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.