

## FIXED-POINT ANALYSIS OF WEIGHTED INTEGRATED GRADIENTS

SAIDJON KAMOLOV<sup>1\*</sup>

**ABSTRACT.** Feature attribution methods like Integrated Gradients (IG) are widely used to interpret deep neural networks. While typical extensions of IG introduce continuous weighting kernels along the integration path, these approaches are usually evaluated based on static axiomatic properties. We propose a theoretical framework analyzing Weighted Integrated Gradients (WIG) as a continuous mathematical operator acting on the model's function space. By investigating the sequence generated when iteratively applying the explanation to itself, we use Fixed-Point Theory and Taylor spectral decomposition to show that WIG functions as a spectral filter on model complexity. Our analysis identifies three regimes. First, standard IG acts as an identity operator, effectively functioning as an all-pass filter. Second, input-weighted explanations act as expansion operators or high-pass filters, amplifying higher-order nonlinearities. Third, baseline-weighted explanations act as contraction mappings or low-pass filters. Iterative application of a baseline-weighted explanation operator converges to a linear surrogate, establishing a formal equivalence between baseline-weighted attribution and model distillation.

*Keywords.* Explainable AI; Integrated Gradients; Feature Attribution; Fixed-Point Theory; Spectral Decomposition; Operator Theory; Knowledge Distillation.

*2020 Mathematics Subject Classification.* Primary 68T07; Secondary 47H10

### 1. INTRODUCTION

The deployment of complex deep learning architectures has motivated the ongoing development of explainability techniques. Integrated Gradients (IG) is a well-established method for computing feature attributions by integrating the gradients of a model along a linear path from a baseline input to the instance of interest [18]. The method is frequently utilized due to its adherence to axiomatic properties such as sensitivity and completeness.

Various modifications of IG apply non-uniform weighting distributions along the integration path. Weighted Integrated Gradients (WIG) enable smooth interpolation between global, baseline-centric explanations and local, gradient-centric

---

*Date:* Received: Dec 29, 2025; Revised: Feb 25, 2026; Accepted: Mar 17, 2026.

\* Corresponding author

© The Author(s) 2025. This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of the licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

explanations. Much of the theoretical analysis of WIG to date has focused on geometric properties and cooperative game theory formulations.

Our work reframes feature attribution by treating the explanation process as a continuous mathematical operator that maps an original neural network to a new scalar field representing the sum of its attributions. This framing enables an analysis of the iterated sequence generated when repeatedly explaining a model.

Using Fixed-Point Theory alongside Taylor spectral decomposition, we compute the exact eigenvalues of the WIG operator. This leads to a categorization of weighting functions based on their spectral filtering behavior. We mathematically demonstrate that specific WIG formulations natively function as iterative distillation operators, thereby connecting local interpretability mechanisms with global model compression.

## 2. RELATED WORK

The formalization of feature attribution as a cooperative game is heavily grounded in the Aumann-Shapley value [1]. Integrated Gradients [18] provides a continuous analog to Shapley values for differentiable models. Subsequent work has expanded upon this axiomatic framework; for example, Expected Gradients [6] incorporates baseline distributions to improve attribution priors, while additive importance measures have been formalized to unify local and global feature contributions [4].

Extensions to standard IG often adjust the integration schedule or blur the baseline to account for scale and spatial variations [20]. The impact of baseline selection on feature attributions has been systematically evaluated to understand its effect on explanation variance [17]. Guided Integrated Gradients adaptively conditions the integration path to mitigate baseline noise [14]. Discretized versions have also been proposed to handle non-continuous spaces like language [15]. Recent adaptations introduce path-sampled weighting schemes [13] and path-weighted integrated gradients tailored for medical domains such as dementia classification [12]. Other empirical schemes seek to stabilize shattered gradients directly through noise injection [16].

The structural vulnerabilities of feature attributions also require careful consideration. Explanations can be extremely fragile to adversarial perturbations [7], and specific geometric properties of neural networks allow explanations to be manipulated without altering model predictions [5]. These limitations have prompted researchers to more rigorously define faithfulness and interpretability evaluation metrics in complex architectures like transformers [8, 3].

Methodologically, our analysis draws upon fixed-point theory in machine learning. Fixed-point iterations are central to the analysis of implicit layers in deep equilibrium models [2] and monotone operator networks [19]. The rigorous mathematical formulation of machine learning processes has seen increasing attention across various domains, ranging from systematic reviews of mathematical methods in feature selection [11] to the analytical modeling of data sampling techniques like SMOTE [9, 10]. However, the application of fixed-point contraction mappings

directly to the functional space of post-hoc explanations has not been previously formalized.

### 3. THE EXPLANATION OPERATOR FRAMEWORK

Let  $\mathcal{F}$  be the space of continuously differentiable scalar functions  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ . We normalize the input space such that the baseline  $x'$  is positioned at the origin, and  $F(0) = 0$ .

Weighted Integrated Gradients compute the attribution for feature  $i$  by integrating the partial derivative along the path  $\tau x$ , modified by a continuous probability density function  $\omega(\tau)$  on  $[0, 1]$ :

$$\text{WIG}_i(x, 0; \omega) = x_i \int_0^1 \omega(\tau) \frac{\partial F(\tau x)}{\partial x_i} d\tau. \quad (3.1)$$

Summing the attributions over all  $d$  features maps the original model  $F$  to a new scalar field, producing the total attribution map at  $x$ . We define this mapping as the Explanation Operator.

**Definition 3.1.** For a given weight function  $\omega(\tau)$ , the Explanation Operator  $T_\omega : \mathcal{F} \rightarrow \mathcal{F}$  is:

$$T_\omega[F](x) := \sum_{i=1}^d \text{WIG}_i(x, 0; \omega) = \int_0^1 \omega(\tau) \langle \nabla F(\tau x), x \rangle d\tau. \quad (3.2)$$

A fixed point of this operator is a model  $F^*$  satisfying  $T_\omega[F^*] = F^*$ . Such a model is exactly reconstructed by its total WIG attribution map.

### 4. SPECTRAL DECOMPOSITION OF FEATURE ATTRIBUTIONS

Analyzing the dynamics of iteratively applying  $T_\omega$ , where  $F_{n+1} = T_\omega[F_n]$ , requires analyzing its spectral properties. Under Taylor's theorem, a continuously differentiable model  $F(x)$  can be expanded into a sum of homogeneous polynomials  $F(x) = \sum_{k=1}^{\infty} F^{(k)}(x)$ , with  $F^{(k)}$  being homogeneous of degree  $k$ . Consequently,  $F^{(k)}(\lambda x) = \lambda^k F^{(k)}(x)$  for any scalar  $\lambda$ .

These homogeneous polynomials natively diagonalize the Explanation Operator.

**Lemma 4.1.** *Every homogeneous function  $F^{(k)}(x)$  of degree  $k \geq 1$  is an eigenfunction of the Explanation Operator  $T_\omega$ , with corresponding eigenvalue:*

$$\lambda_k(\omega) = \int_0^1 \omega(\tau) k \tau^{k-1} d\tau = k \mathbb{E}_{\tau \sim \omega}[\tau^{k-1}]. \quad (4.1)$$

*Proof.* According to Euler's Homogeneous Function Theorem, the inner product of the gradient and the input vector is  $k$  times the function value:  $\langle \nabla F^{(k)}(x), x \rangle = k F^{(k)}(x)$ . The gradient vector field  $\nabla F^{(k)}$  is itself homogeneous of degree  $k - 1$ .

Applying the operator yields:

$$\begin{aligned} T_\omega[F^{(k)}](x) &= \int_0^1 \omega(\tau) \langle \nabla F^{(k)}(\tau x), x \rangle d\tau \\ &= \int_0^1 \omega(\tau) \tau^{k-1} \langle \nabla F^{(k)}(x), x \rangle d\tau \\ &= \left( k \int_0^1 \omega(\tau) \tau^{k-1} d\tau \right) F^{(k)}(x) = \lambda_k(\omega) F^{(k)}(x). \end{aligned}$$

□

This result demonstrates that the explanation operator acts independently across different polynomial degrees of the model’s function space.

## 5. FIXED-POINT THEOREMS AND FILTERING REGIMES

Since  $T_\omega$  operates linearly on  $\mathcal{F}$ , the convergence of the sequence  $F_n = T_\omega^n[F]$  is determined by the magnitude of the eigenvalues  $\lambda_k(\omega)$ . Following the Banach Fixed Point Theorem, the shape of the weighting kernel  $\omega(\tau)$  dictates the mapping characteristics. We categorize these into three spectral regimes.

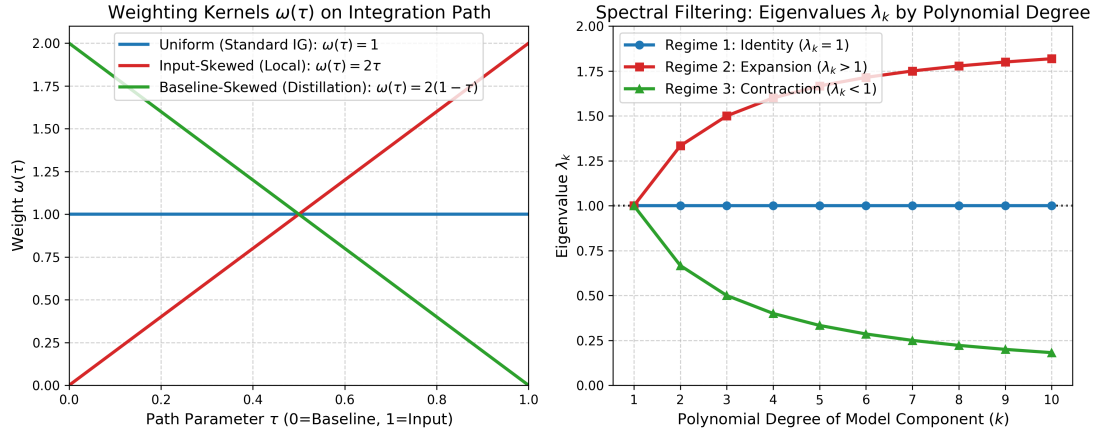


FIGURE 1. Left: The three families of WIG weighting kernels. Right: The corresponding eigenvalues  $\lambda_k$ . Note that local weighting acts as a high-pass filter ( $\lambda_k > 1$ ), while baseline weighting acts as a low-pass filter ( $\lambda_k < 1$  for  $k \geq 2$ ).

**5.1. Standard IG Fixed Point.** The standard Integrated Gradients formulation uniformly weights the integration path, which corresponds to  $\omega(\tau) = 1$ .

**Theorem 5.1.** *For uniform weighting  $\omega(\tau) = 1$ , the Explanation Operator  $T_{uniform}$  is the identity operator on  $\mathcal{F}$ .*

*Proof.* Applying Lemma 4.1 with  $\omega(\tau) = 1$  evaluates the eigenvalues as:

$$\lambda_k(\text{Uniform}) = k \int_0^1 \tau^{k-1} d\tau = k \left[ \frac{\tau^k}{k} \right]_0^1 = 1 \quad \forall k \geq 1.$$

Because all eigenvalues are 1,  $T_{\text{uniform}}[F] = F$  for all functions.  $\square$

Any continuously differentiable function is a stable fixed point of standard IG. From a signal processing perspective, it acts as an all-pass filter, preserving all orders of non-linearity. This provides a spectral verification of the completeness axiom associated with standard IG.

**5.2. Local Weighting and Expansion.** Weighting functions skewed toward  $\tau = 1$  emphasize local gradients near the input. Using a linear example of  $\omega(\tau) = 2\tau$ , the operator’s behavior changes entirely.

**Theorem 5.2.** *For input-skewed weighting  $\omega(\tau) = 2\tau$ , the Explanation Operator acts as a strict expansion mapping on the non-linear components of  $F$ .*

*Proof.* The eigenvalues evaluate to:

$$\lambda_k(\text{Local}) = k \int_0^1 (2\tau)\tau^{k-1} d\tau = \frac{2k}{k+1}.$$

For linear components where  $k = 1$ ,  $\lambda_1 = 1$ . For quadratic components where  $k = 2$ ,  $\lambda_2 = \frac{4}{3} > 1$ . In the limit as  $k \rightarrow \infty$ ,  $\lambda_k \rightarrow 2$ .  $\square$

Because  $\lambda_k > 1$  for all non-linear terms, an iterative application of input-weighted IG inflates the high-degree complexities of the model. The operator functions as a high-pass filter. Focusing strictly on local gradients creates an intrinsic instability in the explanations of highly non-linear models.

**5.3. Baseline Weighting and Distillation.** Functions skewed toward  $\tau = 0$  suppress local gradients and instead prioritize regions near the baseline. A linear example is  $\omega(\tau) = 2(1 - \tau)$ . The eigenvalues for this regime are:

$$\lambda_k(\text{Baseline}) = k \int_0^1 2(1 - \tau)\tau^{k-1} d\tau = 2k \left( \frac{1}{k} - \frac{1}{k+1} \right) = \frac{2}{k+1}. \quad (5.1)$$

For linear terms,  $\lambda_1 = 1$ . However, for all non-linear terms where  $k \geq 2$ ,  $\lambda_k < 1$ . As  $k \rightarrow \infty$ ,  $\lambda_k \rightarrow 0$ .

This contraction property leads to our main theorem, proving that baseline-weighted explanations naturally perform knowledge distillation.

**Theorem 5.3.** *Let  $\omega(\tau)$  be a strictly decreasing probability density function on  $[0, 1]$  satisfying  $\int_0^1 \omega(\tau) d\tau = 1$ . Let  $F \in \mathcal{F}$  be a non-linear model. The Explanation Operator  $T_\omega$  is a strict contraction mapping on the non-linear components of  $F$ .*

*The sequence of iterated explanations  $F_n = T_\omega^n[F]$  converges uniformly to a unique fixed point  $F^*$ :*

$$\lim_{n \rightarrow \infty} T_\omega^n[F](x) = F^{(1)}(x) = \langle \nabla F(0), x \rangle, \quad (5.2)$$

*which is exactly the first-order linear approximation of the model evaluated at the baseline.*

*Proof.* A strictly decreasing  $\omega(\tau)$  places strictly more probability mass on lower values of  $\tau$ . Thus,  $\mathbb{E}_\omega[\tau^{k-1}] < \mathbb{E}_{\text{uniform}}[\tau^{k-1}] = \frac{1}{k}$  for all  $k \geq 2$ . It follows that  $\lambda_k(\omega) = k\mathbb{E}_\omega[\tau^{k-1}] < 1$  for all  $k \geq 2$ .

Since  $\lambda_1 = 1 \cdot \int_0^1 \omega(\tau) d\tau = 1$ , the linear component  $F^{(1)}$  is preserved. Because  $\lambda_k < 1$  for all  $k \geq 2$ , the operator is a strict contraction on the orthogonal complement. Following the Banach Fixed Point Theorem, applying the operator  $n$  times scales the magnitude of  $F^{(k)}$  by  $(\lambda_k)^n$ . As  $n \rightarrow \infty$ , the non-linear terms decay to zero, isolating the invariant linear projection  $F^{(1)}(x)$ .  $\square$

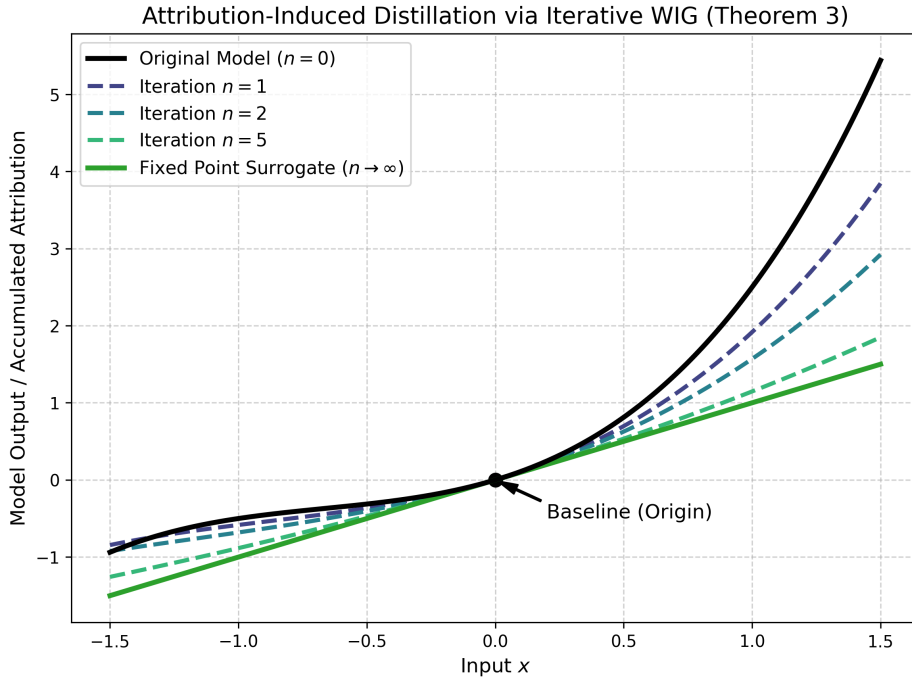


FIGURE 2. Visualization of Theorem 3. Iteratively applying a baseline-weighted Explanation Operator systematically contracts the non-linear components of the original model (black), eventually collapsing the model into its perfect linear surrogate (green).

## 6. DISCUSSION AND IMPLICATIONS FOR XAI

Framing feature attributions as spectral filters provides a distinct mechanism for analyzing interpretability outputs. Theorem 3 shows that baseline-weighted IG functions as a low-pass filter on model complexity. Repeatedly querying the explanation smooths out the shattered decision boundaries typical of deep networks. The fixed point of this procedure is a linear proxy model, mathematically linking post-hoc feature attribution with model distillation techniques.

Conversely, the analysis of the local weighting regime highlights why gradient-based methods evaluated near the input often appear excessively noisy. Input-weighted kernels act as high-pass filters because their eigenvalues exceed 1, amplifying high-frequency variations in the model’s geometry.

These findings suggest that  $\omega(\tau)$  functions can be constructed to yield specific spectral properties. Similar to filter design in signal processing, attribution kernels can be optimized to isolate or suppress particular polynomial degrees of complexity within a neural network.

## 7. CONCLUSION

We formalized the Iterated Explanation Operator, establishing connections between fixed-point theory, spectral filtering, and feature attribution. By projecting Weighted Integrated Gradients onto the polynomial eigenfunctions of the input space, we demonstrated that the choice of weighting along the integration path determines whether an explanation preserves, amplifies, or distills the underlying non-linearities of a model.

## REFERENCES

- [1] R. J. Aumann and L. S. Shapley. *Values of Non-Atomic Games*. Princeton University Press, 1974. DOI: <https://doi.org/10.1515/9781400881964>
- [2] S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. DOI: <https://doi.org/10.48550/arXiv.1909.01377>
- [3] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. DOI: <https://doi.org/10.1109/CVPR46437.2021.00084>
- [4] I. Covert, S. M. Lundberg, and S.-I. Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223, 2020. DOI: <https://doi.org/10.48550/arXiv.2003.00687>
- [5] A.-K. Dombrowski, M. Alber, C. J. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32, 2019. DOI: <https://doi.org/10.48550/arXiv.1906.07983>
- [6] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3(8):620–631, 2021. DOI: <https://doi.org/10.1038/s42256-021-00343-w>
- [7] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019. DOI: <https://doi.org/10.1609/aaai.v33i01.33013681>
- [8] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020. DOI: <https://doi.org/10.18653/v1/2020.acl-main.386>
- [9] F. Kamalov. Asymptotic behavior of SMOTE-generated samples using order statistics. *Gulf Journal of Mathematics*, 17(2):327–336, 2024. DOI: <https://doi.org/10.56947/gjom.v17i2.2343>
- [10] F. Kamalov, S. E. Choutri, and A. F. Atiya. Analytical formulation of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Gulf Journal of Mathematics*, 19(1):400–415, 2025. DOI: <https://doi.org/10.56947/gjom.v19i1.2639>
- [11] F. Kamalov, H. Sulieman, A. Alzaatreh, M. Emarly, H. Chamlal, and M. Safaraliev. Mathematical methods in feature selection: A review. *Mathematics*, 13(6):996, 2025. DOI: <https://doi.org/10.3390/math13060996>

- [12] F. Kamalov, M. A. Falasi, and F. Thabtah. Path-weighted integrated gradients for interpretable dementia classification. *arXiv preprint arXiv:2509.17491*, 2025. DOI: <https://doi.org/10.48550/arXiv.2509.17491>
- [13] F. Kamalov, F. Thabtah, R. Sivaraaj, and N. Abdelhamid. Path-sampled integrated gradients. *Gulf Journal of Mathematics*, 22(1):1–10, 2026. DOI: <https://doi.org/10.56947/gjom.v22i1.4141>
- [14] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5058, 2021. DOI: <https://doi.org/10.1109/CVPR46437.2021.00501>
- [15] S. Sanyal and X. Ren. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, 2021. DOI: <https://doi.org/10.18653/v1/2021.emnlp-main.805>
- [16] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. DOI: <https://doi.org/10.48550/arXiv.1706.03825>
- [17] P. Sturmfels, S. Lundberg, and S.-I. Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. DOI: <https://doi.org/10.23915/distill.00022>
- [18] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. DOI: <https://doi.org/10.48550/arXiv.1703.01365>
- [19] E. Winston and J. Z. Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 10718–10728, 2020. DOI: <https://doi.org/10.48550/arXiv.2002.08587>
- [20] R. Xu, P. Gao, J. C. Chen, E. Tekin, and Y. N. Wu. Attribution in scale and space: A path-based method for interpreting deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020. DOI: <https://doi.org/10.1109/CVPR42600.2020.00970>

<sup>1</sup>FACULTY OF ENGINEERING, TAJIK TECHNICAL UNIVERSITY, DUSHANBE, TAJIKISTAN.  
Email address: [said.kamolov@yahoo.com](mailto:said.kamolov@yahoo.com); [said.kamolov@ttu.tj](mailto:said.kamolov@ttu.tj)