Common Functional Decompositions Can Mis-attribute Differences in Outcomes Between Populations

Manuel Quintero
MIT IDSS
mquint@mit.edu

William T. Stephenson
MIT Lincoln Laboratory[*]
william.stephenson@ll.mit.edu

Advik Shreekumar
MIT Economics
adviks@mit.edu

Tamara Broderick
MIT EECS
tbroderick@mit.edu

Abstract

In science and social science, we often wish to explain why an outcome is different in two populations. For instance, if a jobs program benefits members of one city more than another, is that due to differences in program participants (particular covariates) or the local labor markets (outcomes given covariates)? The Kitagawa-Oaxaca-Blinder (KOB) decomposition is a standard tool in econometrics that explains the difference in the mean outcome across two populations. However, the KOB decomposition assumes a linear relationship between covariates and outcomes, while the true relationship may be meaningfully nonlinear. Modern machine learning boasts a variety of nonlinear functional decompositions for the relationship between outcomes and covariates in one population. It seems natural to extend the KOB decomposition using these functional decompositions. We observe that a successful extension should not attribute the differences to covariates — or, respectively, outcomes given covariates — if those are the same in the two populations. Unfortunately, we demonstrate that, even in simple examples, two common decompositions — the functional ANOVA and Accumulated Local Effects — can attribute differences to outcomes given covariates, even when they are identical in two populations. We provide and partially prove a conjecture that this misattribution arises in any additive decomposition that depends on the distribution of covariates.

1 Introduction

Motivating Example. The mayor of City K compares the results of a job training program to a similar one in City H. She collects data about participants' income after the program (Y), and other covariates like age and employment before the program (X). She notices that program graduates in her city have lower post-program income than those in City H ($E_K[Y] \neq E_H[Y]$). If she can figure out why this difference occurs, perhaps she can modify the job training program or its recruitment strategy to make it more effective. Before committing to a potentially costly causal analysis, can she use the available covariates to develop hypotheses for why program outcomes differ in her city?

Many scientific questions reduce to comparisons between two populations. A common follow-up question to observing differences is why they occur. One reason might be that the populations differ

[*] DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

on meaningful traits. In our example, perhaps the distribution of X is unequal: say, both training programs are best for people who have never been employed, but city K's program enrolled more trainees with past jobs than city H's program. In this case, covariates X drive the difference. Or, perhaps jobs in city K may pay less than those in city H's for trainees without a high school diploma. In this case, outcomes given covariates $Y \mid X$ drive the difference. A useful explanation for mean differences between populations would distinguish between these possibilities, as well as describe which aspects of the covariates or outcomes given covariates explain the difference.

The Kitagawa-Oaxaca-Blinder (KOB) decomposition (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) is widely used in the econometrics literature to solve exactly this problem. The KOB decomposition separates difference of means into components that depend on the distribution of covariates, X, and those that depend the conditional expectation of outcomes given features, $\mathbb{E}[Y|X]$. However, it relies on parametric linear models of the conditional expectation. A natural extension would allow for non-linear or nonparametric models for $\mathbb{E}[Y|X]$. Such an approach could account for shifts in the distribution of X through a generic step-wise transformation that moves the distribution of X from population H to population K. Importance can be assigned to individual features in the change in conditional expectation $\mathbb{E}[Y|X]$ through the use of additive functional decomposition methods.

Such a generalization requires a choice of the functional decomposition. Fortunately, modern machine learning offers multiple options. For example, the functional ANOVA (FANOVA) (Stone, 1994; Huang, 1998; Hooker, 2007) and Accumulated Local Effects (ALE) (Apley and Zhu, 2020) decompositions have been widely used in sensitivity analysis (Chastaing et al., 2012; Antoniadis et al., 2021), machine learning interpretability (Lengerich et al.; Limmer et al., 2024), finance (Liang and Cai, 2022; Belhadi et al., 2021), and environmental and climate sciences (Huang et al., 2023; Peichl et al., 2021; Hill et al., 2023).

The success of such decompositions makes them seem like natural choices for use in explaining the difference in means. However, we demonstrate that common functional decomposition – the FANOVA and ALE – are ill-suited for this task. We further conjecture and partially prove that any functional decomposition method that depends on the covariate distribution in H or K (which includes the FANOVA and ALE) must misattribute differences stemming from a changing distribution of X to differences stemming from changing $Y \mid X$. We thus argue that common functional decomposition methods are, despite their broad success in many areas, inappropriate for use for explaining differences between two populations.

2 Related work and notation

2.1 Kitagawa-Oaxaca-Blinder decomposition

The Kitagawa-Oaxaca-Blinder (KOB) decomposition provides a framework for explaining differences in means between two populations by decomposing them into components attributable to differences in covariates and conditional expectations. In its original form, KOB assumes a linear relationship between the covariates $X \in \mathbb{R}^d$ and the outcome $Y \in \mathbb{R}$:

$$\mathbb{E}_{H_{Y|X}}[Y|X] = X\beta_H \quad \text{and} \quad \mathbb{E}_{K_{Y|X}}[Y|X] = X\beta_K,$$

where $H_{Y|X}$ is the conditional distribution of $Y \mid X$ in population H; similarly for population K. X are finite-dimensional covariates, and β^H and β^K represent the respective coefficients for the linear relationship in each population. KOB decomposes the difference $\mathbb{E}_K[Y] - \mathbb{E}_H[Y]$ as

$$\underbrace{\mathbb{E}_{K_X}[\mathbb{E}_{K_{Y|X}}[Y|X] - \mathbb{E}_{H_{Y|X}}[Y|X]]}_{Y \mid X \text{ effect}} + \underbrace{\mathbb{E}_{K_X}[\mathbb{E}_{H_{Y|X}}[Y|X]] - \mathbb{E}_{H_X}[\mathbb{E}_{H_{Y|X}}[Y|X]]}_{\text{Covariate effect}} \tag{1}$$

$$= \sum_{j=1}^{d} \underbrace{\mathbb{E}_{K_X}[X_j](\beta_j^K - \beta_j^H)}_{Y \mid X \text{ effect for } j\text{th covariate}} + \sum_{j=1}^{d} \underbrace{(\mathbb{E}_{K_X}[X_j] - \mathbb{E}_{H_X}[X_j])\beta_j^H}_{\text{Covariate effect for } j\text{th covariate}}, \tag{2}$$

where H_X and K_X represent the distribution of the covariates for the respective populations. Note that throughout, when referring to a marginal distribution or a conditional distribution, we will specify it as a subscript of the joint distribution of X, Y, which will be denoted as H or K. For instance, H_i represents the marginal distribution of feature x_i, and $H_{1:i}$ refers to the joint distribution of the first i covariates in X for population H.

As we will see in the next section, there is a natural extension of the KOB decomposition that offers a similar interpretation for more general functional decomposition methods. The existing literature provides several such methods; however, we focus on two—FANOVA and ALE—that decompose functions into additive components. Other functional decomposition methods, such as Partial Dependence Plots (Friedman, 2001), do not offer additive decompositions, and thus cannot be immediately used as a part of KOB-like decompositions.

2.2 FANOVA

FANOVA measures the importance of features in determining the output of the function and in identifying underlying additive interactions between subsets of variables (Hooker, 2004). It presents a natural representation of the functional in terms of low-order components (Hooker, 2007) by stating that a square integrable function $f(x)$ with $x \in \mathbb{R}^d$ can be written uniquely as $f(x) = \sum_{S \in 2^{[d]}} \mathcal{L}(f, W, S)(x)$, where $2^{[d]}$ denotes the power set of $[d] = \{1, 2, \ldots, d\}$, W is a general measure of the covariates, and the components are jointly defined by satisfying

$$\{\mathcal{L}(f, W, S) \mid S \in 2^{[d]}\} = \underset{\{\mathcal{L}(f,W,S) \in L^2(\mathbb{R}^d)\}_{S \in 2^{[d]}}}{\arg\min} \int \left(\sum_{s \in 2^{[d]}} \mathcal{L}(f, W, s)(x_s) - f(x) \right)^2 W(x) dx, \tag{3}$$

subject to *hierarchical orthogonality conditions* among them.

2.3 Accumulated Local Effects (ALE)

ALE is another additive functional decomposition method that is particularly suitable for visualizing the effects of predictors (Apley and Zhu, 2020). Although ALE is more generally defined, the case for $d = 2$ with a differentiable $f(x_1, x_2)$ suffices for our illustrative purposes. The ALE component for x_1 is then defined as:

$$\mathcal{L}(f, W, \{1\}) = \int_{x_{\min,1}}^{x_1} \mathbb{E}_{X_2 \sim W_2} \left[\frac{\partial f(X_1, X_2)}{\partial X_1} \middle| X_1 = z_1 \right] dz_1 - \text{constant}, \tag{4}$$

where $\frac{\partial f(x_1, x_2)}{\partial x_1}$ is the partial derivative of the function f with respect to x_1, and $x_{\min,1}$ is a lower bound of the support of W_1. The term $\mathcal{L}(f, W, \{2\})$ is defined similarly; for the definition of $\mathcal{L}(f, W, \{1, 2\})$ and for the $d > 2$ case, see (Apley and Zhu, 2020).

3 Additive decompositions of population differences

As discussed in Section 1, a natural and desirable extension of KOB, Section 2.1, would allow for arbitrary flexible regression models by extending it to non-linear functional forms. Recall the KOB decomposition in Equation 2 separates a difference in means into a $Y \mid X$ effect and a covariate effect. To extend the KOB decomposition to more flexible models, we assume a general functional form for the conditional expectation. Specifically, we assume that a flexible regression model is fitted such that $f^K(X) \approx \mathbb{E}_{K_{Y|X}}[Y|X]$, and similarly for population H. Our goal is to decompose Equation 1 into smaller, interpretable components just as in the KOB decomposition. To achieve this goal, and in the spirit of FANOVA and ALE discussed in Section 2, we assume a generic additive functional decomposition, denoted by $\mathcal{L}$, which operates on arbitrary functions f of the covariates, distributions over the covariates H_X, and subsets of features S. This decomposition yields an additive representation that holds for all $x \in \mathbb{R}^d$.

$$f(x) = \sum_{S \in 2^{[d]}} \mathcal{L}(f, H_X, S)(x), \tag{5}$$

Given such an additive functional decomposition, it is straightforward to extend the KOB decomposition. We define two types of swaps, analagous to the terms in the KOB decomposition. First, we can swap out distributions over covariates one dimension of X at a time; we call such terms *the difference* *due to the changing distribution of X_i*. Second, we can swap out the functional decomposition terms of f^H for those of f^K; we call such terms *the difference due to differences in $Y \mid X$*. We define the KOB extension decomposition for the general case below:

Definition 1. *Let $\mathcal{S}$ define an ordering of all subsets $S \subset 2^{[d]}$; we refer to the ith subset in this ordering as $\mathcal{S}_i$. We define* the importance decomposition *to be:*

$$\mathbb{E}_K[Y]{-}\mathbb{E}_H[Y] = \sum_{i=1}^{|\mathcal{S}|} \delta_{S_i}^{Y|X} + \sum_{j=1}^{d} \delta_j^X, \tag{6}$$

$$\begin{aligned} \textit{where: } \delta_{S_i}^{Y|X} &:= \mathbb{E}_{H_X}\left[\sum_{j=1}^{i} \mathcal{L}(f^K, K_X, S_j) + \sum_{j=i+1}^{|\mathcal{S}|} \mathcal{L}(f^H, H_X, S_j)\right] \\ &\quad - \mathbb{E}_{H_X}\left[\sum_{j=1}^{i-1} \mathcal{L}(f^K, K_X, S_j) + \sum_{j=i}^{|\mathcal{S}|} \mathcal{L}(f^H, H_X, S_j)\right] \\ \delta_j^X &:= \mathbb{E}_{K_{1:j|j+1:d}H_{j+1:d}}\left[f^K(X)\right] - \mathbb{E}_{K_{1:j-1|j:d}H_{j:d}}\left[f^K(X)\right] \end{aligned}$$

Note that the sums in $\delta_{S_i}^{Y|X}$ differ at index i, so that $\delta_{S_i}^{Y|X} = \mathbb{E}_{H_X}[\mathcal{L}(f^K, K_X, S_i) - \mathcal{L}(f^H, H_X, S_i)]$. We therefore call $\delta_{S_i}^{Y|X}$ the difference due to the dependence of $Y \mid X$ on feature subset S_i*. Likewise, the distributions over covariates in δ_j^X differ in whether X_j follows a distribution determined by H or K. We therefore call δ_j^X* the difference due to the change in distribution of covariate j*.*

Definition 1 is an extension of the KOB decomposition from Section 2, which also defines differences from swapping out distributions of covariates, as well as differences in swapping out (a model for) $Y \mid X$. The main difference is that Definition 1 uses a generic additive decomposition of $Y \mid X$, whereas the KOB decomposition assumes a linear model.

This decomposition – like the KOB decomposition – makes a series of specific choices: first swapping S_1, then S_2, ... then finally swapping $S_{|S|}$, and then swapping covariate one, then covariate two, etc. Why not swap S_2 first? Why not swap covariate three immediately after S_1? In general, there is no reason to prefer any one ordering, and different orderings will produce different results. With no preferred ordering of swaps, one may prefer to average over all possible orderings and report the resulting averages as the definitions of $\delta_{S_i}^{Y|X}$ and δ_j^X.[2] Our results here apply to any fixed order; we leave the extension to averaging over all orderings as future work.

4 Failure of existing decompositions

Once a user has specified the functional forms of $f^H(X)$ and $f^K(X)$, the only decision to be made before using Definition 1 is the choice of functional decomposition $\mathcal{L}$. At first glance, options such as ALE or FANOVA from Section 2.2 and 2.3 seem like excellent choices: they provide additive decompositions of generic functions with properties that make them well-suited for understanding functions in other applications. However, we conjecture that a broad class of functional decompositions, including FANOVA and ALE, are inappropriate for explaining population differences in the sense of Definition 1, despite their great success in other applications. In particular, we conjecture that such decompositions incorrectly state that differences stem from changes in $Y \mid X$.

Recall that Definition 1 defines $\delta_{S_i}^{Y|X}$ to be the *difference due to the dependence of $Y \mid X$ on feature subset S_i*. Suppose that the distributions $Y \mid X$ are in fact identical across the two populations H and K, and thus $f^K = f^H = f$. In such situations, any reasonable decomposition should lead us to believe there is no difference due to differences in $Y \mid X$; that is, $\delta_{S_i}^{Y|X} = 0$. Unfortunately, the next example shows that FANOVA can misattribute differences to differences in $Y \mid X$.

Example 1. *To begin with, note that when $f^K = f^H = f$, $\delta_S^{Y|X}$ reduces to*

$$\Delta(f, H_X, K_X, S) := \delta_S^{Y|X} = \mathbb{E}_{H_X}\left[\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S)\right]. \tag{7}$$

[2] Shorrocks (2013) describes such averages as applying logic of Shapley values to functional decompositions.

Consider the case where the fitted model is additive and has two covariates: $f^K = f^H = f(x) = x_1 + x_2$. Suppose that population H has covariates X_1, X_2, with $\mathbb{E}_{H_X}[X_1] = \mathbb{E}_{H_X}[X_2] = 0$, while in population K, $\mathbb{E}_{K_X}[X_1] = \mu$ and $\mathbb{E}_{K_X}[X_2] = 0$ for $\mu \neq 0$. In both populations, X_1 and X_2 are independent and have finite variance.

Then, following the FANOVA decomposition in Equation 3, the components for each subset satisfy the following for each population:

$$\hat{f}^H_{\emptyset}(x) = 0, \quad \hat{f}^H_{\{1\}}(x) = x_1, \quad \hat{f}^H_{\{2\}}(x) = x_2, \quad \hat{f}^H_{\{1,2\}}(x) = 0,$$

$$\hat{f}^K_{\emptyset}(x) = \mu, \quad \hat{f}^K_{\{1\}}(x) = x_1 - \mu, \quad \hat{f}^K_{\{2\}}(x) = x_2, \quad \hat{f}^K_{\{1,2\}}(x) = 0.$$

Hence, the difference in means due to differences in $Y \mid X$ for the component of $S = \{1\}$, is given by,

$$\Delta(f, H_X, K_X, \{1\}) = \mathbb{E}_{H_X}\left[\hat{f}^K_{\{1\}}(x) - \hat{f}^H_{\{1\}}(x)\right] = \mathbb{E}_{H_X}[x_1 - \mu - x_1] = -\mu \neq 0,$$

which is not equal to zero, so FANOVA misattributes effects to $Y \mid X$ in this example.

In Appendix 2, we provide an example for ALE where $f^K = f^H$ but $\delta^{Y|X}_{S_i} \neq 0$. Thus either ALE or FANOVA can misattribute differences to differences in $Y \mid X$.

Without a clear understanding of when such misassignments will occur, the decomposition of Definition 1 is of little value, as practitioners would never know when to trust its outputs. To resolve this problem, we attempt to characterize how properties of the functional decomposition $\mathcal{L}$ cause this misassignment. In particular, we conjecture that if $\mathcal{L}$ depends on its input probability distribution, then it can misassign the effects, and the output of Definition 1 is not to be trusted. This holds true for both ALE and FANOVA; given that both see broad use in the machine learning and statistics literature, we conclude that requirements for understanding population differences are different than requirements in other applications. Further, we conjecture the converse is true: Definition 1 does not have this misassignment only if $\mathcal{L}$ does not depend on its input distribution. To prove this, we need to define what we mean by dependence of $\mathcal{L}$ on the input distribution and misassignment of effects.

Definition 2. *We say that a functional decomposition $\mathcal{L}$ does not depend on its input distribution if for all f, H_X, K_X and S, $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) \equiv 0$.*

Definition 3. *We say that a functional decomposition $\mathcal{L}$ misattributes effects of $Y \mid X$ if $\Delta(f, H_X, K_X, S) \neq 0$ for any f, H_X, K_X, S.*

Therefore, we aim to characterize conditions on the functional $\mathcal{L}$ under which $\Delta(f, H_X, K_X, S)$ does or does not equal zero for all f_X, H_X, K.

5 When do functional decompositions misattribute effects?

We conjecture that, under regularity assumptions on $\mathcal{L}(f, K_X, S)$, the function f, and the densities H_X and K_Y, a functional decomposition $\mathcal{L}$ does not misattribute the effects of $Y \mid X$ if and only if it does not depend on its input distribution.

Suppose $\mathcal{L}(f, K_X, S)$ is a continuous functional in its first argument f, Lebesgue measurable in its second argument, K_X, and square integrable, in the L^2 sense, for all triplets (f, K_X, S). For example, our first condition is satisfied in cases, such as in FANOVA, when $\mathcal{L}$ is the integral operator with respect to any probability measure or the Lebesgue measure. The third assumption is identical to those in FANOVA and ALE, which both require $\mathcal{L}$ to belong to the space of square integrable functions, L^2. Lastly, we assume that the densities K_X belong to the space of compactly supported functions, denoted by $\mathcal{P}(X)$. Note that not only the definition of ALE decomposition assumes compactly support densities, but also this assumption is fairly mild. In practice, most distributions can be restricted to a compact region (e.g., age, income, and years of education are all bounded).

We parameterize perturbations around a density K_X as $K_X + \phi$, for admissible[3] functions ϕ. We denote by $\mathcal{D}_K$, the set of admissible perturbation functions of K_X. Under an additional condition,

[3] We require that ϕ be square integrable and that $K_X + \phi$ be a valid probability density; see Appendix B.2 for details.

assuming that $\mathcal{L}$ is continuously differentiable as a function of ϕ, we ensure that we can approximate $\mathcal{L}(\cdot, \phi)$ with the linear approximation around zero:

$$\mathcal{L}(\cdot, \phi) = \mathcal{L}(\cdot, 0) + D_{\phi}\mathcal{L}(\cdot, 0)[\phi].$$

Where $D_{\phi}\mathcal{L}(\cdot, 0)[\phi]$ is known as the (Fréchet) derivative of $\mathcal{L}$ with respect to the function ϕ evaluated at the zero function and it operates on ϕ. See Definition 4 and Appendix B.2 for a more rigorous discussion of the perturbation functions. Although we have not yet verified that FANOVA and ALE satisfy continuous differentiability with respect to perturbations, our conditions are mild, and so we conjecture that this is the case.

We now attempt to characterize the behavior of the functional $\mathcal{L}$ under small perturbations of the density K_X. Our main theoretical result depends on the following conjecture:

Conjecture 1. *Assume the above regularity conditions on $\mathcal{L}$ (Assumptions 1 and 2 in the Appendix). Let $K_X \in \mathcal{P}(X)$ and $\mathcal{D}_K$ denote the set of admissible perturbation function of K_X.*

If,

$$\mathbb{E}_{x\sim K_X+\phi}\left[\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x)\right] = 0, \textit{ for all } \phi \in \mathcal{D}_K.$$

Then,

$$D_{\phi}\mathcal{L}(\cdot, 0)[\phi] = 0,$$

and therefore $\mathcal{L}(f, K_X, S)$does not depend on its input distribution. That is,

$$\mathcal{L}(f, K_X, S) = \mathcal{L}(f, K_X + \phi, S), \textit{ for all } \phi \in \mathcal{D}_K.$$

While we have not yet fully proved Conjecture 1, we feel it is intuitively sensible: if a decomposition $\mathcal{L}$ does not misassign effects of transport for *any* distribution, then it must be constant with respect to its input distribution. See Appendix B.3 for a partial proof. Regardless, if Conjecture 1 is true, it leads to our main result:

Theorem 1. *Under the assumptions of Theorem 1, a functional decomposition $\mathcal{L}(f, K_X, S)$ does not misattribute effects of $Y \mid X$ if and only if it does not depend on its input distribution.*

Proof. The "if" part is straightforward: by definition, if $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = 0$ for all f, K_X, H_X, S, then $\Delta(f, K_X, H_X, S) = 0$. The proof of the "only if" relies on Theorem 1. Since K_X was chosen arbitrarily in Definition 3, this result must hold over the entire defined probability space, which implies that $\mathcal{L}(f, K_X, S) - \mathcal{L}(f, H_X, S) = 0$, for all f, K_X, H_X. □

Conditional on Conjecture 1, Theorem 1 demonstrates that any decomposition method that depends on its input distribution may misattribute effects. Our Examples 1 and 2, together with Theorem 1, underscore that popular decomposition methods, such as FANOVA and ALE, are not suitable for explaining differences between two populations, highlighting the need to develop novel decomposition techniques to tackle this problem.

6 Conclusion

In this work, we note that functional decompositions like FANOVA and ALE seem at first glance like excellent candidates for decomposing differences in two populations. However, we here provide simple counterexamples showing that both FANOVA and ALE can incorrectly assign differences in the distribution of covariates X to differences in the outcome-given-covariates, $Y \mid X$. We further conjecture that this phenomenon is more general: any functional decomposition method that depends on its input distribution will have this problem. Our result calls for the use of decompositions that do not depend on their input distribution for use in explaining population differences. Fortunately, such decompositions exist. In fact, an early version of the FANOVA did not depend on the input distribution (Hooker, 2004). While this non-dependence is often cited as a detriment to the interpretability of a functional decomposition in one population (Hooker, 2007; Apley and Zhu, 2020), our results give evidence that this non-dependence is necessary when comparing multiple populations. We leave application of such decompositions to explaining population differences as future work.

Acknowledgments

The authors are grateful to Raj Agrawal for his contributions and insights during the initial stages of this project. This research was supported in part by an ONR Early Career Grant. Advik Shreekumar was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

This material is based upon work supported by the Combatant Commands under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Combatant Commands.

References

Anestis Antoniadis, Sophie Lambert-Lacroix, and Jean-Michel Poggi. Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 206:107312, 2021. doi: 10.1016/j.ress.2020.107312. URL https://www.sciencedirect.com/science/article/pii/S0951832020308073.

Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, September 2020. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12377.

V. I. Averbukh and O. G. Smolyanov. The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, 22:201–258, 1967. doi: 10.1070/RM1967v022n02ABEH001223.

Amine Belhadi, Swapnil S. Kamble, V. Mani, et al. An ensemble machine learning approach for forecasting credit risk of agricultural smes' investments in agriculture 4.0 through supply chain finance. *Annals of Operations Research*, 2021. doi: 10.1007/s10479-021-04366-9.

Alan S. Blinder. Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, 8(4):436, 1973. ISSN 0022166X. doi: 10.2307/144855.

Gaelle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis. *Electronic Journal of Statistics*, 6(none), January 2012. ISSN 1935-7524. doi: 10.1214/12-EJS749.

Ward Cheney. *Analysis for Applied Mathematics*, volume 208 of *Graduate Texts in Mathematics*. Springer, 2001. doi: 10.1007/978-1-4613-0101-2. Chapter 3: Calculus in Banach Spaces.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, New York, 2nd edition, 1999. Theorem 3.8, "The Lebesgue-Radon-Nikodym Theorem".

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232, 2001.

Kathryn E. Hill, Stuart C. Brown, Alice Jones, Damien Fordham, and Robert S. Hill. Modelling climate using leaves of Nothofagus cunninghamii—overcoming confounding factors. *Sustainability*, 15(9):7603, 2023. doi: 10.3390/su15097603. URL https://doi.org/10.3390/su15097603.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580. ACM, 2004.

Giles Hooker. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, September 2007. ISSN 1061-8600, 1537-2715. doi: 10.1198/106186007X237892.

Feini Huang, Wei Shangguan, Qingliang Li, Lu Li, and Ye Zhang. Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. *Environmental Modelling & Software*, 167: 105762, 2023. ISSN 1364-8152. doi: 10.1016/j.envsoft.2023.105762. URL https://www.sciencedirect.com/science/article/pii/S1364815223001482.

Jianhua Z. Huang. Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, 26(1), February 1998. ISSN 0090-5364. doi: 10.1214/aos/1030563984.

Evelyn M Kitagawa. Components of a Difference Between Two Rates. *Journal of the American Statistical Association*, pages 1168–1194, 1955.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying Interaction Effects with the Functional ANOVA: An Efficient Algorithm for Recovering Identifiable Additive Models.

Longyue Liang and Xuanye Cai. Time-sequencing european options and pricing with deep learning – analyzing based on interpretable ale method. *Expert Systems with Applications*, 187:115951, 2022. doi: 10.1016/j.eswa.2021.115951.

Steffen Limmer, Steffen Udluft, and Clemens Otte. Neural-anova: Model decomposition for interpretable machine learning, Aug 2024. URL https://www.researchgate.net/publication/383308412_Neural-ANOVA_Model_Decomposition_for_Interpretable_Machine_Learning.

Ronald Oaxaca. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, 14 (3):693, October 1973. ISSN 00206598. doi: 10.2307/2525981.

M. Peichl, S. Thober, L. Samaniego, B. Hansjürgens, and A. Marx. Machine-learning methods to assess the effects of a non-linear damage spectrum taking into account soil moisture on winter wheat yields in germany. *Hydrology and Earth System Sciences*, 25:6523–6545, 2021. doi: 10.5194/hess-25-6523-2021. URL `https://doi.org/10.5194/hess-25-6523-2021`.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality*, 11(1):99–126, March 2013. ISSN 1569-1721, 1573-8701. doi: 10.1007/s10888-011-9214-z.

Charles J. Stone. The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. *The Annals of Statistics*, 22(1), March 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325361.

Eberhard Zeidler. *Nonlinear Functional Analysis and Its Applications: I: Fixed-Point Theorems*. Springer-Verlag, New York, 1986. ISBN 978-0387964992.

A Appendix / supplemental material

A.1 ALE misattributes the effect of transport

We present an example where $f^K = f^H$ but $\delta_S^{Y|X} \neq 0$ for the ALE decomposition, meaning it misattributes the effects of transport of $Y|X$.

Example 2. *Let $f^K = f^H = f(x) = x_1 x_2$, and for population H, $x_1 \sim N(1,1)$, $x_2 \sim N(0,1)$, and for population K, $x_1 \sim N(0,1)$, $x_2 \sim N(\mu, 1)$, where $\mu \neq 0$. Assume we observed data $\{(x^j_{i,1}, x^j_{i,2})\}_{i=1}^n$, with n sufficiently large, for $j = H, K$. The ALE (Apley and Zhu, 2020) main effect for X_1 is given by:*

$$\begin{aligned}\hat{f}_{1,ALE}(x_1) &\equiv \int_{x_{\min,1}}^{x_1} \mathbb{E}[f^1(X_1, X_2)|X_1 = z_1] dz_1 - c_1 \\ &= \int_{x_{\min,1}}^{x_1} \int p_{2|1}(x_2|z_1) f^1(z_1, x_2) dx_2 dz_1 - c_1,\end{aligned}$$

where c_1 is a centering constant and we define the effects for X_2 analogously. Let,

$$x^j_{\min,1} = \min\{x^j_{i,1}\}_{i=1}^n, \quad x^j_{\min,2} = \min\{x^j_{i,2}\}_{i=1}^n \quad \textit{for} \quad j = H, K.$$

We compute the main effects for each population:

$$\begin{aligned}
f^H_{1,ALE}(x_1) &= \int_{x^H_{\min,1}}^{x_1} \mathbb{E}[X_2 \mid X_1 = z_1]\, dz_1 = \int_{x^H_{\min,1}}^{x_1} 0\, dz_1 = 0, \\
c^H_1 &= \frac{1}{n}\sum_{i=1}^n f^H_{1,ALE}(x^H_{i,1}) = 0, \\
f^K_{1,ALE}(x_1) &= \int_{x^K_{\min,1}}^{x_1} \mu\, dz_1 = \mu(x_1 - x^K_{\min,1}), \\
c^K_1 &= \frac{1}{n}\sum_{i=1}^n \mu(x^K_{i,1} - x^K_{\min,1}) = \mu\left(\frac{1}{n}\sum_{i=1}^n x^K_{i,1} - x^K_{\min,1}\right) \approx \mu(0 - x^K_{\min,1}) = -\mu x^K_{\min,1}, \\
f^H_{2,ALE}(x_2) &= \int_{x^H_{\min,2}}^{x_2} \mathbb{E}[X_1 \mid X_2 = z_2]\, dz_2 = x_2 - x^H_{\min,2}, \\
c^H_2 &= \frac{1}{n}\sum_{i=1}^n f^H_{2,ALE}(x^H_{i,2}) = \frac{1}{n}\sum_{i=1}^n (x^H_{i,2} - x^H_{\min,2}) = \frac{1}{n}\sum_{i=1}^n x^H_{i,2} - x^H_{\min,2} = 0 - x^H_{\min,2} = -x^H_{\min,2}, \\
f^K_{2,ALE}(x_2) &= \int_{x^K_{\min,2}}^{x_2} \mathbb{E}[X_1 \mid X_2 = z_2]\, dz_2 = \int_{x^K_{\min,2}}^{x_2} 0\, dz_2 = 0, \\
c^K_2 &= \frac{1}{n}\sum_{i=1}^n f^K_{2,ALE}(x^K_{i,2}) = \frac{1}{n}\sum_{i=1}^n 0 = 0.
\end{aligned}$$

Centering the ALE Effects

$$\begin{aligned}
\hat{f}^H_{1,ALE}(x_1) &= f^H_{1,ALE}(x_1) - c^H_1 = 0,\\
\hat{f}^H_{2,ALE}(x_2) &= f^H_{2,ALE}(x_2) - c^H_2 = (x_2 - x^H_{\min,2}) - (-x^H_{\min,2}) = x_2,\\
\hat{f}^K_{1,ALE}(x_1) &= f^K_{1,ALE}(x_1) - c^K_1 = \mu(x_1 - x^K_{\min,1}) - (-\mu x^K_{\min,1}) = \mu x_1,\\
\hat{f}^K_{2,ALE}(x_2) &= f^K_{2,ALE}(x_2) - c^K_2 = 0.
\end{aligned}$$

Finally, we compute $\Delta(f, H_X, K_X, S)$:

$$\Delta(f, H_X, K_X, \{1\}) = \mathbb{E}_{H_X}[f^K_{1,\textit{ALE, centered}}(x_1) - f^H_{1,\textit{ALE, centered}}(x_1)] = \mathbb{E}_{H_X}[\mu x_1 - 0] = \mu\mathbb{E}_{H_X}[x_1] = \mu\cdot 1 = \mu \neq 0.$$

$$\Delta(f, H_X, K_X, \{2\}) = \mathbb{E}_{H_X}[f^K_{2,\textit{ALE, centered}}(x_2) - f^H_{2,\textit{ALE, centered}}(x_2)] = \mathbb{E}_{H_X}[0 - x_2] = -\mathbb{E}_{H_X}[x_2] = 0.$$

Therefore, for $S = \{1\}$, ALE misattributes the effects of $Y \mid X$.

B Mathematical framework

In this section, we describe in detail the mathematical framework defined in Section 5, along with the additional necessary notation and assumptions required to prove our main theorem. We also explain the mathematical definition of an admissible perturbation and prove that such perturbations exist for any compactly supported density.

B.1 Additional notation

Let $X \subseteq \mathbb{R}^d$ be a compact set of features and $\mathcal{C}_0(X)$ the set of continuous functions over X. In this work, we focus on continuous compactly supported densities and everywhere positive on X. Let $\mathcal{P}(X)$ represent the space of such probability densities. That is,

$$\mathcal{P}(X) = \left\{ p(x) : p(x) \in \mathcal{C}_0(X), \int_X p(x)\,dx = 1, \text{ and } p(x) > 0, \text{ for all } x \in X \right\}$$

We can think of these densities as the Radon-Nikodym derivative of probability measures that are absolutely continuous with respect to an underlying measure, typically the counting measure or the Lebesgue measure (λ). Here, we focus on the Lebesgue measure, though we conjecture that our work holds for any underlying measure.

We denote by $\mathcal{M}(X)$ the space of Lebesgue measurable functions on the covariates (X) representing flexible regression models for the conditional expectation of $Y \mid X$, that is,

$$\mathcal{M}(X) = \{f : X \to \mathbb{R} | f \text{ is } \lambda\text{-measurable}\}$$

Note that, since functions in $\mathcal{M}(X)$ are λ-measurable, they are also measurable with respect to the probability measures that give rise to densities in $\mathcal{P}(X)$.

We make the following regularity and basic assumptions on the functional decomposition $\mathcal{L}(f, K_X, S)$.

Assumptions 1. *The following hold:*

1. ***Continuity**: For any (K_X, S), the map $f \to \mathcal{L}(f, K_X, S)$ is continuous for almost all $f \in \mathcal{M}(X)$.*
2. ***Measurability**: For any (f, S), the map $K_X \to \mathcal{L}(f, K_X, S)$ is Lebesgue measurable for all $K_X \in \mathcal{P}(X)$.*
3. ***Integrability**: The map $(f, K_X, S) \mapsto \mathcal{L}(f, K_X, S)$ belongs to $L^2(X, \lambda)$, for all $(f, K_X, S) \in \mathcal{M}(X) \times \mathcal{P}(X) \times 2^{[d]}$.*

Where we have denoted by the usual notation $L^2(X, \lambda)$ the space of square integrable functions over X with respect to the Lebesgue measure (λ), as in this work we will focus on the Lebesgue measure only, we omit it from further notation.

B.2 Admissible perturbation functions

To define the admissible perturbation functions mentioned in Section 5, we first need to define Fréchet differentiability (Cheney, 2001).

Definition 4 (Fréchet differentiability). *Let $f : D \to Y$ be a mapping from an open set D in a normed linear space X into a normed linear space Y. Let $x \in D$. If there exists a bounded linear map $A : X \to Y$ such that*

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} = 0,$$

then f is said to be Fréchet differentiable at x, or simply differentiable at x. Furthermore, A is called the (Fréchet) derivative of f at x.

We now define a subspace of functions in $L^2(X)$ that will be useful to define perturbation functions.

Definition 5 (Zero-mean square integrable functions). *We denote by $W(X)$, the space of zero-mean functions in $L^2(X)$ with respect to the Lebesgue measure.*

$$W(X) = \left\{ \phi \in L^2(X) : \int_X \phi(x)\, dx = 0 \right\}.$$

Where the $L^2(X)$ has the usual inner product and norm, $\|\phi\|_{L^2}^2 := \int_X \phi(x)^2\, dx < \infty$.

Definition 6 (Admissible perturbation function). *We say a continuous function $\phi \in L^2(X)$ is an admissible perturbation for a density $K_X \in \mathcal{P}(X)$ if $K_X + \phi \in \mathcal{P}(X)$ and has full support everywhere X.*

We denote by $\mathcal{D}_K$ the set of admissible perturbation functions of K_X: $\mathcal{D}_K = \{\phi \in L^2(X) : K_X + \phi \in \mathcal{P}(X)\}$, and we show that $\mathcal{D}_K \neq \{0\}$ for all $K_X \in \mathcal{P}(X)$.

Lemma 1. *For any density $K_X \in \mathcal{P}(X)$, there exist an admissible perturbation function different than zero.*

Proof. Let any smooth compactly supported function $\psi \in L^2(X)$. Then we can take,

$$\tilde{\phi}(x) = \psi(x) - \frac{1}{\lambda(X)} \int_X \psi(y) dy,$$

such that $\tilde{\phi}(x) \in W(X)$, that is, $\int_X \tilde{\phi}(x)\, dx = 0$. To ensure the positivity requirement, we can take a function $\phi(x) = \varepsilon \tilde{\phi}(x)$, for $\varepsilon > 0$, which still is in $L^2(X)$ and integrates to zero. Such $\varepsilon > 0$ must satisfy that for a given density $K_X(x)$,

$$K_X(x) + \varepsilon \tilde{\phi}(x) > 0 \iff \varepsilon \tilde{\phi}(x) > -K_X(x),\ \forall\, x \in X. \tag{8}$$

Whenever $\tilde{\phi}(x) > 0$, Equation 8 is always satisfied. Thus, the only relevant case is when $\tilde{\phi}(x) < 0$, for which Equation 8 is satisfied if and only if,

$$\varepsilon < \frac{-K_X(x)}{\tilde{\phi}(x)},\ \forall\, x \in X \text{ such that } \tilde{\phi}(x) < 0.$$

Or equivalently,

$$\varepsilon \leq \frac{\inf_{x \in X} K_X(x)}{\sup_{x \in X} |\tilde{\phi}(x)|},$$

where by assumption the right hand side is strictly positive. Thus $\phi(x)$ is an admissible perturbation function of $K_X(x)$. □

Note that given a probability density K_X, we can parameterize the functional decomposition in terms of $\phi(x)$ as follows: $\mathcal{L}(f, \phi, S) = \mathcal{L}(f, K_X + \phi, S) : \mathcal{M}(X) \times \mathcal{D}_{\mathcal{K}}(X) \times 2^{[d]} \to L^2(\mathbb{R}^S)$. For this parameterization, in addition to Assumption 1, we need to assume the continuous differentiability of $\mathcal{L}$ as a function of ϕ (see Assumption 2) to ensure that $\mathcal{L}$ is Fréchet differentiable as a map from the Banach space $L^2(X)$ into the Banach space $L^2(X_S)$ (Zeidler, 1986; Averbukh and Smolyanov, 1967), where $X_S \in X$ denotes the subset of covariates indexed by S .

Assumptions 2. *The map $\phi \to \mathcal{L}(\cdot, \phi(x))$ is continuously differentiable as a map from $L^2(X)$ into $L^2(X_S)$.*

Under this new assumption, we can linearly approximate $\mathcal{L}(\cdot, \phi)$ around $\phi = 0$ with a linear and bounded functional.

$$\mathcal{L}(\cdot, \phi) = \mathcal{L}(\cdot, 0) + D_\phi \mathcal{L}(\cdot, 0)[\phi] + o(\|\phi\|_{L^2}).$$

Where $D_\phi \mathcal{L}(\cdot, 0)[\phi]$ is the Fréchet derivative of $\mathcal{L}$ with respect to the function ϕ evaluated at the zero function and $o(\|\phi\|_{L^2})$ represents a higher-order functional that vanishes faster than $\|\phi\|_{L^2}$ as $\phi \to 0$. More formally, for any $\delta > 0$, there exists a $\tau > 0$ such that if $\|\phi\|_{L^2} < \tau$, then $|o(\|\phi\|_{L^2})| \leq \delta \|\phi\|_{L^2}$.

Remark 1. *The Fréchet derivative is a linear and bounded functional which operates on functions $\phi \in L^2(X)$. That is, there exist a constant $C > 0$ such that,*

$$\|D_\phi \mathcal{L}(\cdot, 0)[\phi]\|_{L^2} \leq C \|\phi\|_{L^2}$$

B.3 Proof of Theorem 1

We first show some lemmas that will be useful through the proof of Theorem 1.

Lemma 2. *Given our assumptions, for any $K_X \in \mathcal{P}(X)$ and $\phi \in W(X)$, the following integrals are finite.*

$$\left| \int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\phi](x)\right)\ \phi(x)\, dx \right| < \infty, \tag{9}$$

$$\left| \int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\phi](x)\right) \phi(x)\, K_X(x)\, dx \right| < \infty. \tag{10}$$

Furthermore,

$$\left| \int_X o(\|\phi\|_{L_2})(x)(K_X(x) + \phi(x))\, dx \right| = o(\|\phi\|_{L^2}) \tag{11}$$

Proof. K_X is continuous and compactly supported on X, then by a direct consequence of the extreme value theorem, it is bounded: there exists a $B > 0$ such that $\sup_{x\in X} |K_X(x)| \leq B_k < \infty$; by a similar argument, $\sup_{x\in X} |\phi(x)| \leq B_\phi < \infty$. We first show Equation 9:

$$\begin{aligned}\left| \int_X \left(D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right) K_X(x)\, dx \right| &\leq \int_X \left|D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right| K_X(x)\, dx \\ &\leq \left(\int_X \left(D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right)^2\, dx \right)^{1/2} \left(\int_X K_X(x)^2\, dx \right)^{1/2} \\ &\leq C \cdot \|\phi\|_{L^2} \cdot B_K \cdot \sqrt{\lambda(X)} \\ &\leq C \cdot B_\phi \cdot B_K \cdot \lambda(X) < \infty.\end{aligned}$$

To show Equation 10:

$$\begin{aligned}\left| \int_X \left(D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right) \phi(x)\, dx \right| &\leq \int_X \left|D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right| |\phi(x)|\, dx \\ &\leq \left(\int_X \left(D_\phi \mathcal{L}(\cdot; 0)[\phi](x)\right)^2\, dx \right)^{1/2} \left(\int_X \phi(x)^2\, dx \right)^{1/2} \\ &\leq C \cdot \|\phi\|_{L^2} \cdot B_\phi \sqrt{\lambda(X)} \\ &\leq B_\phi^2 \cdot C \cdot \lambda(X).\end{aligned}$$

To show Equation 11: For any $\delta > 0$, there exist a $\tau > 0$ such that if $\|\phi\|_{L_2} < \tau$, then $o(\|\phi\|_{L^2}) \leq \delta \|\phi\|_{L_2}$, thus:

$$\begin{aligned}\left| \int_X o(\|\phi\|_{L_2})(x)(K_X(x) + \phi(x))\, dx \right| &\leq \int_X |o(\|\phi\|_{L_2})(x)|(K_X(x) + \phi(x))\, dx \\ &\leq (B_K + B_\phi) \int_X |o(\|\phi\|_{L_2})|\, dx \\ &\leq (B_K + B_\phi) \cdot \delta \|\phi\|_{L_2} \lambda(X) \\ &= o(\|\phi\|_{L_2}).\end{aligned}$$

□

Lemma 3. *Let $X \subset \mathbb{R}^d$ be a measurable set with finite Lebesgue measure $\lambda(X) < \infty$. Then, the orthogonal complement of $W(X)$ in $L^2(X)$ is the space of constant functions on X; that is,*

$$W(X)^\perp = \left\{ f \in L^2(X) : f(x) = c, \text{ a.e. on } X \right\}.$$

Proof. Let

$$V = \left\{ f \in L^2(X) : f(x) = c \text{ a.e. on } X \right\}.$$

We will prove that $W(X)^\perp = V$ by first showing that $V \subseteq W(X)^\perp$. Let $f \in V$; then, for any $\psi \in W(X)$, we have

$$\int_X f(x)\psi(x)\, dx = c \int_X \psi(x)\, dx = 0.$$

It remains to show that $W(X)^\perp \subseteq V$. Let $f \in W(X)^\perp$, then $\int_X f(x)\psi(x)\, dx = 0$ for any $\psi(x) \in W(X)$. In particular, we can take an arbitrary measurable set $A \subset X$ and define

$$\psi_A(x) = \chi_A(x) - \frac{\lambda(A)}{\lambda(X)}$$

where $\chi_A(x)$ is the indicator function over A and λ is the Lebesgue measure. Thus,

$$\begin{aligned} 0 = \int_X f(x)\psi_A(x)\,dx &= \int_X f(x)\chi_A(x)\,dx - \int_X f(x)\frac{\lambda(A)}{\lambda(X)}\,dx \\ \Leftrightarrow \int_A f(x)\,dx &= \int_X f(x)\frac{\lambda(A)}{\lambda(X)}\,dx = \lambda(A)\left(\frac{\int_X f(x)\,dx}{\lambda(X)}\right) \end{aligned} \tag{12}$$

Define $\mu(A) = \int_A f(x)\,dx$, which is a signed measure absolutely continuous with respect to the Lebesgue measure. On one hand, by the Radon-Nikodym Theorem for signed measures (Folland (1999); Theorem 3.8), $f(x)$ is the Lebesgue integrable Radon-Nikodym derivative. On the other, by Equation 12:

$$\mu(A) = \lambda(A) \cdot c, \text{ for any measurable set A} \subset X, \tag{13}$$

where $c = \left(\frac{\int_X f(x)\,dx}{\lambda(X)}\right)$. By the Lebesgue almost everywhere uniqueness of the Radon-Nikodym derivative, we have form Equation 13 and definition of μ that

$$f(x) = c, \text{ a.e. } x \in X.$$

Therefore, $f \in V$ and $W(X)^{\perp} \subseteq V$. □

We can now proceed to prove our main theorem. This is a work in progress, but there are several approaches to consider for this statement. Some might require additional constraints, or we may need to restrict to specific spaces.

Theorem 1. *Take Assumptions 1 and 2. Let $K_X \in \mathcal{P}(X)$. If,*

$$\mathbb{E}_{x\sim K_X+\phi}\left[\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x)\right] = 0, \textit{ for all } \phi \in \mathcal{D}_K.$$

Then,

$$D_\phi \mathcal{L}(\cdot, 0)[\phi] = 0, \textit{ for all } \phi \in \mathcal{D}_K,$$

and therefore $\mathcal{L}(f, K_X, S)$ is invariant under perturbations of concentration. That is,

$$\mathcal{L}(f, K_X, S) = \mathcal{L}(f, K_X + \phi, S), \textit{ for all } \phi \in \mathcal{D}_K.$$

Proof. By assumption $\mathbb{E}_{x\sim K_X+\phi}\left[\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x)\right] = 0,$ for all $\phi \in \mathcal{D}_K$. i.e.,

$$\begin{aligned} 0 &= \int_X \left(\mathcal{L}(f, K_X, S)(x) - \mathcal{L}(f, K_X + \phi, S)(x)\right)\left(K_X(x) + \phi(x)\right)dx, \\ &= \int_X \left(\mathcal{L}(\cdot, 0)(x) - \mathcal{L}(\cdot, \phi)(x)\right)\left(K_X(x) + \phi(x)\right)dx, \\ &= -\int_X \left[D_\phi \mathcal{L}(\cdot, 0)[\phi](x) + o(\|\phi\|_{L^2}(x))\right]\left(K_X(x) + \phi(x)\right)dx, \\ \Longleftrightarrow\ 0 &= \int_X \left[D_\phi \mathcal{L}(\cdot, 0)[\phi](x) + o(\|\phi\|_{L^2})(x)\right]\left(K_X(x) + \phi(x)\right)dx. \end{aligned} \tag{14}$$

Then, by Lemma 2, we can split the integrals, and rewrite Equation 14 as:

$$\int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\phi](x)\right)K_X(x)\,dx + \int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\phi](x)\right)\phi(x)\,dx = -\int_X o(\|\phi\|_{L^2})(x)(K_X(x) + \phi(x))\,dx.$$

Since this equation must hold for all $\phi \in \mathcal{D}_K$, we can proceed as in the proof of Lemma 1. Specifically, let $\phi(x) = \varepsilon\psi(x)$ for sufficiently small $\varepsilon > 0$ and $\psi(x) \in W(X)$, where $W(X)$ is the set of mean zero functions in L^2 (see Definition 5). Furthermore, by Lemma 2, we know the following: $\int_X o(\|\phi\|_{L^2})(x)(K_X(x) + \phi(x))\,dx = o(\|\phi\|_{L^2})$. Note also that $o(\|\varepsilon\psi\|_{L^2}) = o(\varepsilon\|\psi\|_{L^2}) = o(\varepsilon)$ since $\|\psi\|_{L^2} < \infty$, then the above equation simplifies to:

$$\int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\varepsilon\psi](x)\right)K_X(x)\,dx + \int_X \left(D_\phi \mathcal{L}(\cdot, 0)[\varepsilon\psi](x)\right)\varepsilon\psi(x)\,dx = o(\varepsilon).$$

Where by $o(\varepsilon)$ we mean a constant that goes to zero faster than ε. By linearity of the Fréchet derivative, we can take ε out of the operator, divide by it, and since $\frac{o(\varepsilon)}{\varepsilon} = o(1)$, we obtain:

$$\int_X \left(D_\phi \mathcal{L}(\cdot,0)[\psi](x)\right) K_X(x)\, dx + \varepsilon \int_X (D_\phi \mathcal{L}(\cdot,0)[\psi](x))\psi(x)\, dx = o(1).$$

Taking $\varepsilon \to 0$, we get that the first integral is equal to zero:

$$\int_X \left(D_\phi \mathcal{L}(\cdot,0)[\psi](x)\right) K_X(x)\, dx = 0. \tag{15}$$

Equation 15, states that for any ψ, the evaluation functional $D_\phi \mathcal{L}(\cdot,0)[\psi](x))$ is orthogonal to the density K_X.

The proof is not yet complete, but we observe a very specific constraint on the functional $\mathcal{L}$, namely that its derivative must be zero when weighted by the density function. There are potential avenues to explore here, such as working with RKHS to investigate specific forms of kernels and determine in which cases such kernels would be equivalent to the zero function or impose additional constraints on the functional. These constraints would still apply in many cases and indicate that the operator is locally constant.

$\square$