Toward a Principled Theory of XAI via Spectral Analysis

Amir Mehrpanah[1], Matteo Gamba[1], Hossein Azizpour[1,2]
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]SciLifeLab, Stockholm, Sweden
{amirme, mgamba, azizpour}@kth.se

Abstract

Establishing trust in AI systems requires explanations for their decisions. However, AI models that are practically effective often express highly nonlinear functions of the data, in turn resulting in complex explanations. Humans, by contrast, have a cognitive preference for low-complexity explanations. Consequently, there have been various efforts to simplify explanations of non-linear models.

*A central dilemma in Explainable AI (XAI) arises at this point: should simplification be pursued ante-hoc (*i.e.*, designing models yielding simple explanations once trained) or post-hoc (*i.e.*, designing explanation methods that work with arbitrarily complex models)?*

Crucially, both strategies rely heavily on heuristics and implicit assumptions, lacking a rigorous theoretical foundation. This prevents a principled analysis of the fundamental trade-offs inherent to XAI.

This position paper advocates for spectral analysis as a promising framework for deeper theoretical analysis of XAI. Using image data as a case study, we examine the challenges of both ante- and post-hoc approaches and outline future research directions.

We retrospectively analyze both approaches, uncovering their implicit assumptions via two fundamental questions: the source of explanation complexity and the necessity of model complexity for a task. These questions provide a principled basis for choosing between the two approaches.

Regardless of one's stance, we argue that spectral methods offer a valuable foundation for formal XAI and can inform efforts across other modalities.

1. Introduction

Explainability is a fundamental requirement for deploying deep learning models in sensitive domains such as healthcare, autonomous driving, and legal decision-making. As black-box models such as deep neural networks become increasingly powerful, the need to understand their internal mechanisms and decisions has never been more critical [2].

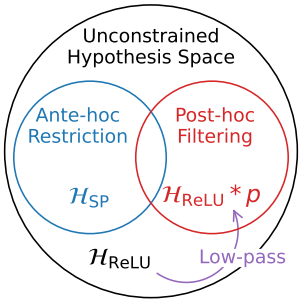
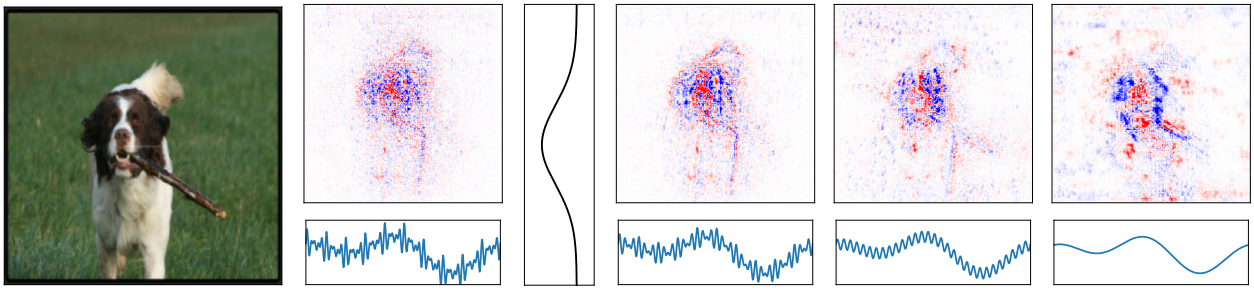


Figure 1. **Post-hoc Filtering *vs*. Ante-hoc Restriction of Hypothesis Space.** This Venn diagram shows the relationships between post-hoc filtering (b→c→d in Fig. 2), and ante-hoc, (e and f in Fig. 2), restriction of the hypothesis space (see Sec. 2 for discussion). Post-hoc methods search a broad space $\mathcal{H}_{\text{ReLU}}$ (black set) and then apply a low-pass filter via perturbation (purple arrow) to remove complexity (red set), while ante-hoc methods restrict $\mathcal{H}_{\text{ReLU}}$ to low-frequency functions before training (blue set). Note that depending on specific design parameters, the resulting spaces may be nested subsets of each other.

As models become more intricate, their explanations often grow so elaborate, surpassing human cognitive limits [8]. Empirical studies further show that humans consistently favor concise accounts over detailed yet unwieldy ones [15]. In response, the XAI literature has converged on two broad approaches for reducing attribution complexity [25], each grounded in a distinct philosophical stance:

- **Ante-hoc simplification (model-centric)** Designing architectures or training regimes that induce simpler functions (see $\mathcal{H}_{\text{SP}}$ in Fig. 1). The guiding assumption is that a simpler model will naturally yield simpler explanations, sacrificing expressivity for interpretability.
- **Post-hoc simplification (explanation-centric)** Such techniques approximate the complex model in a lower-complexity space by manipulating the attribution method ($\mathcal{H}_{\text{ReLU}} * p$ in Fig. 1). They rely on heuristics and implicit surrogates with typically unknown properties (*e.g.*, performance, faithfulness). The underlying belief is that faithfulness can be sacrificed for interpretability.



(a) Input (b) VG - ReLU (c) p (d) SG - ReLU (e) VG - $\mathrm{SP}_{\beta=10}(x)$ (f) VG - $\mathrm{SP}_{\beta=5}(x)$

Figure 2. **Examples From Post-hoc and Ante-hoc.** This figure shows the examples of post-hoc and ante-hoc methods for controlling the frequency content of attribution maps. Panel (a) shows the input image. Panel (b) displays the attribution map produced by VanillaGrad [21] (VG), which exhibits high-frequency noise. Panel (d) shows the result of post-hoc smoothing using SmoothGrad [22] (SG), which effectively reduces high-frequency components. This smoothing can be interpreted as a low-pass filter, implemented using a Gaussian kernel, as visualized in (c) and analyzed in [16]. In contrast, panel (e) presents attribution maps generated by a model whose hypothesis space has been restricted ante-hoc—prior to training—through an architectural choice, in this case, a smoother activation, SoftPlus (SP). Such models tend to produce inherently smoother explanations without requiring post-hoc filtering. The plots beneath (b), (d), and (e) are typical 1D examples representing functions belonging to the corresponding hypothesis spaces: $\mathcal{H}_{\mathrm{ReLU}}$, $\mathcal{H}_{\mathrm{ReLU}} * p$, and $\mathcal{H}_{\mathrm{SP}(\beta)}$ (see Fig. 1).

Both strategies share the same objective–making deep-networks' local behavior human interpretable—yet they proceed from fundamentally different premises about where the explanation complexity originates from, and where it should be controlled: in the model itself or in the interpretive lens applied afterward.

We analyze the trade-offs involved in each of these approaches and argue that the post-hoc *vs*. ante-hoc dilemma depends ones' position on two fundamental questions, regarding the *source and necessity of complexity*.

These questions do not favor either approach but seek to reveal their underlying assumptions and provide a principled basis for the post-hoc *vs*. ante-hoc dilemma.

We focus on image data and ReLU networks, given their strong theoretical basis [18], though the arguments generalize to other modalities.

We argue that the spectral perspective enables deeper theoretical analysis, offering a promising framework to address limitations in current XAI research.

Outline In Sec. 2, we review post-hoc methods through the lens of spectral analysis, identifying a unifying principle: *most operate as implicit low-pass filters on the network.*

In Sec. 3, we leverage a spectral analysis that characterize the relationship between complexity of networks and their attributions. Showing that explanation complexity originates from the model and not the explanation method.

In Sec. 4, we discuss whether such complexity is necessary for task performance. We show that a substantial portion stems from architectural choices—such as activation functions—rather than from intrinsic task complexity.

Finally, in Sec. 5, we outline our view on XAI's trajectory, arguing that the choice between post- and ante-hoc depends on one's stance on the fundamental questions. Regardless, spectral analysis remains a promising direction. We also highlight concrete challenges in each approach.

Contributions This work (1) offers a principled basis for choosing between post- and ante-hoc approaches via two fundamental questions, (2) introduces spectral analysis as a promising framework with initial results, and (3) highlights theoretical gaps and directions for future research.

2. The Emergence of Implicit Surrogates

As neural networks grow increasingly complex, the need for explanations of their outputs become evident to both researchers and practitioners. However, there is no consensus on what constitutes an explanation, leaving room for exploratory and heuristic approaches—an advantageous flexibility in the early stages.

Possibly inspired by interpretability in linear models, early efforts [21] focused on input gradients $\nabla_x f$ of a trained model f, also known as VanillaGrad. However, the input gradient often appears noisy and difficult to interpret due to high frequency information [10] (see Fig. 2b). Through empirical exploration, it has been observed that aggregating explanations over perturbed inputs reduces the complexity of explanations [4, 22] (as in Fig. 2d).

This insight give rise to a broad class of perturbation-based methods, later extended beyond inputs to model parameters and architectures, encouraging diverse experimental post-hoc designs [3, 5, 7, 12, 13, 19, 20, 23, 24].

A formal spectral view to explanations by [16] unifies many post-hoc perturbation-based methods casting them as implicit low-pass filters, attenuating high-frequency com-

ponents of the decision function (see Fig. 1 and the 1D analogy in the second rows of Figs. 2b and 2d). We denote this technique as the convolution of the network's gradient field, $\nabla_x f$, with the perturbation distribution p, $\nabla_x f * p = \nabla_x (f * p) = \nabla_x \widetilde{f}$, acknowledging that it excludes other forms of perturbation, yet remains conceptually aligned with our perspective.

Although low-pass filtering is a well-established concept in signal processing, its significance as an underlying principle of post-hoc explainability has not been explicitly recognized. Crucially, perturbation-based explanations have been shown to implicitly construct surrogate models via this filtering process. The resulting implicit surrogate belongs to a restricted hypothesis space, which forms a subset of the original hypothesis space $\mathcal{H}_{\text{ReLU}}$:

$$\mathcal{H}_{\text{ReLU}*p} = \{f * p : f \in \mathcal{H}_{\text{ReLU}}\} \subseteq \mathcal{H}_{\text{ReLU}}. \tag{1}$$

Understanding the properties of these implicit surrogates present several challenges. Indeed, their faithfulness and inference-time performance, are typically unknown a priori, and are sensitive to heuristically chosen hyperparameters [1, 14]. Since the surrogate's attributions differ from those of the original model, a trade-off arises between faithfulness and complexity. In practice, post-hoc methods generally reduce explanation complexity at the cost of faithfulness.

Recent work suggests that there are more principled alternatives for managing this trade-off [6]. We examine such alternatives in the following section from our spectral view.

3. Is Explanation Complexity Intrinsic?

Some researchers implicitly assume that attribution complexity stems from the attribution method itself, and may not originate from the model complexity.

This viewpoint motivates the development of diverse explanation methods—effectively different forms of low-pass filtering—to reduce perceived complexity. However, this raises a fundamental question (**Q1**):

Is the complexity of the explanation intrinsic to the model, or is it an artifact of the explanation method?

In this section, we address this question by focusing on the image modality. We argue that explanation complexity is not solely an artifact of the explanation method, but instead a *fundamental property rooted in the spectral characteristics of the neural network.*

While it is commonly believed that more complex models yield more complex explanations, [17] approaches this issue from a signal processing perspective. They interpret the visual "noisiness" of attributions as a manifestation of high-frequency components in their spatial power spectrum.

Their analysis reveals a direct connection between the spectral properties of the model and that of the attribution, which we iterate in the following informal statement:

Theorem 1 (Informal)**.** *For models trained on spatially correlated data (*e.g.*, images), the tail behavior of the model function's spatial power spectrum governs the tail behavior of the gradient's spatial power spectrum.*

Refer to Appendix A for technical details.

Intuitively, this result implies that if a model relies heavily on high-frequency features for prediction, then any gradient-based explanation will inherently exhibit high spatial frequency—*i.e.*, complexity.

In simpler terms, models that exhibit sharp variations in their input-output mapping inevitably produce complex explanations, irrespective of the explanation method employed. Thus, the complexity intrinsic to the model naturally propagates to the explanation. Consequently, this complexity must be addressed either during model design, *i.e.* ante-hoc, or post hoc through low-pass filtering techniques (this is visualized in Fig. 1 as two alternatives for (b)).

This theorem establishes a critical link: understanding explanation complexity requires understanding the complexity of the model itself. The complexity arises from the model's decision function, not from the attribution method.

Revisiting assumptions in the post-hoc approach, it becomes evident that altering attribution methods to mask model complexity is often a last resort. A more principled approach is to reduce complexity at the modeling phase.

Next, we address the second key question: can intrinsic model complexity be mitigated, or are post-hoc methods the only viable option?

4. Is All Model Complexity Necessary?

In the previous section, a spectral view allowed us to formalize the extent to which model complexity influences explanation complexity. We established—both intuitively and theoretically—that explanation complexity arises from the model itself and not the explanation method.

While some researchers rely on the assumption that such complexity is an artifact of the explanation technique to justify post-hoc methods, another common belief is that, while explanation complexity originates from the model, the model complexity is necessary for performance. This leads us to the second fundamental question (**Q2**):

Is all of the model's complexity necessary for the task?

As we show in Theorem 1 one of the contributors to the complexity is the use of ReLU activations. The non-differentiable point at zero introduces piecewise-linear behavior, resulting in sharp transitions in the network's decision function and consequently high-frequency components in gradient-based attributions.

To evaluate whether this sharpness—and thus the resulting explanation complexity—is necessary, we consider a

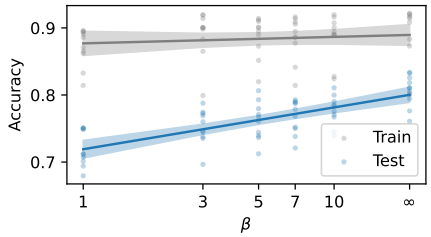


Figure 3. **Effect of Smoothness on Generalization Performance.** The figure displays experimental results obtained on the Imagenette dataset (reproduced the results from [17]). Neural networks employing smoothed variants of the ReLU activation function were trained with different smoothness levels, controlled by the parameter β (x-axis). The limit $\beta \to \infty$ recovers the conventional ReLU. The results indicate that constraining the network's capacity to represent high-frequency components leads to only a minor decline in validation accuracy.

smooth parameterization of ReLU, the SoftPlus function (SP), defined as

$$\mathrm{SP}_\beta(x) = \frac{1}{\beta}\log(1 + e^{\beta x}),$$

where $\beta > 0$ controls the sharpness. As $\beta \to \infty$, SP converges to ReLU; as β decreases, it becomes smoother.

This interpolation enables a direct investigation into the effect of architectural sharpness on explanation complexity. Empirical results (compare Fig. 2f and Fig. 2b) demonstrate that sharper activations lead to increased explanation complexity, confirming that such complexity is inherently tied to model design.

More formally, for positive real numbers $q < r$, one expects a nesting of hypothesis spaces of the form:

$$\mathcal{H}_{\mathrm{SP}(\beta=q)} \subset \mathcal{H}_{\mathrm{SP}(\beta=r)} \subset \mathcal{H}_{\mathrm{SP}(\beta=\infty)} = \mathcal{H}_{\mathrm{ReLU}} \qquad (2)$$

Crucially, this reduction in complexity can be achieved with minimal degradation in performance (see Fig. 3). While post-hoc methods may also reduce explanation complexity, they operate on implicit surrogates whose properties remain largely uncharacterized. In contrast, architectural modifications provide direct, interpretable control over the source of complexity.

These findings suggest that some model complexity present in modern networks is a legacy of early developments in deep learning, prioritizing expressivity and capability over interpretability. Revisiting such design choices with explainability in mind reveals new opportunities to mitigate complexity at its source, and relying on post-hoc remedies as a last resort.

5. Future Directions

The challenges one faces in XAI depends on their answers to the fundamental questions regarding the *source and necessity of complexity*. Explicitly stating researchers' stance on (**Q1** and **Q2**) prevents method–goal mismatch and clarifies the trade-off between complexity and simplicity.

A call for theory Irrespective of the ante-hoc *vs*. post-hoc dilemma, XAI currently lacks mathematical frameworks, and formal grounding, which we believe spectral analysis can be promising in addressing this need.

5.1. Ante-hoc Research Directions

- **Function-space restrictions** Develop principled regularizers and architectures that bound high-frequency behavior while preserving task-relevant expressivity.
- **Optimization reachability** Deviations from conventional designs, often fall outside the scope of well-established theoretical analyses, leaving convergence in the restricted hypothesis space an open question.
- **Inductive-bias formalization** Connect architectural biases (*e.g*. convolutional weight sharing, activation smoothness, batch norm, skip connections) to spectral properties of the learned function and, by extension, to explanation complexity.

5.2. Post-hoc Research Directions

- **Design explicit low-pass filters** As post-hoc approaches share the low-pass filtering mechanism, we may replace implicit surrogate construction with explicit, parameter-controlled low-pass filters whose faithfulness–complexity trade-off can be quantified *a priori*. This leads to techniques similar to pruning or schemes that "low-pass" the network itself, yielding a model with reduced explanation complexity.
- **Low-pass filtering in Other Domains** The spectral view and explicit low-pass filtering allows us to extend post-hoc explainability framework to other domains where Fourier analysis is defined such as graphs and groups.
- **Benchmark design** Borrow evaluation techniques from classical signal processing to create benchmarks that measure both frequency attenuation and faithfulness.

Summary The practitioner's stance on **Q1** and **Q2** determines their route in ante-hoc/post-hoc dilemma. Formalism is the most important need of the current XAI research. One approach to such formalism is the spectral view of networks and explanation methods.

To highlight the promise in this view, we presented a unifying view to ante-hoc and post-hoc methods, that is, restriction of hypothesis space before or after training, respectively. We identified the source of complexity in attributions, *i.e*. the network, and suggested an ante-hoc approach for restriction of hypothesis space. We presented previous works adopting this view that unify post-hoc approaches under one mechanism, *i.e*. low-pass filtering.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps, 2020. arXiv:1810.03292 [cs, stat]. 3

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 2019. arXiv:1910.10045 [cs]. 1

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015. 2

[4] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations, 2020. arXiv:2005.00631 [cs]. 2

[5] Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. NoiseGrad: Enhancing Explanations by Introducing Stochasticity to Model Weights, 2022. arXiv:2106.10185 [cs]. 2

[6] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos Networks: Alignment is All We Need for Interpretability, 2022. 3

[7] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise Explanations of Neural Networks using Adversarial Training, 2020. arXiv:1810.06583 [cs]. 2

[8] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by Removing: A Unified Framework for Model Explanation, 2022. arXiv:2011.14878 [cs, stat]. 1

[9] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the Similarity between the Laplace and Neural Tangent Kernels, 2020. arXiv:2007.01580 [cs]. 6

[10] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks, 2019. arXiv:1806.10758 [cs, stat]. 2

[11] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks, 2020. arXiv:1806.07572. 6

[12] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided Integrated Gradients: An Adaptive Path Method for Removing Noise, 2021. arXiv:2106.09788 [cs]. 2

[13] Beomsu Kim, Junghoon Seo, SeungHyun Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps, 2019. arXiv:1902.04893 [cs, stat]. 2

[14] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of saliency methods, 2017. arXiv:1711.00867 [cs, stat]. 3

[15] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:59–67, 2019. 1

[16] Amir Mehrpanah, Erik Englesson, and Hossein Azizpour. On Spectral Properties of Gradient-Based Explanation Methods. In *Computer Vision – ECCV 2024*, pages 282–299, Cham, 2025. Springer Nature Switzerland. 2

[17] Amir Mehrpanah, Matteo Gamba, Kevin Smith, and Hossein Azizpour. On the Complexity-Faithfulness Trade-off of Gradient-Based Explanations, 2025. arXiv:2508.10490 [cs]. 3, 4, 6

[18] Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the Spectrum of the NTK via a Power Series Expansion, 2023. arXiv:2211.07844 [cs]. 2

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. arXiv:1610.02391 [cs]. 2

[20] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, 2017. arXiv:1605.01713 [cs]. 2

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014. arXiv:1312.6034 [cs]. 2

[22] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, 2017. arXiv:1706.03825 [cs, stat]. 2

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, 2015. arXiv:1412.6806 [cs]. 2

[24] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, 2017. arXiv:1703.01365 [cs]. 2

[25] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based Feature Attribution in Explainable AI: A Technical Review, 2024. arXiv:2403.10415 [cs]. 1

A. Proofs and Technical Background

This section summarizes key background and proof elements adapted from [17], combining results from kernel theory, NTK analysis, and spectral decay properties relevant to gradient-based explanations.

A.1. Kernel Methods and RKHS

A kernel is a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K_{ij} = k(x_i, x_j)$ is positive semidefinite for any finite $\mathcal{X} = \{x_1, \ldots, x_n\}$. Common examples include the Laplace kernel $k(x, x') = \exp(-\|x - x'\|)$ and the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2)$. Each kernel k defines a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ in which smoothness is dictated by k; for instance, $\mathcal{H}_{\text{Gaussian}} \subset \mathcal{H}_{\text{Laplace}}$.

For shift-invariant kernels $k(\Delta)$, $\Delta = \|x - x'\|$, the RKHS admits a Fourier characterization:

$$\mathcal{H}_k = \left\{ f : \int \frac{|\mathcal{F}\{f\}(\omega)|^2}{\mathcal{F}\{k\}(\omega)} \, d\omega < \infty \right\}.$$

Thus, the allowable sharpness of f is bounded by the spectral decay of k.

A.2. Neural Tangent Kernels and Laplace Equivalence

The Neural Tangent Kernel (NTK) [11] describes similarity in terms of network weight gradients:

$$\hat{k}_\ell(x, z) = \left\langle \frac{\partial f(x)}{\partial W^{(\ell)}}, \frac{\partial f(z)}{\partial W^{(\ell)}} \right\rangle,$$

which is related to the pre-activation tangent kernel (PTK) $\mathcal{K}^{(\ell)}$ by

$$\hat{k}_\ell(x, z) = \mathcal{K}^{(\ell)}(x, z) \cdot x_\ell^\top z_\ell.$$

Empirical and theoretical evidence [9] shows that NTKs often closely resemble Laplace kernels in the spectral tail, allowing Laplace kernels to serve as tractable surrogates for analysis.

A.3. Spectral Tail Bound for Input Gradients

Let $f(x) = \sum_{i \in \mathcal{I}} \alpha_i k(x, x_i)$ and $\nabla f(x) = \sum_{i \in \mathcal{I}} \alpha_i \nabla k(x, x_i)$. For shift-invariant k, the Fourier spectrum of ∇f satisfies:

$$|\mathcal{F}\{\nabla f\}(\omega)|^2 \in \mathcal{O}\big(n\,\omega^2 |\hat{k}(\omega)|^2\big).$$

Under fixed dataset size, high spatial autocorrelation, and at least one intersection between training and explanation trajectories, a local linearization shows that the Fourier decay of the explanation trajectory derivative $x_e'(\tau)$ satisfies

$$|\mathcal{F}_\tau\{x_e'(\tau)\}|^2 \in \mathcal{O}\big(\omega^2 |\hat{k}(\omega)|^2\big).$$

Thus, the NTK tail decay directly determines the sharpness of input-gradient–based explanations.

A.4. Effect of Activation Smoothing on NTK Sharpness

For an activation ϕ, define its smoothed form $\phi_\beta = \phi * g_\beta$ where g_β is a Gaussian of precision β. In the NTK τ-transform framework, this smoothing modifies the kernel covariance, yielding faster spectral decay as β increases. The $K^{(1)}$ term in NTK is similarly smoothed. In practice, ϕ_β is approximated by a SoftPlus with parameter proportional to β, interpolating between smooth (large β) and sharp (ReLU) kernels.

Overall, these results link kernel smoothness, NTK sharpness, and the spectral decay of explanation gradients, forming the theoretical backbone for the spectral perspective presented in this work.