. Expected Gradients (EG) averages attributions over empirical baseline images. Canonical Integrated Gradients (CIG) uses weighted average attributions over images close to the centered-logit neutral manifold.

of average absolute pixel importance, the two distributional baselines, EG and CIG, together with the mean-image point baseline remain close, but the agreement is no longer nearly exact. The Pearson correlation is 0.989 between EG and CIG and 0.990 between the mean-image baseline and CIG, while the correlation between the all-black baseline and CIG is only about 0.75. Thus the main separation is still between the black-image baseline and the remaining methods, but within the set of realistic baselines there is now some visible dispersion. In particular, the mean-image baseline is no longer obviously farther from CIG than EG is; both remain highly correlated with CIG in aggregate, but neither is effectively identical to it.

Aggregate feature-importance rankings, however, average over both pixels and observations and can mask meaningful differences at the level of individual predictions. To assess those differences, we compare the attribution vectors image by image.

Figure 4 summarizes the overlap of the top 5 attributed pixels with CIG. The black-image baseline is clearly the outlier: it typically shares only about 2 or 3 of the top 5 pixels with CIG. The mean-image and EG baselines are much closer, but still far from identical. Their average top-5 overlap with CIG is approximately 0.75, so they agree on about 4 of the 5 most important pixels for a typical image. Exact top-5 agreement with CIG is comparatively rare, occurring only for a minority of images. The same conclusion appears in the broader case-level metrics: the median cosine similarity is about 0.95 for EG versus CIG and about 0.94 for mean-image versus CIG, while the most important pixel agrees with CIG in only about 55% of images for EG

Figure 4: Digit Attribution Overlap



The figure shows the overlap between the top 5 features when comparing CIG to alternative baselines: EG, the mean image, and a black-image, respectively. The figure analyzes the attribution overlap for the 540 individual test images in the sample.

and 59% for the mean-image baseline. Thus, even though the aggregate feature rankings are highly correlated, the prediction-level explanations differ meaningfully across reasonable baselines.

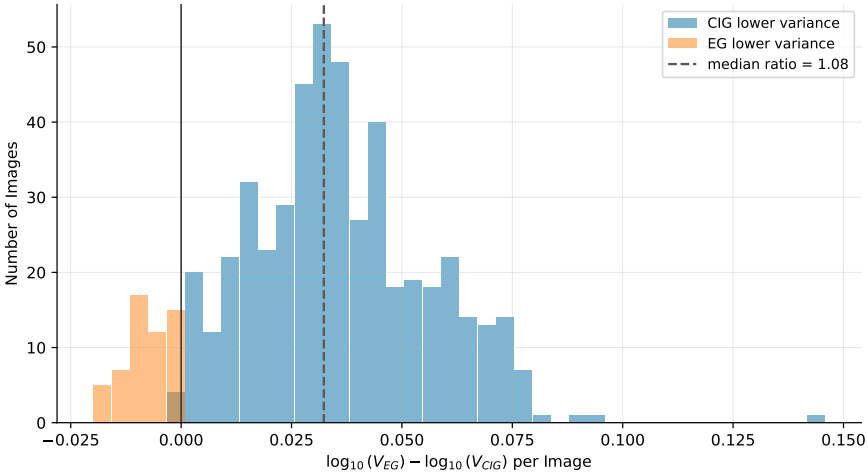
We can also compare EG and CIG from an efficiency perspective by taking A_{CIG} as the estimand and treating EG as a competing estimator. For each test image we compute the per-baseline variance

$$V_Q(x) = \sum_i w_i \, \|IG(x_i, x) - A_Q(x)\|^2 \tag{37}$$

under the full neutral weights ($Q = Q_N$) and under uniform weights ($Q = P_X$), reusing the same per-baseline attributions. A m-baseline Monte Carlo estimator then has variance $V_Q(x)/m$, so the ratio V_{EG}/V_{CIG} measures how many more baselines EG needs to match the precision of CIG.

Figure 5 shows that CIG is modestly more efficient than EG in this example. Across the 540 test images, the median variance ratio is approximately 1.08, with an interquartile range of roughly 1.04 to 1.11. Thus EG requires more baselines than CIG to achieve slightly lower precision, but CIG's gain is smaller compared to the synthetic example. This modest variance advantage is consistent with the digits problem being a relatively benign setting in which the neutral reference overlaps materially with the empirical image distribution, so EG and CIG average over fairly similar families of baseline images.

Figure 5: Attribution Efficiency



Taking the canonical attribution $A_{CIG}(x) = E_{Q_N}[IG(X_0, x)]$ as the estimand, Expected Gradients is a competing estimator whose bias is small here (Figure 4). So the variance comparison is an efficiency comparison. For a fixed target x, $V_Q(x) = \mathrm{Var}_{X_0 \sim Q}[IG(X_0, x)]$ is the per-baseline variance under reference distribution Q ($Q = P_X$ for EG, $Q = Q_N$ for CIG); a k-baseline estimate has variance $V_Q(x)/k$, so V_{EG}/V_{CIG} is the factor by which EG needs more baselines to match CIG's precision at equal cost. The figure shows the distribution of $\log_{10}(V_{EG}) - \log_{10}(V_{CIG})$ across 540 test images.

These findings help clarify the role of baseline selection in image attribution. The all-black image is not a typical observation from this data distribution and does not correspond to a prediction-neutral reference point. It answers a different question: how the prediction changes relative to the absence of any image signal. By contrast, the mean-image baseline, EG, and CIG all use reference images that remain close to the empirical data manifold, which explains why they agree much more closely in aggregate. At the same time, the prediction-level comparisons show that these methods are not interchangeable. Even when aggregate feature-importance rankings are very similar, the explanation of a particular image can depend materially on the choice of baseline distribution.

The digit-classification example occupies an intermediate regime. Canonical Integrated Gradients does not produce a qualitatively different attribution story from other realistic baselines, as it does relative to the all-black image. But neither does it collapse to EG or to the mean-image baseline. Instead, it delivers a prediction-neutral reference distribution that remains close to the data manifold, produces aggregate explanations similar to other realistic baselines, and still differs enough at the case level to matter for the interpretation of individual predictions.

4.3 Ames Housing Regression

We next study CIG applied to the Ames housing dataset (De Cock, 2011), a common benchmark for tabular regression. There are 2,930 total observations and 80 base features, including a mix of numerical variables and categorical attributes (e.g., neighborhood, quality ratings). We preprocess all variables using median imputation, standardization, and one-hot encoding. After one-hot encoding the categorical features, there are about 290 features.[5] When reporting feature importance, we group the contributions from all the dummies corresponding to a single categorical feature.

We train a nonlinear PyTorch regression model for log sale prices on a random half of the sample and then run attribution on the remaining 1,465 observations in the hold-out sample. The model fits the held-out sample reasonably well, with an out-of-sample R^2 of 0.785. The neutral prediction is the training-sample mean log price, $f_0 = 12.014$. The mean training prediction is slightly lower, 11.989, while the mean evaluation prediction is 12.006. Thus the conventional mean-feature baseline, Expected Gradients, and Canonical Integrated Gradients decompose slightly different prediction gaps.

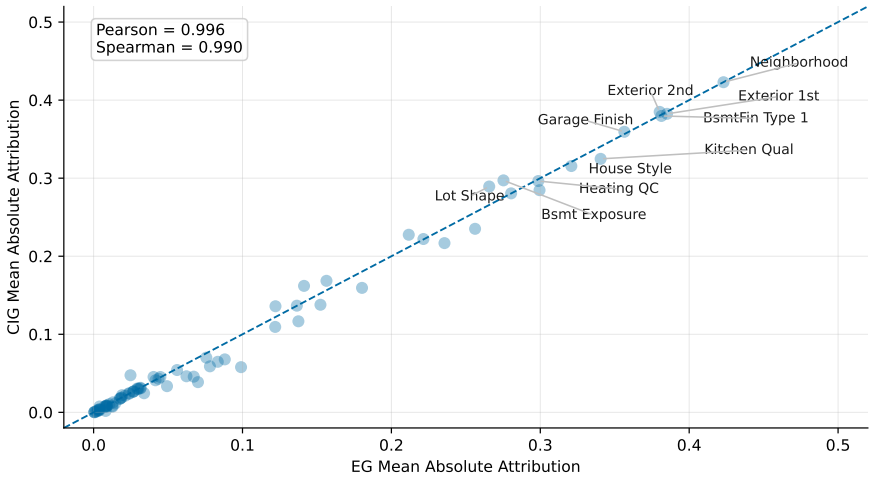
The neutrality diagnostics show that the CIG reference distribution satisfies exact neutrality in finite sample. The full neutral weights have effective sample size 122.5, and the sparse reference set retains 123 baselines with effective sample size 108.0. The resulting weighted baseline prediction equals f_0 to numerical precision. By contrast, Expected Gradients uses 1,000 empirical baselines whose mean prediction is 12.002, close to but not exactly equal to the neutral level. The mean-feature baseline is less neutral, with baseline prediction 11.948.

Aggregate feature-importance rankings are highly similar across methods. Canonical IG has Pearson correlation 0.996 with Expected Gradients, 0.995 with mean-baseline IG, and 0.995 with SHAP when feature importance is measured by mean absolute attribution. These results indicate that, for this fitted model and dataset, the principal feature-importance conclusions are largely insensitive to the choice among reasonable reference distributions. Figure 6 shows that CIG and EG produce very similar average absolute feature rankings.

High aggregate agreement does not imply identical explanations for individual predictions. The aggregate importance statistics average across features and observations, masking some case-level differences. Comparing explanations observation-by-observation, the median cosine similarity

[5] The number of expanded features varies slightly across samples because we apply one-hot encoding after the train-test split, and rare categorical levels may be absent in a given training sample.

Figure 6: Feature Importance Correlation



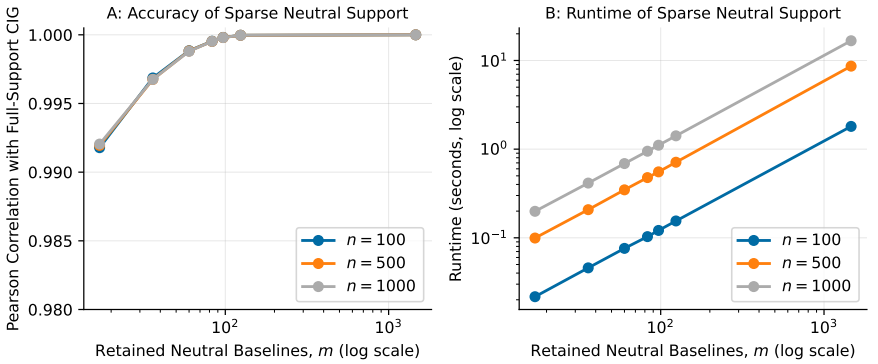
The figure compares individual feature importance for CIG and EG in the Ames housing regression example. The regression uses a PyTorch neural network and the analysis is based on 1,465 test samples.

between Canonical IG and Expected Gradients is 0.990, while the median similarity between Canonical IG and mean-baseline IG is 0.988. Relative attribution differences are approximately 15-16 percent of the average attribution norm, and the identity of the most important feature differs in roughly 17-20 percent of observations. Thus, even in a benign regression problem where aggregate feature rankings are nearly identical, baseline choice can affect the explanation of individual predictions.

The timing results show the computational advantage of concentrating on the neutral reference set. For 1,000 target observations, Expected Gradients requires 18.7 seconds using 1,000 baselines, while Canonical IG requires 1.5 seconds using the sparse neutral support. SHAP DeepExplainer requires 10.0 seconds, and mean-baseline IG is fastest because it uses only one baseline. Thus Canonical IG produces results close to Expected Gradients in this example, but at roughly one-twelfth of the runtime while preserving exact neutrality.

A support-size experiment confirms that the sparse approximation is not fragile. Full-support Canonical IG uses all 1,465 training observations and takes 17.6 seconds for 1,000 target observations. Retaining 99 percent of the squared-weight mass uses only 124 baselines, takes 1.5 seconds, and has Pearson correlation 0.99997 with the full-support neutral attribution. Even much smaller supports remain accurate. Retaining 50 percent of squared-weight mass uses 36 baselines, takes 0.45 seconds, and still has Pearson correlation 0.9968 with the full-support neutral result. Retaining 25

Figure 7: Effects of Support Size



Panel A shows the correlation between average CIG attributions as a function of the number of included baseline inputs. CIG retains the baseline inputs that produce predictions closest to the neutral prediction f_0. The three lines show correlations averaged over random samples with n sample points.

Panel B shows computational cost for CIG as a function of the number of included baseline inputs. The lines correspond to sample sizes of n points.

percent uses only 17 baselines, takes 0.21 seconds, and keeps the correlation above 0.992.

Figure 7 illustrates that CIG attributions are materially correlated for a broad range of support choices but that smaller supports produce faster runtimes.

Overall, the Ames example represents a stable, approximately linear setting in which all reasonable reference distributions produce similar explanations. Aggregate feature-importance rankings are nearly identical, and even individual explanations remain highly correlated. In this setting, the principal advantages of Canonical IG are not materially different attributions but rather a principled neutral reference distribution, exact neutrality by construction, diagnostic information about the effective baseline population, and substantial computational savings through concentration on a limited neutral support.

These results suggest that baseline sensitivity is often modest when reasonable baselines are used, while providing a framework for identifying situations in which baseline choice is consequential.

5 Summary

Integrated gradients explain predictions relative to a reference distribution. The central question is not how to choose a baseline input, but how to choose a reference population. We argue that the natural attribution question is: *why does the model predict* $f(x)$ *rather than a neutral prediction* f_0*?* For regression, the neutral prediction is the unconditional mean outcome; for classification,

it is the unconditional class-probability vector. For nonlinear models, neutral predictions generally correspond to a manifold of inputs rather than a unique reference point because $f(E[X]) \neq E[f(X)]$.

This perspective leads to Canonical Integrated Gradients (CIG), defined as the expected integrated gradient over the data distribution restricted to the neutral manifold,

$$A_{CIG}(x) = E\left[\, IG(X_0, x) \mid f(X_0) = f_0 \,\right]. \tag{38}$$

We estimate this quantity by weighting observed inputs according to the proximity of their predictions to f_0 and then enforcing exact neutrality through a minimal chi-square adjustment. The resulting estimator is consistent, requires only the fitted model and observed data, and avoids modeling the feature distribution.

Viewing attribution methods through the lens of reference distributions unifies fixed-baseline integrated gradients, mean-input baselines, Expected Gradients, and CIG. For linear calibrated models they decompose the same total prediction gap when their reference distributions are neutral, but component-level agreement requires stronger conditions on the reference means. More generally, the methods differ only through the reference distributions they induce. When those distributions are similar, as in the Ames housing and digit-classification examples, the resulting explanations are often similar as well. When they differ substantially, the differences take two forms: a wrong-gap error when the reference is not neutral, as for the mean-input and all-black baselines, and a reallocation of a correctly sized gap when the reference is neutral but is not the conditional Q_N, as for Expected Gradients in the synthetic example.

The empirical results suggest that baseline sensitivity is often less severe than implied by extreme examples, but they also illustrate the value of a principled reference distribution. CIG provides exact neutrality, diagnostic information about the effective baseline population, substantial computational savings relative to large-scale averaging schemes such as Expected Gradients, and a clear interpretation tied directly to neutral predictions. In this sense, CIG is best viewed not as a replacement for all existing baselines, but as a canonical neutral reference against which other baseline choices can be understood and evaluated.

6 References

Aas, Kjersti, Martin Jullum, and Anders Løland, 2021, Explaining individual predictions when the features are dependent: More accurate approximations to Shapley values, *Artificial Intelligence* 298.

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, Joseph D. Janizek, Scott Lundberg, and Su-In Lee, 2020, True to the model or true to the data?, Working paper, University of Washington, Seattle, WA.

De Cock, Dean, 2011, Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project, *Journal of Statistics Education* 19 (3), 1–15.

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

Erion, Gabriel, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee, 2021, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, *Nature Machine Intelligence* 3 (7), 620–631.

Frye, Christopher, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige, 2021, Shapley explainability on the data manifold, in *International Conference on Learning Representations (ICLR)*.

Hentschel, Ludger, 2026, TreeIG: Exact integrated gradients for tree-based models, Working paper, Versor Investments, New York, NY.

Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum, 2020, Feature relevance quantification in explainable AI: A causal problem, in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, 2907–2916 (PMLR).

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).

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

Sturmfels, Pascal, Scott Lundberg, and Su-In Lee, 2020, Visualizing the impact of feature attribution baselines, *Distill* 5 (1), e22.

Sundararajan, Mukund, and Amir Najmi, 2020, The many Shapley values for model explanation, in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 9269–9278 (PMLR).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan, 2017, Axiomatic attribution for deep networks, in Doina Precup, and Yee Whye Teh, eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328 (PMLR).

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

A Efficient Computation

The neutral-manifold attribution $\widehat{A}_N(x) = \sum_i w_i^* IG(x_i, x)$ of Section 3 costs one integrated-gradient evaluation per observation that carries positive weight. With the full support this is order n evaluations—the same cost as Expected Gradients—even though the kernel weights concentrate and most observations carry negligible mass. This appendix gives an estimator whose cost is set by the required precision of the attribution rather than by the sample size, and which never evaluates integrated gradients at all n observations. Throughout, $r_i = f(x_i) - f_0$ is the prediction gap and w^* the neutral-manifold weights from (20).

A.1 Screening

The weights w_i^* form a probability distribution with known scale: each is comparable to the uniform share $1/n$. The closed-form projection (21) gives $w_i^* = q_i\big(1 - \mu_q(r_i - \mu_q)/\sigma_q^2\big)$, which already absorbs the neutrality tilt, so ordering by w_i^* ranks importance better than ordering by the bare gap $|r_i|$. Screening on the weights is the natural first step; the question is where to stop.

A mass-coverage rule—retain weights above c/n, or retain the largest until they cover a fixed fraction of the total—ties the retained count to the effective number of baselines $n^* = 1/\sum_i (w_i^*)^2$. When the weights are diffuse, spread over a band of n_{band} near-neutral observations of comparable size, $n^* \approx n_{band}$, which grows linearly with n because the density of predictions near f_0 is fixed. A mass-coverage cut then retains order n observations and the speed advantage is lost in the diffuse case. The right target is the precision of the attribution, which is n-independent, not the weight mass, which is not: $\widehat{A}_N(x)$ is an average over near-neutral references whose individual attributions vary smoothly, so a few references estimate that average well.

A.2 Sizing by Precision

Sample the support rather than thresholding it, so that the precision statement is exact. Draw baselines $i_1, \ldots, i_k$ independently with probability w_i^* and form

$$\widehat{A}_{N,k(x)} = \frac{1}{k} \sum_{t=1}^{k} IG(x_{i_t}, x), \tag{39}$$

an unbiased estimate of $\widehat{A}_N(x)$ with standard error $\sqrt{V/k}$, where $V = \mathrm{Var}_{w^*}\big(IG(x_i, x)\big)$ is the variance of the attribution across references—the

reference-sensitivity discussed in Section 2.4—not the numerical error within a single path integral. Since V is unknown, size k in two stages: compute integrated gradients for a pilot batch of k_0 sampled baselines, form the componentwise sample variances s_j^2, and set

$$k^* = \max_j \left\lceil \frac{t_{k_0-1}^2 \, s_j^2}{tol_j^2} \right\rceil, \tag{40}$$

where the tolerances tol_j are set relative to $f(x) - f_0$ so that adequate precision means precision relative to the quantity being decomposed. If $k^* > k_0$, draw the additional baselines and verify; one top-up is almost always sufficient. This is Stein's two-stage procedure: the pilot variance yields the required count in one step. Because V is small when the near-neutral references are homogeneous—as they typically are—k^* is small and independent of n; when references genuinely disagree, k^* rises accordingly.

A.3 A Control Variate from the Linear Attribution

The count k^* can be reduced further by extracting the part of the attribution that is cheap to compute exactly. Split each integrated gradient into a linear proxy and a nonlinear residual,

$$IG(x_i, x) = L_i + R_i, \qquad L_i = \nabla f(x) \odot (x - x_i), \tag{41}$$

where $\odot$ denotes the elementwise product and L_i requires only the gradient at x, shared across all baselines. The weighted linear part collapses to a single evaluation,

$$\sum_i w_i^* L_i = \nabla f(x) \odot (x - \overline{x}_w), \qquad \overline{x}_w = \sum_i w_i^* x_i, \tag{42}$$

so the neutral-manifold attribution becomes

$$\widehat{A}_N(x) = \nabla f(x) \odot (x - \overline{x}_w) + \sum_i w_i^* R_i, \tag{43}$$

with only the residual term requiring path integrals. Sizing the sample by (40) on the residual variance $\mathrm{Var}_{w^*}(R)$ rather than V requires far fewer evaluations: the residual vanishes identically when f is linear and is a small, low-variance correction when f is mildly nonlinear. The linear backbone costs one gradient evaluation for the entire reference set; path integrals are spent only on the curvature the model actually exhibits.

A.4 Procedure

Writing S for the sampled support of size k^*:

1. Compute the kernel weights q from (**??**) and the neutral projection w^* from (21) using predictions alone; no gradient evaluations are required.
2. Compute the linear backbone $\nabla f(x) \odot (x - \overline{x}_w)$ with a single gradient at x.
3. Draw a pilot of k_0 baselines proportional to w^*, evaluate their integrated gradients and residuals, and set k^* from (40) using the residual variance; top up to k^* if needed.
4. Refit the weights on S to restore exact neutrality, $\sum_{i \in S} \widetilde{w}_i f(x_i) = f_0$, so that completeness $\sum_k [\widehat{A}_N(x)]_k = f(x) - f_0$ holds exactly; sampling proportional to w^* concentrates on both sides of f_0, so S generically brackets f_0 and the refit is feasible.
5. Report $\widehat{A}_N(x) = \nabla f(x) \odot (x - \overline{x}_{\widetilde{w}}) + \sum_{i \in S} \widetilde{w}_i R_i$.

The total number of integrated-gradient evaluations is k^*, governed by the residual variance and the target precision, independent of the sample size, in place of the order-n cost of the full neutral-manifold estimator or of Expected Gradients.