A Cohomological Global-to-Local Inference Framework

via Spectral Embedding and Harmonic Gluing

Kazuo Fukiya

Abstract

We propose a global-to-local inference framework for sequence representations that combines spectral analysis on token embeddings with a cohomological regularization principle over overlapping local windows. Starting from discrete Fourier analysis along the token axis, we interpret low-frequency components as carriers of global semantic intent and high-frequency components as local variations. We introduce a graph-based cochain complex over window covers, define a coboundary-induced inconsistency loss, and employ a Hodge-theoretic harmonic projection to extract maximally consistent global representations. The resulting objective integrates task loss, spectral entropy preservation, and cohomological consistency. We provide explicit derivations, optimization schemes, and a concrete integration into Transformer architectures. This work positions cohomology not as a purely topological invariant, but as a computationally tractable regularizer for semantic consistency in deep sequence models.

---

1. Introduction

Large language models rely on locally computed representations whose global semantic consistency is only implicitly enforced. While attention mechanisms provide soft global interactions, they do not explicitly penalize semantic inconsistencies arising across overlapping local contexts. In parallel, recent work has shown that spectral analysis of token embeddings can separate global intent from local fluctuations.

This paper unifies these perspectives. We combine (i) spectral decomposition of token sequences, used to extract a low-frequency global intent, with (ii) a cohomology-inspired framework over local windows, used to regularize inconsistencies among local inferences. Our central contribution is a *harmonic gluing principle*: among all local representations consistent with task loss, we select those

minimizing coboundary energy, and project onto the harmonic subspace of a graph Laplacian induced by the window cover.

Importantly, the proposed framework is not a strict implementation of algebraic topology on sheaves, but a linearized and computationally feasible analogue tailored to neural representations.

---

2. Spectral Formulation on Token Sequences

2.1 Notation and Discrete Fourier Transform

Let a token sequence of length T be represented by an embedding matrix [V ^{d T},] where the t-th column $v_t \in \mathbb{R}^d$ is the embedding of token t.

Let $F \in \mathbb{C}^{T\times T}$ denote the unitary DFT matrix acting along the token axis. The spectral representation is [V = V F,] where the n-th column $\tilde{v}_n$ corresponds to frequency n.

The spectral energy at frequency n is defined as [E_n = | v_n |_2^2,] and the normalized spectral distribution [p_n = .]

2.2 Spectral Entropy and Band Selection

We define a spectral entropy functional [H(p) = - *{n=1}^T p_n p_n.] Low-frequency components are assumed to encode global semantic structure, while high-frequency components capture local variations and noise. Given a frequency band $\Omega \subset \{1, \dots, T\}$, we define the truncated distribution [p_n^{()} = .] Band selection is performed by minimizing the information loss [D{KL}(p | p^{()}),] or equivalently by retaining the minimal Ω such that [*{n } E_n* {n=1}^T E_n,] for a fixed energy ratio $\rho \in (0,1)$.

2.3 Global Intent Reconstruction

The global intent embedding is reconstructed via inverse DFT restricted to Ω: [G = V_{[:,]} F^{-1}_{[,:]}.] Depending on the application, G may be interpreted as a position-dependent low-frequency reconstruction, or further averaged to obtain a position-invariant global intent vector.

---

3. Local Windows and Cochain Structure

3.1 Window Cover and Local Sections

Let $\mathcal{U} = \{U_i\}_{i=1}^{N}$ be a collection of overlapping windows, where each $U_i \subset \{1, \dots, T\}$ is a contiguous interval. For each window, a local inference module produces a representation [s_i ^r.] The collection $s = (s_1, \dots, s_N)$ defines a 0-cochain over the cover.

3.2 Graph Structure and Coboundary Operator

We define an undirected graph $\mathcal{G}$ whose nodes correspond to windows U_i, with an edge (i, j) whenever $U_i \cap U_j \neq \varnothing$. Assigning an arbitrary orientation, we define the incidence matrix $B \in \mathbb{R}^{|E| \times N}$.

The coboundary operator acting on 0-cochains is [s = (B I_r) s,] which assigns to each edge the difference of local sections.

3.3 Inconsistency Energy and Laplacian

The total inconsistency energy is [| s |_2^2 = s^(L I_r) s,] where $L = B^{\top} B$ is the graph Laplacian. This term penalizes semantic disagreement across overlapping windows.

---

4. Hodge-Theoretic Interpretation

The Laplacian admits an eigendecomposition [L = U U^.] The nullspace $\ker L$ corresponds to harmonic 0-cochains, i.e., assignments s that are globally consistent across the cover.

The harmonic projection is given by [P_{} = U_0 U_0^,] where U_0 spans $\ker L$. Applying this projection yields the maximally consistent component of local representations.

We emphasize that nontrivial cohomological obstructions (corresponding to higher-degree cohomology) are not eliminated by this projection; they manifest as irreducible inconsistencies that must be reduced through learning rather than projection.

---

5. Unified Optimization Objective

The full objective integrates task performance, spectral preservation, and cohomological consistency: [L = *i L*{}(s_i) + , s^(L I_r) s + , D_{KL}(p | p^{()}) + , _i | s_i - P_i(G) |_2^2.]

Here P_i extracts the restriction of the global intent G to window U_i.

---

6. Optimization and Linearized Update

Assuming a second-order approximation of local losses, [L_{}(s_i) s_i^H_i s_i - b_i^s_i,] we obtain a global linear system [(H + L I_r + I) s = b + P(G),] which can be solved efficiently using conjugate gradients due to sparsity.

Optionally, after each update, s may be projected onto $\ker L$ to extract the harmonic component.

---

7. Integration into Transformer Architectures

7.1 Spectral Module

At each layer and head, short token windows are transformed via FFT, low-frequency components are reconstructed, and the resulting global intent is injected as a residual: [h_t^{()} h_t^{()} + g_t^{()}.]

7.2 Cohomology Regularizer

Intermediate representations extracted per window are regularized by the Laplacian loss, whose gradients backpropagate through attention and feed-forward blocks.

---

8. Evaluation Protocol

We propose the following metrics: - Perplexity or task loss. - Local consistency score based on cosine similarity over window intersections. - Spectral entropy change ΔH.

Ablations remove spectral, cohomological, or global-prior terms to isolate their contributions.

---

9. Limitations and Scope

The framework relies on linearized consistency and frequency-based assumptions. True semantic contradictions correspond to nontrivial cohomological obstructions that cannot be removed by harmonic projection alone. Moreover, the DFT basis is order-dependent; extensions to graph-based spectral bases are a natural direction for future work.

---

10. Conclusion

We presented a mathematically explicit framework that combines spectral global intent extraction with cohomology-inspired local consistency regularization. By framing semantic agreement as a harmonic condition on a window graph, we obtain a principled and computationally feasible method for global-to-local inference in sequence models. This work establishes a bridge between spectral methods, Hodge theory, and modern neural architectures, offering both theoretical insight and practical implementation pathways.