A Markov Categorical Framework for Language Modeling

Yifan Zhang

Abstract

Autoregressive language models factorize sequence probabilities as $P_\theta(\mathbf{w}) = \prod_t P_\theta(w_t \mid \mathbf{w}_{<t})$ and are trained by minimizing negative log-likelihood (NLL). We present a divergence-enriched Markov Categorical framework that models the single-step map $\mathbf{w}_{<t} \mapsto P_\theta(\cdot \mid \mathbf{w}_{<t})$ as a composite kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in `Stoch`, enabling diagrammatic reasoning about information flow and geometry. Conceptually, we (i) identify an *information surplus* relevant to multi-token/speculative decoding and prove it equals the conditional mutual information $I(H_t; W_{t+1:t+K-1} \mid W_t)$, sharpening DPI-based arguments; (ii) show that for $D = D_{\text{KL}}$ the categorical entropy of the language model (LM) head equals the Shannon conditional entropy $H(W_t \mid H_t)$, clarifying the notion of "intrinsic stochasticity"; and (iii) formalize conditions under which NLL acts as an implicit spectral objective by aligning representation geometry with a predictive similarity operator defined on the representation space via disintegration. These results consolidate standard facts into a compositional calculus that yields clean identities and operator-theoretic statements about autoregressive language models.

`Date:` Revised V2: August 18, 2025

`Project Page:` https://github.com/asiresearch/lm-theory

1 Introduction

Autoregressive language models (AR LMs), particularly those based on the Transformer architecture (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020), have achieved remarkable success, defining the state-of-the-art in natural language generation and demonstrating impressive few-shot learning capabilities. These models operate by sequentially predicting the next token in a sequence based on the preceding context. Formally, given a sequence $\mathbf{w} = w_1 \dots w_L$ with tokens w_i from a finite vocabulary $\mathbb{V}$, the model learns a parameterized probability distribution P_θ that factorizes as:

$$P_\theta(\mathbf{w}) = \prod_{t=1}^{L} P_\theta(w_t|\mathbf{w}_{<t}), \tag{1.1}$$

where $\mathbf{w}_{<t} \coloneqq w_1 \dots w_{t-1}$ is the context sequence, and θ denotes the model parameters, typically optimized by minimizing the negative log-likelihood (NLL) on vast text corpora. The core computational step is the mapping from a context $\mathbf{w}_{<t}$ to the conditional probability distribution $P_\theta(\cdot|\mathbf{w}_{<t})$ over $\mathbb{V}$ for the next token w_t.

Despite their empirical triumphs, a deep theoretical understanding of their internal mechanisms remains incomplete (Manning et al., 2020; Elhage et al., 2021; Yuan, 2023). Current analysis

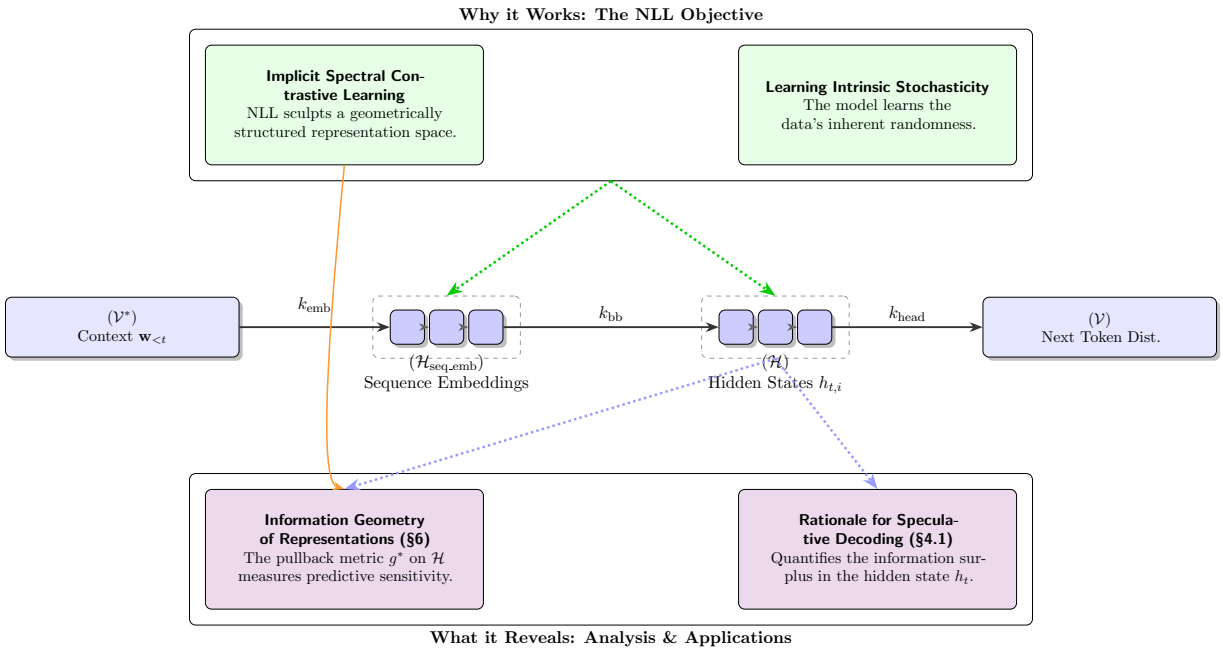

Figure 1 A conceptual overview of our framework. **Center:** The core thesis models the Autoregressive generation step as a composition of Markov kernels $k_{\text{gen}} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category `Stoch`. This separates the deterministic context encoding $(k_{\text{emb}}, k_{\text{bb}})$ from the stochastic prediction (k_{head}). **Top:** This compositional lens reveals the deeper mechanisms of the NLL objective, which we re-frame as minimizing the average KL divergence between the model and true data kernels. We prove that this single objective implicitly forces the model to learn two key properties of the data: its intrinsic conditional stochasticity (measured via categorical entropy) and its underlying geometric structure, which we show is equivalent to performing spectral contrastive learning on a predictive similarity graph. **Bottom:** The framework yields new analytical tools. By pulling back the Fisher-Rao metric, we endow the representation space $\mathcal{H}$ with an information geometry that quantifies predictive sensitivity. This provides a formal basis for analyzing the information surplus in hidden states, giving a rigorous theoretical foundation for modern speculative decoding methods.

often relies on empirical probes (Hewitt and Manning, 2019) or studies of specific components like attention heads (Olsson et al., 2022). While insightful, these methods can be fragmented and often lack a unified mathematical language to describe the model's compositional and stochastic nature as a whole. A fundamental open question is why the simple NLL objective is so effective, leading to representations that capture complex linguistic and world knowledge. Another critical challenge is improving the slow, sequential nature of AR generation. Recent advances in speculative decoding, such as EAGLE (Li et al., 2024), have achieved significant speedups by predicting multiple tokens in parallel, suggesting that the final hidden state h_t contains far more information than is needed for predicting only the single next token w_t. However, a formal understanding of this information surplus is lacking.

This paper addresses this gap by introducing a unifying analytical framework for AR LMs. Our central thesis is that the language of *Markov Categories (MCs)* (Cho and Jacobs, 2019; Fritz, 2020) provides the natural mathematical setting to provide a single, unified mathematical language to formally connect several concepts that are usually discussed separately: information flow through the model's components, the geometry of the learned representation space, and the structural effects

of the NLL training objective.

While many individual mathematical tools we employ—such as the pullback of the Fisher-Rao metric or the connection between NLL and KL divergence—are well-established, our primary contribution is their novel synthesis and application to dissect AR LMs. The originality of this work lies in using the category `Stoch` to formally model compositional information flow, leading to new insights uniquely enabled by this perspective. Unlike information-theoretic analyses that treat models as monolithic black boxes analyzing external behavior (e.g., the entropy of the output sequence), our framework uses categorical information theory to analyze the internal transformations and the learned geometry of the representation space at each stage of processing.

This paper introduces an analytical framework focused on the internal mechanics of the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$. We leverage the category `Stoch`, a canonical MC whose objects are standard Borel spaces (like the continuous representation space $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$) and whose morphisms are Markov kernels (Kallenberg and Kallenberg, 1997; Fritz, 2020). We provide the first formal model of the AR generation step as a composite kernel in `Stoch`:

$$k_{\text{gen},\theta} \coloneqq k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V})). \tag{1.2}$$

Here, k_{emb} and k_{bb} represent the typically deterministic context embedding and backbone transformations that produce the final hidden state $h_t \in \mathcal{H}$, while k_{head} is the generally stochastic kernel mapping h_t to the predictive distribution $P_\theta(\cdot|\mathbf{w}_{<t})$.

A crucial aspect of our framework is enriching `Stoch` with a statistical divergence D (e.g., D_{KL}) (Baez et al., 2016; Perrone, 2023a,b). This allows for defining intrinsic, categorical information measures like entropy $\mathcal{H}_D$ and mutual information I_D (Perrone, 2023a), which automatically satisfy the Data Processing Inequality (DPI). Leveraging this unified framework, this paper makes the following contributions:

1. **A Formal Compositional Model for Information Flow (Section 3):** We formally model the AR generation step as a composite kernel, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. This compositional structure is a powerful tool for reasoning about how information is transformed, preserved, or lost at each distinct stage of processing.

2. **Information-Theoretic Rationale for Speculative Decoding (Section 4.1):** We leverage the DPI to provide a formal, quantitative explanation for the success of methods like EAGLE (Li et al., 2024). We show that the information surplus these methods exploit—the information in a hidden state H_t about multiple future tokens—can be rigorously quantified within our framework.

3. **A Unified View of the NLL Objective (Sections 5 and 7):** We show how the MC framework unifies three critical interpretations of NLL training under one theoretical roof:

   - **Compression:** NLL as KL minimization, equivalent to optimal source coding.
   - **Learning Stochasticity:** NLL forces the model to learn the data's inherent randomness, a process we formalize using categorical entropy. We show that optimizing NLL implies that the model's learned stochasticity converges to that of the data (Theorem 5.2).
   - **Implicit Spectral Contrastive Learning:** We prove that NLL acts as an implicit spectral contrastive objective. By analyzing the information geometry of the prediction head, we

demonstrate that NLL forces the model to learn a geometrically structured representation space by aligning representations with the eigenspectrum of a predictive similarity operator, effectively separating representations of predictively dissimilar contexts (Theorem 7.6).

The primary goal of this paper is to establish the theoretical framework itself, which is a substantial contribution that provides the essential groundwork for future empirical investigation. While estimating some of our proposed information-theoretic quantities in high dimensions is a known challenge (discussed in Section 4), the framework's immediate value lies in providing a new, more powerful lens for theoretical analysis. By formalizing how information is transformed (Section 4), how predictive sensitivity is encoded in representation geometry (Section 6), and how the NLL objective implicitly structures representations (Section 7), we can move towards more principled approaches to model design, interpretation, and control.

The paper is organized as follows. Sections 2 and 3 introduce the MC framework and our compositional model of LMs. Section 4 uses the framework to analyze information flow, providing a rationale for speculative decoding. Section 5 connects the NLL objective to learning the data's intrinsic stochasticity. Sections 6 and 7 present our main theoretical result, showing how NLL performs implicit spectral learning by shaping the geometry of the representation space. Sections 8 and 9 discuss related work and conclude.

2 Background

This section reviews the essential mathematical concepts forming the foundation of our framework: the definition of Markov Categories and the specific category `Stoch`, followed by the enrichment of `Stoch` with statistical divergences leading to categorical information measures.

2.1 Markov Categories and `Stoch`

Markov Categories provide an axiomatic framework for probability and stochastic processes using category theory (Fritz, 2020).

Definition 2.1 (Markov Category (Fritz, 2020))**.** A Markov category $(\mathcal{C}, \otimes, I)$ is a symmetric monoidal category where each object X is equipped with a commutative comonoid structure $(\Delta_X : X \to X \otimes X, !_X : X \to I)$ that is natural in X, and the monoidal unit I is a terminal object (the *causality* axiom: $!_X$ is the unique map $X \to I$).

Morphisms $k : X \to Y$ are interpreted as stochastic processes. Composition $h \circ k$ is sequential processing, while $k \otimes h$ is parallel processing. The comonoid maps Δ_X (copy) and $!_X$ (discard) abstractly model the duplication and deletion of information. The causality axiom enforces that discarding information is deterministic and ultimately reflects probability normalization ($\int k(x, \mathrm{d}y) = 1$) in concrete examples like `Stoch`. States (probability distributions) on an object X are represented as morphisms $p : I \to X$.

The key example for our purposes is the category `Stoch`.

Definition 2.2 (Category `Stoch` (Fritz, 2020; Perrone, 2023a))**.** The Markov category `Stoch` is defined by:

- **Objects**: Standard Borel spaces $(X, \mathscr{B}(X))$. These are general measure spaces that include finite sets (like a vocabulary $\mathbb{V}$), countable sets, and continuous spaces like Euclidean space $\mathbb{R}^d$ or other Polish spaces. This ensures the framework can handle both discrete tokens and continuous representations. The monoidal unit I is a singleton space $(\{\star\}, \{\emptyset, \{\star\}\})$.
- **Morphisms**: Markov kernels $k : X \to Y$. A map $k : X \times \mathscr{B}(Y) \to [0, 1]$ where $k(x, \cdot)$ is a probability measure on Y for each $x \in X$, and $k(\cdot, A)$ is a measurable function on X for each $A \in \mathscr{B}(Y)$.
- **Composition**: Given $k : X \to Y$ and $h : Y \to Z$, the composite $h \circ k : X \to Z$ is $(h \circ k)(x, C) \coloneqq \int_Y h(y, C)\, k(x, \mathrm{d}y)$ (Chapman-Kolmogorov). Identity $\mathrm{id}_X(x, A) = \delta_x(A)$.
- **Monoidal Product ($\otimes$)**: Product space $(X \times Y, \mathscr{B}(X) \otimes \mathscr{B}(Y))$ with the product σ-algebra. Product kernel $(k \otimes h)((x, y), \cdot) \coloneqq k(x, \cdot) \otimes h(y, \cdot)$ (product measure).
- **Symmetry**: Swap map $\sigma_{X,Y} : X \otimes Y \to Y \otimes X$ is $\sigma_{X,Y}((x, y), \cdot) = \delta_{(y,x)}$.
- **Comonoid Structure**: Copy $\Delta_X : X \to X \otimes X$ is $\Delta_X(x, \cdot) = \delta_{(x,x)}$. Discard $!_X : X \to I$ maps to the unique point measure on I, $!_X(x, \{\star\}) = 1$.
- **Causality**: I is terminal, $!_Y \circ k = !_X$ holds, reflecting probability normalization.

Remark 2.3 (Interpretation)**.** In `Stoch`, objects represent the types of random outcomes (e.g., sequences, vectors, tokens). Morphisms represent stochastic processes or channels mapping inputs to probability distributions over outputs. Deterministic functions $f : X \to Y$ correspond to deterministic kernels $k_f(x, \cdot) = \delta_{f(x)}$. States $p : I \to X$ correspond bijectively to probability measures $\mu_p \in \mathcal{P}(X)$ via $\mu_p(A) = p(\star, A)$. Marginalization arises from discarding information, e.g., for a joint state $p : I \to X \otimes Y$, the X-marginal is $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$.

2.2 Divergence Enrichment and Categorical Information Measures

The structure of `Stoch` is particularly powerful when enriched with a statistical divergence D, quantifying the dissimilarity between probability measures (states) $p, q : I \to X$, written $D_X(p\|q)$ (Perrone, 2023a). Examples include KL divergence (D_{KL}), Total Variation (d_{TV}), Rényi divergences (D_α), and the broad class of f-divergences (D_f) (Amari and Nagaoka, 2000; Nowozin et al., 2016).

A fundamental property linking divergences and Markov kernels is the Data Processing Inequality (DPI), which holds for most standard divergences (e.g., f-divergences, Rényi $\alpha \in [0, \infty]$).

Theorem 2.4 (Data Processing Inequality (DPI))**.** Let D be a statistical divergence satisfying the DPI. For any Markov kernel $k : X \to Y$ in `Stoch` and any pair of states $p, q : I \to X$:

$$D_Y(k \circ p\|k \circ q) \leq D_X(p\|q) \tag{2.1}$$

Processing through the channel k cannot increase the D-divergence between the distributions.

Based on this, Perrone (Perrone, 2023a) introduced categorical definitions of entropy and mutual information intrinsically tied to the divergence D and the MC structure.

Definition 2.5 (Categorical Entropy (Perrone, 2023a))**.** Let $(\mathtt{Stoch}, D)$ be enriched with a DPI-satisfying divergence D.

1. The **Categorical Entropy** of a kernel $k : X \to Y$ measures its intrinsic stochasticity:

$$\mathcal{H}_D(k) \coloneqq D_{Y \otimes Y} \left(\Delta_Y \circ k \quad \| \quad (k \otimes k) \circ \Delta_X \right) \tag{2.2}$$

   Intuitively, this compares two processes: (1) applying k to an input x and then copying the output to get (y, y); versus (2) copying the input x and applying k independently to each copy to get (y_1, y_2). The more stochastic k is, the greater the divergence. If k is deterministic, the entropy is zero.

2. The **Categorical Mutual Information** for a joint state $p : I \to X \otimes Y$ is defined as $I_D(p) \coloneqq D_{X \otimes Y}(p \| p_X \otimes p_Y)$, where p_X and p_Y are the marginal states. It measures the dependence between X and Y.

Remark 2.6 (Properties and Connections)**.** When $D = D_{\mathrm{KL}}$, $I_{D_{\mathrm{KL}}}(p)$ recovers the standard Shannon mutual information $I(X;Y)$. $\mathcal{H}_{D_{\mathrm{KL}}}(k)$ provides an intrinsic measure of the kernel's stochasticity, related to but distinct from average conditional Shannon entropy (Perrone, 2023a). Crucially, these definitions automatically satisfy the DPI. For instance, for a state $p : I \to X \otimes Y$ and a kernel $h : Y \to Z$, the DPI implies $I_D(X;Y) \geq I_D(X;Z)$, reflecting that processing $(Y \to Z)$ cannot create information about X. Information geometry (Amari and Nagaoka, 2000) arises naturally, as the Fisher-Rao metric is induced by the local quadratic approximation of KL divergence.

3 Autoregressive Language Models as Composed Kernels

We now apply the Markov Category framework established in Section 2 to model Autoregressive language models. Specifically, we model the single-step generation mapping $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a composition of Markov kernels within the category `Stoch`.

The relevant measurable spaces (objects in `Stoch`) are:

- Input context space: $(\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) = (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$, where $\mathbb{V}^*$ is the set of finite sequences over the vocabulary $\mathbb{V}$, equipped with a suitable σ-algebra making it standard Borel (e.g., considering it as a disjoint union of finite products $\mathbb{V}^n$).
- Initial sequence representation space: $(\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) = (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$, the space of initial vector sequences (e.g., $\bigcup_n (\mathbb{R}^{d_{\text{model}}})^n$), also equipped with a standard Borel structure.
- Final hidden state space: $(\mathcal{H}, \mathscr{B}(\mathcal{H})) = (\mathcal{H}, \mathscr{B}(\mathcal{H}))$, typically $(\mathbb{R}^{d_{\text{model}}}, \mathscr{B}(\mathbb{R}^{d_{\text{model}}}))$.
- Output vocabulary space: $(\mathbb{V}, \mathcal{P}(\mathbb{V})) = (\mathbb{V}, \mathcal{P}(\mathbb{V}))$, a finite measurable space.

Standard Borel spaces are chosen because they form a well-behaved class of measurable spaces (isomorphic to Borel subsets of Polish spaces) closed under countable products, sums, and containing standard examples like $\mathbb{R}^d$ and finite sets, ensuring measure-theoretic regularity (Kallenberg and Kallenberg, 1997).

The generation process decomposes into three kernels (morphisms in `Stoch`):

1. **Embedding Layer Kernel (**$k_{\text{emb}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$**)**: This kernel encapsulates the initial processing of the discrete input sequence $\mathbf{w}_{<t} \in \mathbb{V}^*$. It typically involves applying a token embedding function $\mathcal{E} : \mathbb{V} \to \mathbb{R}^{d_{\text{model}}}$ to each token w_i and potentially incorporating absolute

positional encodings. Let $f_{\text{emb}} : \mathbb{V}^* \to \mathcal{H}_{\text{seq\_emb}}$ denote the overall deterministic function computing the initial sequence representation $E_{<t}$. Since this mapping is deterministic, the kernel k_{emb} is defined via the Dirac measure $\delta_\cdot$:

$$k_{\text{emb}}(\mathbf{w}_{<t}, A) \coloneqq \delta_{f_{\text{emb}}(\mathbf{w}_{<t})}(A) = \mathbf{1}_A(f_{\text{emb}}(\mathbf{w}_{<t})), \quad \text{for } A \in \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}). \tag{3.1}$$

This is a valid morphism in `Stoch`.

2. **Backbone Transformation Kernel ($k_{\text{bb}} : (\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}})) \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$)**: This kernel represents the core computation, usually a deep neural network like a Transformer stack. Let $f_{\text{bb}} : \mathcal{H}_{\text{seq\_emb}} \to \mathcal{H}$ be the function mapping the initial sequence representation $E_{<t}$ to the final hidden state $h_t \in \mathcal{H}$ (often the output vector at the last sequence position). This function incorporates complex operations like multi-head self-attention and feed-forward layers. Relative positional information, such as Rotary Position Embeddings (RoPE) (Su et al., 2024), is implemented within the function f_{bb} by modifying attention computations based on token positions. Assuming the backbone computation is deterministic for a given $E_{<t}$ and parameters θ, the kernel k_{bb} is also deterministic:

$$k_{\text{bb}}(E_{<t}, B) \coloneqq \delta_{f_{\text{bb}}(E_{<t})}(B) = \mathbf{1}_B(f_{\text{bb}}(E_{<t})), \quad \text{for } B \in \mathscr{B}(\mathcal{H}). \tag{3.2}$$

This is also a morphism in `Stoch`.

3. **LM Head Kernel ($k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$)**: This final kernel maps the summary hidden state $h_t \in \mathcal{H}$ to a probability distribution over the finite vocabulary $\mathbb{V}$. Typically, h_t is passed through a linear layer ($f_{\text{head}} : \mathcal{H} \to \mathbb{R}^{|\mathbb{V}|}$) producing logits $\mathbf{z} = f_{\text{head}}(h_t)$, followed by the softmax function: $P(w|h_t) = [\text{softmax}(\mathbf{z})]_w$. This defines a genuinely stochastic Markov kernel:

$$k_{\text{head}}(h, A) \coloneqq \sum_{w \in A} [\text{softmax}(f_{\text{head}}(h))]_w \quad \text{for } h \in \mathcal{H}, A \subseteq \mathbb{V}. \tag{3.3}$$

This kernel maps each point h in the representation space to a probability measure on the discrete space $\mathbb{V}$, satisfying the required measurability conditions. It is a morphism in `Stoch`.

The overall single-step generation kernel $k_{\text{gen},\theta} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ is the composition $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ in the category `Stoch`. This composition precisely represents the model's learned conditional probability map $P_\theta(\cdot|\mathbf{w}_{<t})$. It is crucial to note that this formalism applies to **any** AR model, including Transformers. The attention mechanism provides a powerful, history-dependent parameterization of this single-step kernel. The subsequent sections use this representation to analyze the model's behavior.

4 Information-Theoretic Analysis via Categorical Metrics

The MC framework allows us to define principled metrics for internal analysis and to formally reason about information flow. We focus on two key applications that are central to the paper's main arguments: quantifying the information surplus exploited by speculative decoding and measuring the intrinsic stochasticity of the prediction head.

We operate within the probabilistic setting induced by a distribution P_{ctx} over input contexts, corresponding to an initial state $p_{W_{<t}} : I \to (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$. Processing this state through the composed kernels induces distributions over the hidden state H_t (state $p_{H_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$) and the next token W_t (state $p_{W_t} : I \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$).

4.1 Information Flow Bounds and Rationale for Speculative Decoding

The Data Processing Inequality (DPI) is native to our framework, but a sharper account of multi-token/speculative decoding arises from the *chain rule* for mutual information. Let $W_{t:t+K-1} = (W_t, \dots, W_{t+K-1})$. Then

$$I(H_t; W_{t:t+K-1}) = I(H_t; W_t) + I\big(H_t; W_{t+1:t+K-1} \mid W_t\big). \tag{4.1}$$

We therefore define the *information surplus* as the conditional term

$$\mathsf{Surplus}_K \coloneqq I\big(H_t; W_{t+1:t+K-1} \mid W_t\big), \tag{4.2}$$

which admits immediate bounds $0 \le \mathsf{Surplus}_K \le H(W_{t+1:t+K-1} \mid W_t) \le (K-1)\log|\mathbb{V}|$. This identifies, *exactly*, the quantity exploited by K-token drafting: information about future tokens in H_t that is not exhausted by W_t alone.

Our framework can model the mechanism of speculative decoding explicitly. These models introduce a second, lightweight "drafting" kernel, $k_{\text{draft}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}^K, \mathcal{P}(\mathbb{V}^K))$, which operates in parallel with the original "verification" head, which we now denote $k_{\text{verify}} \equiv k_{\text{head}}$. Both kernels take the same hidden state H_t as input, creating two parallel information pathways:

$$\begin{array}{ccccc} \mathbb{V}^* & \xrightarrow{k_{\text{enc}} \coloneqq k_{\text{bb}} \circ k_{\text{emb}}} & \mathcal{H} & \xrightarrow{k_{\text{verify}}} & \mathbb{V} \\ & & \Big\downarrow{\scriptstyle k_{\text{draft}}} & & \\ & & \mathbb{V}^K & & \end{array}$$

This compositional diagram makes the structure clear. The framework allows for a principled analysis of these pathways. For instance, one could compare the average categorical entropy (equation (4.4)) of the two heads. It is plausible that an effective drafting head $\bar{\mathcal{H}}_D(k_{\text{draft}})$ would be higher than that of the verification head $\bar{\mathcal{H}}_D(k_{\text{verify}})$, reflecting that its role is to propose a diverse set of candidate sequences, while the verifier's role is to make a high-confidence final selection.

The key insight is that our framework moves beyond a DPI inequality to the identity (4.1), isolating the exact conditional mutual information resource available to parallel proposal mechanisms.

4.2 Metric: LM Head Categorical Entropy (Prediction Stochasticity)

To quantify the intrinsic stochasticity or uncertainty associated with the final prediction step, embodied by the LM head kernel $k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$. This metric is crucial for understanding how NLL training forces the model to learn the data's inherent randomness (Section 5). For a given input $h \in \mathcal{H}$, the categorical entropy quantifies the stochasticity of the output distribution $k_{\text{head}}(h, \cdot)$. We obtain a single summary statistic by averaging this value over the distribution of hidden states p_{H_t}.

Definition 4.1 (Categorical Entropy of k_{head})**.** The Categorical Entropy of k_{head} is defined using equation (2.2) with $X = \mathcal{H}$, $Y = \mathbb{V}$, and $k = k_{\text{head}}$:

$$\mathcal{H}_D(k_{\text{head}}) : \mathcal{H} \to \mathbb{R}_{\ge 0} \coloneqq D_{\mathbb{V}\otimes\mathbb{V}}\left(\Delta_{\mathbb{V}} \circ k_{\text{head}} \quad \| \quad (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}\right). \tag{4.3}$$

The categorical entropy $\mathcal{H}_D(k_{\text{head}})(h)$ measures the divergence between generating a correlated pair (W, W) versus an independent pair (W_1, W_2), where $W, W_1, W_2 \sim k_{\text{head}}(h, \cdot)$. This quantifies how far the output distribution for a given h is from a deterministic point mass. To obtain a single metric for the LM head's overall stochasticity, we compute its expectation with respect to the hidden state distribution p_{H_t}:

$$\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) \coloneqq \mathbb{E}_{h \sim p_{H_t}} \left[D_{\mathbb{V} \otimes \mathbb{V}} \left(\sum_{w \in \mathbb{V}} k_{\text{head}}(h, \{w\}) \delta_{(w,w)} \quad \| \quad k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot) \right) \right]. \quad (4.4)$$

Interpretation. This metric measures the intrinsic conditional stochasticity of the LM head mapping. If k_{head} were deterministic (i.e., for each h, it mapped to a single specific w_h, so $p_h = \delta_{w_h}$), then both measures inside the divergence would be $\delta_{(w_h, w_h)}$, and the entropy would be $D(\delta_{(w_h,w_h)} \| \delta_{(w_h,w_h)}) = 0$. A higher value of $\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t})$ indicates greater average uncertainty or "spread" in the output distribution $p_h = k_{\text{head}}(h, \cdot)$, meaning the kernel is inherently more stochastic. It quantifies how far the prediction process is from a deterministic assignment, measured in the geometry of $\mathbb{V} \otimes \mathbb{V}$ induced by D.

For the specific case $D = D_{\text{KL}}$ and finite $\mathbb{V}$, the categorical entropy *equals* the Shannon conditional entropy:

Proposition 4.2 (KL case reduces to Shannon)**.** For the LM head kernel $k(h, \cdot) = p_h$ over a finite vocabulary, $\mathcal{H}_{D_{\text{KL}}}(k_{\text{head}})(h) = H(p_h), \qquad \bar{\mathcal{H}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t}) = H(W_t \mid H_t)$.

Proof. Since $\sum_w p_h(w) \delta_{(w,w)}$ is supported on the diagonal, a direct calculation gives $D_{\mathbb{V} \otimes \mathbb{V}}\big(\sum_w p_h(w) \delta_{(w,w)} \,\big\|\, p_h \otimes p_h\big) = -\sum_w p_h(w) \log p_h(w)$, and averaging over H_t yields $H(W_t \mid H_t)$. □

5 Pretraining Objective, Compression, and Learning Intrinsic Stochasticity

A central question surrounding large language models is why the seemingly simple Autoregressive objective of next-token prediction, trained via minimizing cross-entropy loss (equivalently, negative log-likelihood or NLL), yields such powerful and versatile capabilities, often exhibiting behaviors associated with understanding and reasoning. The framework of Markov Categories and categorical entropy provides a lens through which to interpret this phenomenon, connecting it to fundamental ideas about compression and learning the inherent stochasticity of the data generating process.

Let $k_{\text{data}} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the (potentially unknown) Markov kernel representing the true data-generating process, such that $k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ corresponds to the true conditional probability measure $P_{\text{data}}(\cdot | \mathbf{w}_{<t})$ on the vocabulary $\mathbb{V}$. Let $p_{W_{<t}}$ denote the marginal probability measure on the context space $(\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*))$, derived from the underlying joint distribution P_{data} over sequences observed in the training corpus.

The standard pretraining objective for an AR LM parameterized by θ is to minimize the negative log-likelihood (NLL) of the next token w_t given the preceding context $\mathbf{w}_{<t}$, averaged over the training data distribution P_{data}. This is equivalent to minimizing the cross-entropy:

$$L_{\text{CE}}(\theta) = -\mathbb{E}_{(\mathbf{w}_{<t}, w_t) \sim P_{\text{data}}}[\log P_\theta(w_t | \mathbf{w}_{<t})] \quad (5.1)$$

where $P_\theta(w_t|\mathbf{w}_{<t})$ is the model's predicted probability. Let $k_{\text{gen},\theta} : (\mathbb{V}^*, \mathscr{B}(\mathbb{V}^*)) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the model's overall generation kernel, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$, such that $P_\theta(\cdot|\mathbf{w}_{<t}) = k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot)$. The objective can be precisely stated in terms of Kullback-Leibler (KL) divergence between the data kernel and the model kernel.

Theorem 5.1 (NLL Minimization as Average KL Minimization)**.** It is a well-known result in information theory that minimizing the cross-entropy loss $L_{\text{CE}}(\theta)$ (equation (5.1)) is equivalent to minimizing the average KL divergence between the true and model conditional distributions. We state it here in the language of our framework to ground the subsequent analysis.

$$\operatorname*{argmin}_\theta L_{\text{CE}}(\theta) = \operatorname*{argmin}_\theta \mathcal{L}_{\text{KL}}(\theta) \coloneqq \operatorname*{argmin}_\theta \mathbb{E}_{\mathbf{w}_{<t}\sim p_{W_{<t}}}[D_{\text{KL}}(k_{\text{data}}(\mathbf{w}_{<t}, \cdot) \,\|\, k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot))] \tag{5.2}$$

where the expectation is taken over contexts $\mathbf{w}_{<t}$ drawn according to the data's marginal context distribution $p_{W_{<t}}$. The minimum value of $\mathcal{L}_{\text{KL}}(\theta)$ is non-negative. If the model class $\{k_{\text{gen},\theta} \mid \theta \in \Theta\}$ is sufficiently expressive to contain k_{data} (i.e., $k_{\text{data}} = k_{\text{gen},\theta_{\text{true}}}$ for some $\theta_{\text{true}} \in \Theta$), then the minimum value is 0, achieved if and only if $k_{\text{gen},\theta^*}(\mathbf{w}_{<t}, \cdot) = k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ for $p_{W_{<t}}$-almost every context $\mathbf{w}_{<t}$.

Proof Sketch. The equivalence arises from rewriting cross-entropy as

$$L_{\text{CE}}(\theta) = \mathcal{L}_{\text{KL}}(\theta) + H(W_t|W_{<t})_{\text{data}},$$

where the conditional entropy $H(W_t|W_{<t})_{\text{data}} = \mathbb{E}_{\mathbf{w}_{<t}}[-\sum_{w_t} P_{\text{data}}(w_t|\mathbf{w}_{<t}) \log P_{\text{data}}(w_t|\mathbf{w}_{<t})]$ is independent of θ. Since $H(W_t|W_{<t})_{\text{data}}$ is constant during optimization, minimizing $L_{\text{CE}}(\theta)$ is identical to minimizing $\mathcal{L}_{\text{KL}}(\theta)$. The non-negativity and condition for achieving zero follow from the fundamental properties of KL divergence. □

This theorem frames NLL training as driving the model kernel $k_{\text{gen},\theta}$ to match the data kernel k_{data}. The connection to compression arises from Shannon's source coding theorem. The minimal average code length required to losslessly encode the next token w_t, given the context $\mathbf{w}_{<t}$ and using an optimal code based on the true distribution $P_{\text{data}}(\cdot|\mathbf{w}_{<t})$, is the conditional Shannon entropy $H(W_t|W_{<t})_{\text{data}}$. The cross-entropy loss $L_{\text{CE}}(\theta)$ achieved by the model represents the average code length when using a code based on the model's distribution $P_\theta(\cdot|\mathbf{w}_{<t})$. Therefore, minimizing NLL (theorem 5.1) is equivalent to finding a model that provides the most efficient compression of the training data sequences, achieving an average code length that approaches the theoretical minimum $H(W_t|W_{<t})_{\text{data}}$. The widely discussed hypothesis that "compression implies understanding" posits that achieving high compression rates on complex data like natural language necessitates learning the underlying structure, rules, and statistical regularities, which may manifest as emergent capabilities.

Beyond matching the predictive distributions point-wise on average, successful NLL training implies that the model also learns to replicate the intrinsic stochasticity or uncertainty inherent in the data generation process at the prediction step. Within our framework, this intrinsic conditional stochasticity can be quantified using the concept of average categorical entropy (equation (4.4)). Recall that for a kernel $k : X \to Y$ and an input distribution p_X, the average categorical entropy with respect to divergence D is:

$$\bar{\mathcal{H}}_D(k; p_X) \coloneqq \mathbb{E}_{x\sim p_X}\left[D_{Y\otimes Y}\left(\sum_{y\in Y} k(x, \{y\})\delta_{(y,y)} \quad \| \quad k(x, \cdot) \otimes k(x, \cdot)\right)\right]. \tag{5.3}$$

This measures the average divergence between deterministically copying the output $y \sim k(x, \cdot)$ versus generating two independent outputs $y_1, y_2 \sim k(x, \cdot)$. It quantifies the average "spread" or non-determinism of the kernel k.

Let $k_{\text{head},\theta} : \mathcal{H} \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ be the LM head kernel corresponding to parameters θ. Let $p_{H_t,\theta}$ be the distribution over hidden states $h_t \in \mathcal{H}$ induced by processing contexts $\mathbf{w}_{<t} \sim p_{W_{<t}}$ through the model's encoder $k_{\text{bb}} \circ k_{\text{emb}}$ (parameterized by θ). The following theorem formalizes the idea that convergence in average KL divergence implies convergence in the learned average intrinsic stochasticity.

Theorem 5.2 (Convergence of Average Categorical Entropy via NLL Minimization)**.** Let D be a divergence on $\mathcal{P}(\mathbb{V} \times \mathbb{V})$ with $\mathbb{V}$ finite, and define $\Psi_D(h, p) \coloneqq D_{\mathbb{V}\otimes\mathbb{V}}\big(\sum_w p(w)\delta_{(w,w)} \,\|\, p \otimes p\big)$. Suppose a sequence θ_n satisfies $\mathcal{L}_{\text{KL}}(\theta_n) \to \inf_\theta \mathcal{L}_{\text{KL}}(\theta)$ and that the *joint* laws of $(H_t^{(n)}, P^{(n)})$, where $H_t^{(n)} \sim p_{H_t,\theta_n}$ and $P^{(n)} = k_{\text{head},\theta_n}(H_t^{(n)}, \cdot)$, converge weakly to $(H_t^\star, P^\star)$ in $\mathcal{H} \times \Delta^{|\mathbb{V}|-1}$. If Ψ_D is bounded and continuous in its second argument, then

$$\lim_{n\to\infty} \mathbb{E}\big[\Psi_D\big(H_t^{(n)}, P^{(n)}\big)\big] = \mathbb{E}\big[\Psi_D\big(H_t^\star, P^\star\big)\big]. \tag{5.4}$$

Furthermore, if $k_{\text{gen},\theta^\star} = k_{\text{data}}$ almost surely under $p_{W_{<t}}$, then the limit equals the corresponding data quantity.

$$\bar{\mathcal{H}}_D(k_{\text{head},\theta^\star}; p_{H_t,\theta^\star}) = \bar{\mathcal{H}}_D(k_{\text{head, data}}; p_{H_t,\text{data}}). \tag{5.5}$$

Proof Sketch. Consider the measurable map $(h, p) \mapsto \Psi_D(h, p)$. By the continuous mapping theorem under joint weak convergence and boundedness/continuity in p, we have $\Psi_D(H_t^{(n)}, P^{(n)}) \Rightarrow \Psi_D(H_t^\star, P^\star)$ and uniform integrability is trivial by boundedness, hence expectations converge. The identification with the data quantity follows from equality of kernels almost surely. Full details in Appendix A.4. □

Theorem 5.2 provides a formal basis for the claim that NLL training compels the model to learn not just the most likely next token, but also the degree of uncertainty or stochasticity associated with that prediction, as dictated by the data. By minimizing the average KL divergence $\mathcal{L}_{\text{KL}}(\theta)$, the model $k_{\text{gen},\theta}$ must align its output distributions $k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot)$ with the data distributions $k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$. This alignment necessarily includes matching the "shape" or "spread" of these distributions, which is precisely what is quantified by the average categorical entropy $\bar{\mathcal{H}}_D$. The parameters θ and the compositional structure $k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ thus become a compressed representation capturing both the predictive dependencies and the inherent conditional randomness of the language source. This suggests that learning the correct level of stochasticity is an integral part of the compression process driven by the NLL objective, contributing to the model's ability to generate realistic and diverse text sequences.

6 Information Geometry of Representation and Prediction Spaces

The Markov Category framework, particularly $(\texttt{Stoch}, D)$ enriched with a divergence like D_{KL}, provides a natural bridge to Information Geometry (Amari and Nagaoka, 2000; Perrone, 2023b). This allows for a geometric analysis of the spaces involved in AR language modeling, particularly the representation space $\mathcal{H}$ and the space of next-token distributions $\mathcal{P}(\mathbb{V})$.

The space $\mathcal{P}(\mathbb{V})$ of probability distributions over the finite vocabulary $\mathbb{V}$ forms a $(|\mathbb{V}|-1)$-dimensional simplex $\Delta^{|\mathbb{V}|-1}$. This space possesses a well-defined Riemannian geometry induced by the Fisher-Rao information metric g^{FR}, whose components in a local coordinate system $\xi = (\xi_1, \ldots, \xi_{|\mathbb{V}|-1})$ for a distribution $p_\xi \in \mathcal{P}(\mathbb{V})$ are given by:

$$g^{\mathrm{FR}}_{ij}(\xi) = \sum_{w \in \mathbb{V}} p_\xi(w) \frac{\partial \log p_\xi(w)}{\partial \xi_i} \frac{\partial \log p_\xi(w)}{\partial \xi_j} = \mathbb{E}_{W \sim p_\xi}\left[\frac{\partial \log p_\xi(W)}{\partial \xi_i} \frac{\partial \log p_\xi(W)}{\partial \xi_j} \right]. \tag{6.1}$$

This metric quantifies the local distinguishability between nearby probability distributions, measuring the distance in terms of expected squared log-likelihood ratio gradients. The geometry of $\mathcal{P}(\mathbb{V})$ also includes dual affine connections ($\pm\alpha$-connections) related to the KL divergence, providing a richer dually flat structure (Amari and Nagaoka, 2000).

The LM Head kernel $k_{\mathrm{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathbb{V}, \mathcal{P}(\mathbb{V}))$ corresponds to a deterministic mapping from a hidden state $h \in \mathcal{H} \cong \mathbb{R}^{d_{\mathrm{model}}}$ to a probability distribution $p_h := k_{\mathrm{head}}(h, \cdot) \in \mathcal{P}(\mathbb{V})$. Let $g_{\mathrm{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ denote this mapping, $p_h = g_{\mathrm{head}}(h)$. Typically, this involves a linear layer followed by softmax: $g_{\mathrm{head}}(h) = \mathrm{softmax}(Wh)$ where $W \in \mathbb{R}^{|\mathbb{V}| \times d_{\mathrm{model}}}$. This mapping g_{head} allows us to pull back the geometric structure from $\mathcal{P}(\mathbb{V})$ onto the representation space $\mathcal{H}$.

Specifically, the Fisher-Rao metric g^{FR} on $\mathcal{P}(\mathbb{V})$ induces a (generally degenerate) Riemannian metric tensor $g^* = g^*_{\mathrm{head}} g^{\mathrm{FR}}$ on $\mathcal{H}$. At a point $h \in \mathcal{H}$, the components of this pullback metric are given by:

$$g^*_{ab}(h) = \sum_{i,j} g^{\mathrm{FR}}_{ij}(g_{\mathrm{head}}(h)) \frac{\partial (g_{\mathrm{head}}(h))_i}{\partial h_a} \frac{\partial (g_{\mathrm{head}}(h))_j}{\partial h_b}, \quad a, b \in \{1, \ldots, d_{\mathrm{model}}\}, \tag{6.2}$$

where h_a, h_b are coordinates of $h \in \mathcal{H}$, and $(g_{\mathrm{head}}(h))_i, (g_{\mathrm{head}}(h))_j$ represent local coordinates of the output distribution $p_h \in \mathcal{P}(\mathbb{V})$ (e.g., probabilities of specific tokens, possibly excluding one due to the sum-to-one constraint). The term $\frac{\partial (g_{\mathrm{head}}(h))_i}{\partial h_a}$ is the Jacobian of the LM head map g_{head} evaluated at h.

Let $J(h)$ denote this Jacobian matrix ($|\mathbb{V}| - 1 \times d_{\mathrm{model}}$ or $|\mathbb{V}| \times d_{\mathrm{model}}$ depending on coordinates). Then $g^*(h) = J(h)^\top g^{\mathrm{FR}}(g_{\mathrm{head}}(h)) J(h)$. The significance of this pullback metric g^* lies in its connection to the local distinguishability of output distributions under perturbations of the input hidden state, as measured by divergences like KL divergence.

Theorem 6.1 (Pullback Metric and Local Divergence)**.** Let $g_{\mathrm{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ be the smooth map corresponding to the LM head kernel. Let $h \in \mathcal{H}$ and $v \in T_h\mathcal{H} \cong \mathcal{H}$. Consider the distributions $p_h = g_{\mathrm{head}}(h)$ and $p_{h+\epsilon v} = g_{\mathrm{head}}(h + \epsilon v)$ for small ϵ. The KL divergence between these output distributions, for small ϵ, is locally approximated by the quadratic form defined by the pullback metric $g^*(h)$:

$$D_{\mathrm{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2} \epsilon^2 g^*(h)(v, v) + O(\epsilon^3) \tag{6.3}$$

where $g^*(h)(v, v) = \sum_{a,b=1}^{d_{\mathrm{model}}} g^*_{ab}(h) v_a v_b$. A similar relationship holds for symmetric KL divergence and, more generally, for any f-divergence D_f where f is sufficiently smooth around 1 with $f''(1) > 0$.

Proof. Let ξ be a local coordinate system for $\mathcal{P}(\mathbb{V})$ around p_h. The KL divergence between two nearby distributions p_ξ and $p_{\xi'}$ can be expanded around p_ξ as (Amari and Nagaoka, 2000):

$$D_{\mathrm{KL}}(p_{\xi'} \,\|\, p_\xi) = \frac{1}{2} \sum_{i,j} g^{\mathrm{FR}}_{ij}(\xi)(\xi'_i - \xi_i)(\xi'_j - \xi_j) + O(\|\xi' - \xi\|^3).$$

Let $\xi(h)$ denote the coordinates of $p_h = g_{\text{head}}(h)$. For $p_{h+\epsilon v}$, the coordinates are $\xi(h+\epsilon v)$. By Taylor expansion in ϵ:

$$\xi_i(h+\epsilon v) = \xi_i(h) + \epsilon \sum_{a=1}^{d_{\text{model}}} \frac{\partial \xi_i}{\partial h_a}(h) v_a + O(\epsilon^2).$$

Thus, $\xi_i(h+\epsilon v) - \xi_i(h) = \epsilon J_{ia}(h) v_a + O(\epsilon^2)$, where $J(h)$ is the Jacobian matrix of the map $h \mapsto \xi(h)$ (i.e., the Jacobian of g_{head} in local coordinates ξ). Substituting this into the KL expansion:

$$\begin{aligned} D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) &= \frac{1}{2} \sum_{i,j} g^{\text{FR}}_{ij}(\xi(h)) \left(\epsilon \sum_a J_{ia}(h) v_a \right) \left(\epsilon \sum_b J_{jb}(h) v_b \right) + O(\epsilon^3) \\ &= \frac{1}{2} \epsilon^2 \sum_{a,b} \left(\sum_{i,j} J_{ia}(h) g^{\text{FR}}_{ij}(\xi(h)) J_{jb}(h) \right) v_a v_b + O(\epsilon^3) \\ &= \frac{1}{2} \epsilon^2 \sum_{a,b} (J(h)^\top g^{\text{FR}}(\xi(h)) J(h))_{ab} v_a v_b + O(\epsilon^3). \end{aligned}$$

The term $J(h)^\top g^{\text{FR}}(\xi(h)) J(h)$ is precisely the matrix representation of the pullback metric $g^*(h)$ in the standard coordinates of $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$, derived from equation (6.2). Thus, $D_{\text{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 g^*(h)(v,v) + O(\epsilon^3)$. The result for other well-behaved f-divergences follows from their similar second-order expansion involving g^{FR}. □

This theorem formally establishes that the pullback metric g^* measures how sensitive the output distribution p_h is to infinitesimal changes in the hidden state h, where sensitivity is gauged by the local divergence (specifically, KL divergence, relating to the Fisher-Rao metric) in the output space $\mathcal{P}(\mathbb{V})$.

Remark 6.2 (Pullback Metric as Expected Score Outer Product)**.** The Fisher-Rao metric g^{FR} is the expected outer product of the score function $\nabla_\xi \log p_\xi(W)$. This property pulls back to $\mathcal{H}$. Let $p_h(w) = k_{\text{head}}(h, \{w\})$. The score vector for token w with respect to the representation is $\nabla_h \log p_h(w) \in \mathcal{H}$. The pullback metric tensor is precisely the expected outer product of this score:

$$g^*(h) = \mathbb{E}_{W \sim p_h}[(\nabla_h \log p_h(W))(\nabla_h \log p_h(W))^\top]. \tag{6.4}$$

This directly connects the information geometry of $\mathcal{H}$ to the sensitivity of log-probabilities to changes in the representation h. This score vector $\nabla_h \log p_h(W)$ is analogous to that used in score-based generative models, but here taken with respect to the conditioning variable h.

The rank of the pullback metric depends on the dimensions of the spaces involved.

Proposition 6.3 (Rank of the Pullback Metric)**.** The rank of the pullback Fisher-Rao metric $g^*(h)$ at a point $h \in \mathcal{H}$ is bounded by the minimum of the representation dimension and the dimension of the probability simplex:

$$\text{rank}(g^*(h)) \leq \min(d_{\text{model}}, |\mathbb{V}| - 1). \tag{6.5}$$

Proof. The pullback metric $g^*(h)$ is defined as $g^*(h) = J(h)^\top g^{\text{FR}}(g_{\text{head}}(h)) J(h)$, where $J(h)$ is the Jacobian of the map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ (represented in appropriate local coordinates). The

dimension of $\mathcal{H}$ is d_{model}, and the dimension of $\mathcal{P}(\mathbb{V})$ is $d_{\text{prob}} = |\mathbb{V}| - 1$. The Jacobian $J(h)$ is a $d_{\text{prob}} \times d_{\text{model}}$ matrix. The Fisher-Rao metric g^{FR} at $g_{\text{head}}(h)$ is a $d_{\text{prob}} \times d_{\text{prob}}$ positive definite matrix (and thus has rank d_{prob}). Using the property that $\text{rank}(A^\top BA) = \text{rank}(A)$ if B is positive definite, we have $\text{rank}(g^*(h)) = \text{rank}(J(h))$. The rank of a matrix is bounded by its dimensions, so

$$\text{rank}(g^*(h)) \leq \min(d_{\text{model}}, d_{\text{prob}}) = \min(d_{\text{model}}, |\mathbb{V}| - 1).$$

□

6.1 Interpretation and Implications

This geometric perspective provides several insights:

- **Sensitivity Analysis:** The quadratic form $g^*(h)(v, v)$ quantifies the local distinguishability (via KL divergence, equation (6.3)) between the output distributions p_h and $p_{h+\epsilon v}$. It measures how sensitive the model's prediction is to perturbations of the hidden state h in a direction v. Directions v with large $g^*(h)(v, v)$ correspond to changes in h that significantly alter the output distribution.
- **Functional Anisotropy:** In modern LMs, the representation dimension is much smaller than the vocabulary size ($d_{\text{model}} \ll |\mathbb{V}|$). By Prop. 6.3, $\text{rank}(g^*(h)) \leq d_{\text{model}}$. This means the pullback metric $g^*(h)$ is necessarily a degenerate tensor on $\mathcal{H}$, but its rank is limited by the representation's capacity. More importantly, the metric is highly *anisotropic*. This geometrically formalizes "functional anisotropy": the learned mapping is highly sensitive in some directions of $\mathcal{H}$ (those with large eigenvalues for $g^*(h)$) but relatively insensitive in others. This is not a flaw but a feature, indicating learned specialization.
- **Spectrum of Sensitivity:** The eigenvalues and eigenvectors of the matrix for $g^*(h)$ reveal the principal directions of predictive sensitivity in $\mathcal{H}$. Directions with large eigenvalues are those where small changes in h induce large changes (geometrically measured by g^{FR}) in the predicted distribution p_h. These are the directions the model has learned are most important for prediction.

If two conditional distributions $p_{H_t|s_1}$ and $p_{H_t|s_2}$ are supported on regions of $\mathcal{H}$ that map to distinct regions in $\mathcal{P}(\mathbb{V})$ via g_{head}, the distance between these regions in $\mathcal{P}(\mathbb{V})$ (measured, e.g., by integrated Fisher-Rao distance or KL divergence) contributes to $\text{RepDiv}_D(s_1 \| s_2)$. The geometry induced by g^* characterizes the local separation capability. Training aims to shape the encoder ($k_{\text{bb}} \circ k_{\text{emb}}$) and the LM head k_{head} such that contexts with different predictive futures ($P_{\text{data}}(\cdot|s_1)$ vs $P_{\text{data}}(\cdot|s_2)$) are mapped to representations h_t whose images under g_{head} are appropriately separated in $\mathcal{P}(\mathbb{V})$, implicitly structuring the manifold $(\mathcal{H}, g^*)$.

7 NLL as Implicit Spectral Contrastive Learning

A central thesis of this work is that the simple objective of minimizing the negative log-likelihood (NLL) of the next token (equation (5.2)) implicitly functions as a powerful form of contrastive learning. While lacking the explicit positive/negative pairs of standard contrastive methods, we prove that NLL optimization inherently structures the learned representation space $\mathcal{H}$ according to predictive similarity. It achieves this by implicitly solving a spectral objective that aligns the

geometry of representations with the underlying predictive structure of the data $P_{\text{data}}(\cdot|x)$, a principle we formalize by connecting NLL to the eigenspectrum of a predictive similarity operator (HaoChen et al., 2021; Tan et al., 2024).

Let $f_{\text{enc}} : \mathbb{V}^* \to \mathcal{H}$ denote the deterministic encoder mapping a context sequence $x = \mathbf{w}_{<t}$ to its hidden representation $h_x = f_{\text{enc}}(x)$, implemented by the composition $k_{\text{bb}} \circ k_{\text{emb}}$. Let $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ be the deterministic mapping from the hidden state to the next-token distribution, corresponding to the LM head kernel k_{head}, such that $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$. The training objective is to minimize the expected KL divergence over the context distribution $\mu_{ctx} = p_{W_{<t}}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, g_{\text{head}}(f_{\text{enc}}(x)))] \tag{7.1}$$

where $P_{\text{data}}(\cdot|x)$ represents the true conditional distribution of the next token given context x, assumed to be derived from the data-generating process.

Successful optimization of $\mathcal{L}(\theta)$ drives the model's output distribution $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ towards the target distribution $P_{\text{data}}(\cdot|x)$ in the sense of minimizing average KL divergence. As we argue below, this fundamental requirement indirectly imposes geometric constraints on the distribution of representations $h_x = f_{\text{enc}}(x)$ in $\mathcal{H}$.

7.1 Constraint on Output Distribution Approximation

Minimizing the NLL loss (equation (7.1)) directly forces the model's predicted distribution $p_\theta(\cdot|x)$ to closely approximate the target distribution $P_{\text{data}}(\cdot|x)$. This closeness can be measured not only by KL divergence but also by other standard metrics on probability distributions, due to well-known inequalities relating them.

Theorem 7.1 (Output Distribution Approximation Constraint)**.** Assume the model parameters θ yield a small average KL divergence $\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))]$, where $p_\theta(\cdot|x) = g_{\text{head}}(f_{\text{enc}}(x))$. Let d_{out} be a metric on $\mathcal{P}(\mathbb{V})$ satisfying a Pinsker-type inequality, such as $d_{\text{out}}(p,q)^k \leq C \cdot D_{\text{KL}}(p\|q)$ for some constants $k, C > 0$. Examples include Hellinger distance (d_H, $k = 2, C = 1/2$) and Total Variation distance (d_{TV}, $k = 2, C = 1$). Then, the expected d_{out}-distance between the model's prediction and the true distribution is also small:

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathcal{L}(\theta). \tag{7.2}$$

Consequently, if the model fits the data well ($\mathcal{L}(\theta)$ is small), the distance between the model's output distributions for any two contexts, $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x'))$, must approximate the distance between the true target distributions, $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$. Specifically, contexts with predictively dissimilar target distributions (large LHS below) must yield model output distributions that are also separated:

$$d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \approx d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')). \tag{7.3}$$

Proof Sketch. The inequality equation (7.2) arises directly from the assumed Pinsker-type inequality $d_{\text{out}}(p,q)^k \leq C \cdot D_{\text{KL}}(p\|q)$. Letting $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$, we have $d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k \leq C \cdot D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))$. Taking the expectation over $x \sim \mu_{ctx}$ on both sides yields the result. For the second part, the triangle inequality states $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) \leq d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + d_{\text{out}}(p_x^\theta, p_{x'}^\theta) +$

$d_{\text{out}}(p^{\theta}_{x'}, p^{\text{data}}_{x'})$. Let $\epsilon_x = d_{\text{out}}(p^{\text{data}}_x, p^{\theta}_x)$. If $\mathcal{L}(\theta)$ is small, $\mathbb{E}_x[\epsilon_x^k]$ is small. By Markov's inequality, $\mathbb{P}(\epsilon_x \geq \delta) \leq \mathbb{E}[\epsilon_x^k]/\delta^k$, implying ϵ_x is small with high probability for typical x. Rearranging the triangle inequality bounds $d_{\text{out}}(p^{\theta}_x, p^{\theta}_{x'})$ by $d_{\text{out}}(p^{\text{data}}_x, p^{\text{data}}_{x'}) \pm (\epsilon_x + \epsilon_{x'})$, leading to equation (7.3). □

This theorem formalizes the intuition that minimizing the NLL objective forces the model's predictions to mirror the structure of the true predictive distributions, specifically in terms of their pairwise distances.

7.2 Consequences for Representation Geometry

Theorem 7.1 establishes that predictively dissimilar contexts x, x' must lead to distinct model output distributions $p_\theta(\cdot|x), p_\theta(\cdot|x')$. Since $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ and $p_\theta(\cdot|x') = g_{\text{head}}(h_{x'})$, this requirement imposes constraints on the corresponding representations $h_x = f_{\text{enc}}(x)$ and $h_{x'} = f_{\text{enc}}(x')$. Specifically, h_x and $h_{x'}$ must differ in ways that are discernible by the head mapping g_{head}. The information geometry of the head mapping, captured by the pullback metric $g^*(h)$ (Section 6), determines which differences in representation space are discernible.

Corollary 7.2 (Implicit Representation Separation)**.** Assume the model fits the data well ($\mathcal{L}(\theta)$ is small). If two contexts x, x' have predictively dissimilar target distributions, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then their learned representations $h_x = f_{\text{enc}}(x)$ and $h_{x'} = f_{\text{enc}}(x')$ must differ in a way that is detectable by the head mapping. Specifically, the difference vector $h_x - h_{x'}$ must have significant components along directions v of high predictive sensitivity (i.e., where the pullback metric quadratic form $g^*(h)(v, v)$ is large).

$$d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')) \text{ large} \implies \int_0^1 g^*(h_t)(h_x - h_{x'}, h_x - h_{x'})dt \text{ is large}, \tag{7.4}$$

where the integral is over the path $h_t = (1-t)h_{x'} + th_x$. Conversely, if contexts x, x' are predictively similar ($d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small), the NLL objective encourages their representations to be close along these same sensitive directions.

Proof Sketch. From Theorem 7.1, if $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be large. The squared distance between two points in a Riemannian manifold is related to the integrated metric along a geodesic. For small distances in output space, we have $d_{\text{out}}(p, q)^2 \approx D_{\text{KL}}(p\|q)$, which from equation (6.3) is related to the pullback metric g^*. A large distance between $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ implies a large integrated path length according to the pullback geometry, forcing h_x and $h_{x'}$ to differ along directions where g^* is large. □

This corollary establishes that NLL minimization implicitly acts like a contrastive learning objective: it pushes representations $h_x, h_{x'}$ apart if their corresponding contexts are predictively dissimilar. This differential pressure based on predictive similarity forms the basis for our connection to spectral methods.

7.3 Predictive Similarity Kernels

The preceding analysis suggests that NLL shapes the representation geometry based on the *dissimilarity* between the true next-token distributions $P_{\text{data}}(\cdot|x)$. To connect this to spectral methods, which operate on similarity structures, we formalize the complementary notion of *predictive similarity*.

Definition 7.3 (Predictive Similarity Kernel)**.** Let $p_x \coloneqq P_{\text{data}}(\cdot|x)$ denote the true conditional distribution for context x. A predictive similarity kernel is a bounded symmetric function $K : \mathbb{V}^* \times \mathbb{V}^* \to \mathbb{R}_{\geq 0}$ quantifying the similarity between p_x and $p_{x'}$. Examples include:

- **Bhattacharyya Coefficient Kernel**: $K_{\text{BC}}(x, x') \coloneqq \text{BC}(p_x, p_{x'}) = \sum_{w \in \mathbb{V}} \sqrt{p_x(w) p_{x'}(w)}$. This measures the cosine similarity between the square-root vectors $(\sqrt{p_x(w)})_w$. It relates directly to the Hellinger distance $d_H^2(p_x, p_{x'}) = 2(1 - K_{\text{BC}}(x, x'))$ and defines a positive semidefinite kernel. High K_{BC} corresponds to low d_H.
- **Hellinger-based Kernel (Gaussian Kernel on $\sqrt{p}$)**: $K_{\text{H}}(x, x') \coloneqq \exp(-\beta d_H^2(p_x, p_{x'}))$ for some scale $\beta > 0$. This explicitly converts the Hellinger distance into a similarity measure via a Gaussian function, yielding a positive semidefinite kernel.
- **Expected Likelihood Kernel (Linear Kernel)**: $K_{\text{Lin}}(x, x') \coloneqq \langle p_x, p_{x'} \rangle = \sum_{w \in \mathbb{V}} p_x(w) p_{x'}(w)$. This is the standard linear kernel (inner product) between probability vectors $p_x, p_{x'}$ and is positive semidefinite. High values indicate significant overlap between the distributions. It can be interpreted as the expected likelihood $p_{x'}(W)$ under $W \sim p_x$.
- **KL-based Kernel**: $K_{\text{KL}}(x, x') \coloneqq \exp(-\beta S_{\text{KL}}(p_x, p_{x'}))$ where S_{KL} is a symmetrized KL divergence (e.g., Jensen-Shannon divergence) and $\beta > 0$. Constructing a positive semidefinite kernel directly from KL divergence requires symmetrization.

In general, high values of $K(x, x')$ indicate high predictive similarity (i.e., small $d_{\text{out}}(p_x, p_{x'})$). To act on *representations*, we disintegrate K through the encoder f_{enc} by defining the induced kernel on $\mathcal{H}$:

$$\widetilde{K}(h, h') \triangleq \mathbb{E}\big[K(X, X') \;\big|\; f_{\text{enc}}(X) = h, f_{\text{enc}}(X') = h'\big], \tag{7.5}$$

whenever the conditional expectation exists. This produces a bounded symmetric measurable kernel on $(\mathcal{H}, \mu)$, where $\mu = (f_{\text{enc}})_{\#}\mu_{ctx}$, suitable for operator-theoretic analysis.

7.4 Connection to Graph Laplacian and Dirichlet Energy Minimization

Consider an undirected graph where contexts $x \in \mathbb{V}^*$ are nodes distributed according to μ_{ctx}, and edge weights are given by a symmetric predictive similarity kernel $K(x, x')$. The quadratic form of the associated graph Laplacian corresponds to the Dirichlet energy, which measures how "smooth" a function ϕ (e.g., a 1D projection of the representations) is over the graph.

$$\mathcal{E}_K(\phi) \coloneqq \frac{1}{2} \iint K(x, x')(\phi(x) - \phi(x'))^2 \, \mu_{ctx}(\mathrm{d}x) \mu_{ctx}(\mathrm{d}x') = \langle \phi, \Delta_K \phi \rangle_{L^2(\mu_{ctx})}. \tag{7.6}$$

Spectral clustering aims to find embeddings (represented by functions ϕ) that minimize this energy subject to constraints, effectively mapping similar contexts close together. The NLL objective, through Corollary 7.2, exerts a related pressure.

Recent work (Park et al., 2024) connects in-context learning to Dirichlet energy minimization on a task-similarity graph. Our work shows this is a foundational principle of NLL pre-training itself, where similarity is defined by the intrinsic next-token distributions $P_{\text{data}}(\cdot|x)$.

Proposition 7.4 (NLL and Implicit Dirichlet Energy Minimization)**.** Let $\phi_v(x) = \langle h_x, v \rangle$ be the projection of the representation $h_x = f_{\text{enc}}(x)$ onto a direction v of high predictive sensitivity (where

$g^*(h)(v, v)$ is large). Minimizing the NLL loss $\mathcal{L}(\theta)$ encourages configurations where the Dirichlet energy $\mathcal{E}_K(\phi_v)$ is small for large-$K(x, x')$ pairs.

Proof Sketch. The Dirichlet energy integrand $K(x, x')(\phi_v(x) - \phi_v(x'))^2$ is large only if both $K(x, x')$ and $(\phi_v(x) - \phi_v(x'))^2$ are large. Corollary 7.2 shows NLL encourages $(\phi_v(x) - \phi_v(x'))^2$ to be small when $K(x, x')$ is high. Conversely, if $K(x, x')$ is low, $(\phi_v(x) - \phi_v(x'))^2$ is encouraged to be large, but the integrand is suppressed by the small $K(x, x')$ term. Thus, NLL minimization discourages configurations that would lead to high Dirichlet energy. □

This proposition formalizes the link: NLL pushes representations towards configurations favored by spectral clustering on the predictive similarity graph.

7.5 NLL as Spectral Objective

We now strengthen this connection, showing that NLL optimization is not merely analogous to spectral methods, but that it implicitly solves a spectral objective. This viewpoint is closer to the analysis in (Tan et al., 2024).

Definition 7.5 (Predictive Similarity Operator)**.** Let $\mu = (f_{\text{enc}})_{\#}\mu_{ctx}$ on $\mathcal{H}$ and $\widetilde{K}$ be as in (7.5). Define the integral operator $M_{\widetilde{K}} : L^2(\mathcal{H}, \mu) \to L^2(\mathcal{H}, \mu)$ by

$$(M_{\widetilde{K}}\psi)(h) \coloneqq \int_{\mathcal{H}} \widetilde{K}(h, h')\, \psi(h')\, \mu(\mathrm{d}h'). \tag{7.7}$$

Under boundedness and symmetry, $M_{\widetilde{K}}$ is compact self-adjoint with a real nonnegative spectrum.

If K is symmetric, M_K is a compact self-adjoint operator whose eigenfunctions capture dominant patterns of predictive similarity. Our key result is that NLL training implicitly aligns the representation geometry with the eigenspace of M_K.

Theorem 7.6 (NLL Objective as Implicit Spectral Alignment under log-linear head)**.** Assume: (i) a log-linear head $p_\theta(w \mid x) \propto \exp\langle g(w), h_x\rangle$ with $\|g(w)\| = 1$; (ii) a norm constraint and whitening on representations $\mathbb{E}[h_x] = 0$, $\mathbb{E}[h_x h_x^\top] = I$; (iii) bounded partition function. Define $\bar{g}_x = \mathbb{E}_{w\sim P_{\text{data}}(\cdot|x)}[g(w)]$ and $K(x, x') = \langle \bar{g}_x, \bar{g}_{x'}\rangle$, with induced $\widetilde{K}$ on $\mathcal{H}$. Then minimizing the expected NLL is equivalent, up to constants and a uniformity term controlled by (ii)–(iii), to maximizing $\mathbb{E}_{x,x'}\big[\langle h_x, h_{x'}\rangle\, K(x, x')\big] = \langle \Phi,\ M_{\widetilde{K}}\Phi\rangle_{L^2(\mu)}$, $\Phi(h) \coloneqq \langle h, \cdot\rangle$, whose maximizers align h along the leading eigenspaces of $M_{\widetilde{K}}$. In particular, any stationary point that saturates the Rayleigh quotient corresponds to h supported on the top eigenfunctions of $M_{\widetilde{K}}$.

Proof Sketch. The softmax NLL decomposes into an *alignment* term $-\mathbb{E}_x\langle h_x, \bar{g}_x\rangle$ plus a *uniformity* log-partition term. Under whitening, maximizing the alignment term is equivalent to maximizing the correlation between $\langle h_x, h_{x'}\rangle$ and $K(x, x')$, yielding a Rayleigh quotient for $M_{\widetilde{K}}$. The uniformity term enforces spread and prevents degenerate scaling; boundedness ensures it contributes a constant under (ii)–(iii). Details in Appendix A.9. □

In summary, NLL optimization does not merely resemble spectral methods; it implicitly is a spectral method. The core objective of matching conditional probabilities imposes geometric constraints on $\mathcal{H}$ that are equivalent to finding a low-dimensional embedding of the predictive similarity graph.

This provides a rigorous, first-principles explanation for why NLL training produces representations that are so semantically and structurally organized.

8 Related Work

The theoretical understanding of representation learning in deep neural networks has been advanced along several parallel, yet largely disconnected, fronts. A significant challenge lies in the absence of a unified mathematical language capable of connecting a model's compositional architecture and its training dynamics to the emergent geometric structure of its learned representations. This work is situated at the confluence of these disparate research programs, aiming to synthesize the algebraic, compositional perspective of categorical probability with the metric, differential-geometric view of information geometry. We structure our review of the related literature into two parts. First, we introduce the foundational languages of probability that our framework unifies: the synthetic view of probability rooted in category theory and the metric view rooted in information geometry. Second, we survey the tools and concepts used to analyze the geometry of learned representations, focusing on the training objectives that guide learning, the spectral methods used to measure the resulting geometry, and the optimization mechanisms that shape it.

8.1 The Languages of Probability

The Synthetic View: Probability as a Category. A burgeoning field of research seeks to reformulate probability theory on a more abstract, algebraic foundation using the language of category theory (Baez et al., 2016; Fong and Spivak, 2018). This synthetic approach, in contrast to the classical analytic approach built upon measure theory, aims to derive probabilistic concepts from a small set of powerful axioms (Fritz, 2020; Perrone, 2023b). The central object of study in this domain is the *Markov category* (Cho and Jacobs, 2019; Fritz, 2020). Formally, a Markov category is a symmetric monoidal category where each object is equipped with a commutative comonoid structure, consisting of morphisms that represent the abstract operations of copying and discarding information (Cho and Jacobs, 2019; Fritz, 2020; Perrone, 2023b). The morphisms in such a category are interpreted as stochastic maps, or Markov kernels, which are probabilistic mappings between objects (Baez et al., 2016; Pardo-Guerra et al., 2025).

The pioneering work of Fritz (2020) has established Markov categories as a robust framework for synthetic probability and statistics. A key advantage of this formalism is its generality; it provides a uniform treatment of vastly different probabilistic settings. For instance, the category `FinStoch`, with finite sets as objects and stochastic matrices as morphisms, and the category `BorelStoch`, with standard Borel spaces as objects and their corresponding Markov kernels as morphisms, are both canonical examples of Markov categories (Fritz, 2020). This high level of abstraction allows for the proof of fundamental statistical theorems—such as the Fisher-Neyman factorization theorem and Kolmogorov's zero-one law—in a purely diagrammatic and synthetic manner, avoiding the low-level complexities of measure theory (Fritz, 2020; Fritz and Rischel, 2020). As Fritz (2020) argues, relying on measure theory is akin to programming in machine code, whereas the categorical approach provides a higher-level language that facilitates reasoning about complex, compositional systems.

This line of inquiry is not merely a formal exercise; it is directly motivated by the challenges of

Table 1 A Comparative Taxonomy of Theoretical Frameworks for Representation Learning. This table situates our proposed framework relative to existing theoretical paradigms, highlighting the unique synthesis of algebraic and geometric perspectives it offers to analyze the representation space directly.

Framework	Core Mathematical Object	Primary Domain of Analysis	Key Insight/Contribution	Limitation Addressed by Our Work	Key Citations
Information Geometry (IG)	Riemannian Manifold (Fisher-Rao Metric)	Model Parameter Space	Provides a metric for the "distance" between models; explains optimization via natural gradients.	Lacks a compositional, algebraic structure; difficult to apply to the dynamics of representations themselves.	Amari and Nagaoka (2000)
Categorical Probability	Markov Category (Symmetric Monoidal Category + Comonoids)	Abstract Probabilistic Systems	Provides a high-level, compositional language for probability, unifying disparate theories via algebraic axioms.	Inherently algebraic and non-metric; does not natively describe the shape or geometry of representation spaces.	Fritz (2020), Perrone (2023b), Cho and Jacobs (2019)
Information Bottleneck (IB)	Mutual Information	Information-Theoretic Channels	Defines representation learning as an optimal trade-off between compression and predictive relevance.	A high-level principle, not a constructive framework; does not specify the geometric mechanism for achieving the bottleneck.	Tishby et al. (2000)
Spectral Graph Theory	Graph Laplacian / Dirichlet Energy	Graph-structured Data / Activations	Quantifies the "smoothness" of functions on a graph; explains oversmoothing in GNNs and representation reorganization in LLMs.	A measurement tool, not a full-fledged theoretical framework; describes *what* geometry emerges but not *why*.	Yuan (2023); Park et al. (2024)
Implicit Bias Theory	Optimization Dynamics (Gradient Flow)	Model Weights & Activations	Explains generalization in overparameterized models; frames NLL as an implicit contrastive, geometry-sculpting process.	Provides a mechanism but lacks a unified language to describe the resulting geometric and probabilistic structures.	Gunasekar et al. (2018)
Our Framework (Proposed)	**Geometric Markov Category**	**Representation Space**	**Unifies the algebraic structure of probability with the metric geometry of representations.**	-	**This Paper**

understanding modern machine learning systems (Yuan, 2023). The compositional structure of deep neural networks, finds a natural description in the language of category theory (Fong and Spivak, 2018; Yuan, 2023; Pardo-Guerra et al., 2025).

The Metric View: Probability as a Manifold. In parallel to the algebraic developments in category theory, the field of *information geometry* (IG) has provided a powerful differential geometric lens for studying machine learning (Amari and Nagaoka, 2000). Foundational work by Amari and Nagaoka (2000) demonstrated that a parametric family of probability distributions can be viewed as a smooth manifold endowed with a canonical Riemannian metric—the Fisher-Rao metric—and a pair of dually-coupled affine connections. This geometric structure is not arbitrary; it can be intrinsically derived from statistical divergence functions, such as the Kullback-Leibler (KL) divergence, which serves as a measure of dissimilarity between distributions.

When applying these geometric tools to deep learning, it is crucial to draw a distinction between IG and the related field of *geometric deep learning* (GDL). GDL is primarily concerned with

generalizing neural network architectures to operate on data that resides in non-Euclidean domains, such as graphs or manifolds; its focus is the geometry of the *input data space*. In contrast, IG has traditionally been used to analyze the geometry of the *parameter space* of a model. By viewing the set of all possible model parameters as a manifold, IG provides sophisticated tools for understanding the dynamics of training, offering a more nuanced perspective on optimization than that afforded by standard ℓ_1 or ℓ_2 regularization (Amari and Nagaoka, 2000).

Towards Categorical Information Geometry. While the algebraic and geometric approaches have largely evolved independently, a new frontier is emerging at their intersection. Recent work has begun to explicitly forge a synthesis, aiming to create a *categorical information geometry* (Perrone, 2023b). This research program, led by researchers such as Perrone (2023b), seeks to enrich the abstract, compositional structures of Markov categories with the metric and quantitative notions central to information theory and geometry, such as entropy and divergence (Perrone, 2023a,b).

This emerging synthesis recognizes that a complete theoretical picture requires both the compositional language of categories and the metric language of geometry. However, to date, the applications of this nascent field have focused primarily on reformulating abstract probability theory. The critical connection to the analysis of *learned representations* in practical, large-scale deep learning models remains largely unexplored. This work aims to bridge that gap, demonstrating that a categorical information geometry provides the ideal framework for analyzing the structure of the representation spaces sculpted by the learning process.

8.2 The Geometry of Learning and Representation

Objectives of Learning Representations. A guiding principle for understanding the purpose of representation learning is the *Information Bottleneck* (IB) theory, introduced by Tishby et al. (2000). The IB principle posits that an optimal representation T of some input data X should be a bottleneck that is maximally informative about a relevant target variable Y while being maximally compressed with respect to the input X (Tishby et al., 2000; Shwartz-Ziv and Tishby, 2017). This trade-off between predictive accuracy and compressional complexity provides a powerful, high-level objective for representation learning.

The IB framework has been particularly influential in the theoretical analysis of deep learning. It led to the hypothesis that the training of deep neural networks proceeds in two distinct phases: an initial fitting phase, where the mutual information $I(T;Y)$ between the representation and the target increases, followed by a "compression" phase, where the mutual information $I(X;T)$ between the input and the representation decreases (Shwartz-Ziv and Tishby, 2017). While the universality of this two-phase dynamic is a subject of ongoing debate, the core intuition—that effective training involves not just memorization but also a form of structured compression—provides a compelling motivation for investigating the geometry of the learned representations.

Measurements of Representation Geometry. To move from high-level principles to concrete analysis, we require quantitative tools to probe the geometric properties of the high-dimensional activation spaces within a neural network. A particularly effective set of tools for this purpose comes from spectral graph theory. Given a graph with adjacency matrix A and degree matrix D, the *Graph Laplacian* is defined as $L = D - A$ (Berahmand et al., 2025). For any function f defined on the nodes of the graph (e.g., a feature activation), the quadratic form $f^\top L f$ defines the graph's

Dirichlet energy. This quantity measures the smoothness of the function with respect to the graph structure; a low Dirichlet energy indicates that connected nodes have similar function values (Park et al., 2024).

This concept has a rich history in machine learning, forming the basis of spectral clustering algorithms (Berahmand et al., 2025) and, more recently, being used to analyze and mitigate the over-smoothing problem in Graph Neural Networks (GNNs). Over-smoothing occurs when stacking many GNN layers causes the representations of all nodes to converge to an indistinguishable point, a phenomenon characterized by the Dirichlet energy of the representations collapsing to zero.

Most critically for the present work, this classical geometric tool has been shown to be deeply relevant to the internal dynamics of the most advanced models. A recent study demonstrates that during in-context learning, Large Language Models (LLMs) dynamically reorganize their internal concept representations in a manner that explicitly minimizes the Dirichlet energy with respect to an implicit graph structure defined by the context (Park et al., 2024). This groundbreaking result elevates Dirichlet energy from a tool for analyzing explicit graphs to a general principle governing the emergent geometry of learned representations. This trend towards spectral analysis is further evidenced by other recent work using methods like Centered Kernel Alignment (CKA) to track representation dynamics and spectral editing of activations (SEA) to control model behavior (Qiu et al., 2024).

Mechanisms of Learning Representations. The geometric structures observed in learned representations are not accidental; they are a direct consequence of the implicit biases of the training algorithm. For modern, highly overparameterized models, the optimization process itself, typically driven by variants of gradient descent, imparts an *implicit bias* or *implicit regularization* on the final solution (Vardi, 2023). Even when multiple parameter settings can achieve zero training error, the optimization algorithm preferentially converges to a "simple" solution that generalizes well. For linear models trained on classification tasks, this bias often corresponds to finding the maximum-margin separator, a classic geometric concept (Gunasekar et al., 2018; Soudry et al., 2018; Chizat and Bach, 2020).

A crucial modern insight is that training with the standard Negative Log-Likelihood (NLL) objective can be understood as a form of *implicit contrastive learning*. The NLL loss for a sample (x, y_{true}), given by $-\log P(y_{\text{true}}|x) = -\log \frac{\exp(z_{\text{true}})}{\sum_j \exp(z_j)}$, is minimized by simultaneously increasing the logit z_{true} for the correct class and decreasing the logits z_j for all incorrect classes. This dynamic is functionally equivalent to a contrastive loss that pulls the representation of x towards a "positive" target (the true class) while pushing it away from all negative targets (the incorrect classes). This connection is profound, as it links the vast body of work on self-supervised contrastive learning—which explicitly aims to sculpt a geometrically structured representation space—to standard supervised training. It provides the causal mechanism for why NLL training produces well-structured, geometrically organized representations: the objective itself is implicitly performing a contrastive optimization that enforces this structure.

9 Conclusion

In this work, we introduced a mathematical framework for analyzing the Autoregressive generation step in language models, leveraging the expressive power of Markov Categories. By modeling the process as a composition of Markov kernels, $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$, we established a foundation for a compositional, information-theoretic analysis. This allowed us to formally quantify the information surplus in hidden states, providing a clear theoretical rationale for the success of modern speculative decoding techniques like EAGLE.

More importantly, our framework provides a unified, first-principles explanation for the remarkable effectiveness of the negative log-likelihood (NLL) objective. We showed that NLL training is not merely about predicting the next token; it is a powerful structure-learning algorithm in disguise. We proved that minimizing NLL forces the model to: (1) achieve optimal data compression by learning the intrinsic conditional stochasticity of the data, a process we measured with categorical entropy; and (2) implicitly perform spectral contrastive learning. By analyzing the information geometry of the prediction head via the pullback Fisher-Rao metric, we demonstrated that NLL sculpts the representation space, aligning its geometry with the eigenspectrum of a predictive similarity operator. This explains how NLL learns semantically organized representations without explicit contrastive pairs.

This compositional, probabilistic, and information-geometric perspective offers a principled, mathematically grounded alternative to purely empirical or heuristic analysis, unifying concepts from information theory, geometry, and spectral methods to reveal the deep principles driving the success of large language models.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

John C Baez, Brendan Fong, and Blake S Pollard. A compositional framework for markov processes. *Journal of Mathematical Physics*, 57(3):033301, 2016.

Kamal Berahmand, Farid Saberi-Movahed, Razieh Sheikhpour, Yuefeng Li, and Mahdi Jalili. A comprehensive survey on spectral clustering with graph structure learning. *arXiv preprint arXiv:2501.13597*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.

Kenta Cho and Bart Jacobs. Disintegration and bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda

Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits, 2021.

Brendan Fong and David I Spivak. Seven sketches in compositionality: An invitation to applied category theory. *arXiv preprint arXiv:1803.05316*, 2018.

Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

Tobias Fritz and Eigil Fjeldgren Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality*, 2:3, 2020. doi: 10.32408/compositionality-2-3.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Jeff Z. HaoChen, Hao Chen, Chen Wei, Adrien Gaidon, and Tengyu Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 14239–14250, 2021.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Olav Kallenberg and Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, volume 29, 2016.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Sebastian Pardo-Guerra, Johnny Jingze Li, Kalyan Basu, and Gabriel A Silva. Neural networks and markov categories. *AppliedMath*, 5(3):93, 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. *arXiv preprint arXiv:2501.00070*, 2024.

Paolo Perrone. Markov categories and entropy. *IEEE Transactions on Information Theory*, 70(3): 1671–1692, 2023a.

Paolo Perrone. Categorical information geometry. In *International Conference on Geometric Science of Information*, pages 268–277. Springer, 2023b.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay Cohen. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. Contrastive Learning Is Spectral Clustering On Similarity Graph. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

Yang Yuan. On the power of foundation models. In *International Conference on Machine Learning*, pages 40519–40530. PMLR, 2023.

Appendix

A Full Proofs of Theorems

A.1 Proof of Theorem 5.1 (NLL Minimization as Average KL Minimization)

Let $p_x(\cdot) \coloneqq k_{\text{data}}(x, \cdot)$ denote the true conditional probability distribution $P_{\text{data}}(\cdot|x)$ for context $x = \mathbf{w}_{<t}$. Let $q_{x,\theta}(\cdot) = k_{\text{gen},\theta}(x, \cdot)$ denote the model's conditional probability distribution $P_\theta(\cdot|x)$. The context distribution is $p_{W_{<t}}$.

The cross-entropy loss is defined as:

$$\begin{aligned} L_{\text{CE}}(\theta) &= -\mathbb{E}_{(x,w)\sim P_{\text{data}}}[\log q_{x,\theta}(w)] \\ &= -\mathbb{E}_{x\sim p_{W_{<t}}}\left[\mathbb{E}_{W\sim p_x(\cdot)}[\log q_{x,\theta}(W)]\right] \\ &= -\mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \end{aligned}$$

The average KL divergence is defined as:

$$\begin{aligned} \mathcal{L}_{\text{KL}}(\theta) &= \mathbb{E}_{x\sim p_{W_{<t}}}[D_{\text{KL}}(p_x(\cdot)\,\|\,q_{x,\theta}(\cdot))] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log\frac{p_x(w)}{q_{x,\theta}(w)}\right] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log p_x(w) - \sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \\ &= \mathbb{E}_{x\sim p_{W_{<t}}}[-H(p_x(\cdot))] - \mathbb{E}_{x\sim p_{W_{<t}}}\left[\sum_{w\in\mathbb{V}} p_x(w)\log q_{x,\theta}(w)\right] \\ &= -H(W_t|W_{<t})_{\text{data}} + L_{\text{CE}}(\theta) \end{aligned}$$

where $H(p_x(\cdot))$ is the Shannon entropy of the distribution $p_x(\cdot)$, and $H(W_t|W_{<t})_{\text{data}} = \mathbb{E}_{x\sim p_{W_{<t}}}[H(p_x(\cdot))]$ is the average conditional Shannon entropy of the data generating process.

Rearranging gives:

$$L_{\text{CE}}(\theta) = \mathcal{L}_{\text{KL}}(\theta) + H(W_t|W_{<t})_{\text{data}}$$

Since $H(W_t|W_{<t})_{\text{data}}$ is a property of the data distribution and does not depend on the model parameters θ, minimizing $L_{\text{CE}}(\theta)$ with respect to θ is equivalent to minimizing $\mathcal{L}_{\text{KL}}(\theta)$.

The KL divergence $D_{\text{KL}}(p\|q) \geq 0$ for any probability distributions p, q, with equality if and only if $p = q$. Therefore, the average KL divergence $\mathcal{L}_{\text{KL}}(\theta) = \mathbb{E}_{x\sim p_{W_{<t}}}[D_{\text{KL}}(p_x(\cdot)\,\|\,q_{x,\theta}(\cdot))]$ is also non-negative, as it is an expectation of non-negative values. The minimum value is achieved if and only if $k_{\text{data}}(x, \cdot) = k_{\text{gen},\theta^*}(x, \cdot)$ for $p_{W_{<t}}$-almost every x. □

A.2 Proof of Proposition 4.2 (Categorical Entropy equals Shannon Entropy)

The categorical entropy for a single input h and $D = D_{\text{KL}}$ is:

$$\mathcal{H}_{D_{\text{KL}}}(k_{\text{head}})(h) = D_{\mathbb{V}\otimes\mathbb{V}}(\Delta_{\mathbb{V}} \circ k_{\text{head}}(h, \cdot) \parallel (k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)))\,.$$

Let $p_h(\cdot) = k_{\text{head}}(h, \cdot)$. The first distribution is $\sum_{w \in \mathbb{V}} p_h(w)\delta_{(w,w)}$, which is supported on the diagonal of $\mathbb{V} \times \mathbb{V}$. The second is the product distribution $p_h \otimes p_h$. The KL divergence is:

$$\begin{aligned}
&\sum_{(w,w') \in \mathbb{V} \times \mathbb{V}} \left(\sum_{v \in \mathbb{V}} p_h(v) \delta_{(v,v)}(w, w') \right) \log \left(\frac{\sum_{v \in \mathbb{V}} p_h(v) \delta_{(v,v)}(w, w')}{p_h(w) p_h(w')} \right) \\
&= \sum_{w \in \mathbb{V}} p_h(w) \log \left(\frac{p_h(w)}{p_h(w) p_h(w)} \right) \\
&= \sum_{w \in \mathbb{V}} p_h(w) \log \left(\frac{1}{p_h(w)} \right) = - \sum_{w \in \mathbb{V}} p_h(w) \log p_h(w) = H(p_h).
\end{aligned}$$

The average categorical entropy is then $\bar{\mathcal{H}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t}) = \mathbb{E}_{h \sim p_{H_t}}[H(p_h)]$, which is precisely the definition of the conditional Shannon entropy $H(W_t \mid H_t)$. □

A.3 Information Surplus Identity (Eq. 4.1)

The chain rule for mutual information states that for random variables A, B, C:

$$I(A; B, C) = I(A; B) + I(A; C \mid B).$$

Let $A = H_t$, $B = W_t$, and $C = W_{t+1:t+K-1}$. Substituting these into the chain rule gives:

$$I(H_t; W_t, W_{t+1:t+K-1}) = I(H_t; W_t) + I(H_t; W_{t+1:t+K-1} \mid W_t).$$

Recognizing that $(W_t, W_{t+1:t+K-1})$ is the sequence $W_{t:t+K-1}$, we have:

$$I(H_t; W_{t:t+K-1}) = I(H_t; W_t) + I(H_t; W_{t+1:t+K-1} \mid W_t).$$

This directly yields the identity in Eq. 4.1, where the information surplus is identified as the conditional mutual information term. □

A.4 Proof of Theorem 5.2 (Convergence of Average Categorical Entropy)

The theorem states that if the joint law of the hidden state and the predicted distribution converges, then the expectation of the categorical entropy also converges. We provide a more detailed argument.

Let $(H_t^{(n)}, P^{(n)})$ be the random variable pair on the product space $\mathcal{H} \times \Delta^{|\mathbb{V}|-1}$, where $H_t^{(n)} \sim p_{H_t, \theta_n}$ and $P^{(n)} = k_{\text{head}, \theta_n}(H_t^{(n)}, \cdot)$ is the corresponding output distribution. The average categorical entropy is the expectation of a function of this pair:

$$\bar{\mathcal{H}}_D(k_{\text{head},n}; p_{H_t, \theta_n}) = \mathbb{E}[\Psi_D(H_t^{(n)}, P^{(n)})].$$

The core assumption is that the sequence of joint laws of these pairs converges weakly to the law of a limiting pair $(H_t^\star, P^\star)$, denoted $(H_t^{(n)}, P^{(n)}) \Rightarrow (H_t^\star, P^\star)$.

Let $g : \mathcal{H} \times \Delta^{|\mathbb{V}|-1} \to \mathbb{R}$ be the function $g(h, p) = \Psi_D(h, p)$. We are given that Ψ_D is bounded and continuous in its second argument p. Since $\mathbb{V}$ is finite, the space of distributions $\Delta^{|\mathbb{V}|-1}$ is a compact metric space. A continuous function on a compact space is uniformly continuous. This ensures that $g(h, p)$ is a well-behaved function for the application of the Continuous Mapping Theorem.

The Continuous Mapping Theorem states that if a sequence of random variables X_n converges in distribution to X ($X_n \Rightarrow X$), and g is a continuous function (at least on the support of X), then $g(X_n) \Rightarrow g(X)$. Applying this to our setting:

$$(H_t^{(n)}, P^{(n)}) \Rightarrow (H_t^\star, P^\star) \implies \Psi_D(H_t^{(n)}, P^{(n)}) \Rightarrow \Psi_D(H_t^\star, P^\star).$$

This establishes that the random variable corresponding to the categorical entropy converges in distribution.

For the expectation to converge, i.e., $\lim \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$, we need convergence in distribution plus an additional condition, typically uniform integrability. A sequence of random variables $\{Y_n\}$ is uniformly integrable if, for any $\epsilon > 0$, there exists a $K < \infty$ such that $\mathbb{E}[|Y_n|\mathbf{1}_{|Y_n|>K}] < \epsilon$ for all n. A simpler, sufficient condition is that the sequence is uniformly bounded.

By assumption, the function $\Psi_D(h, p)$ is bounded. This means there exists a constant $M < \infty$ such that $|\Psi_D(h, p)| \leq M$ for all (h, p). Consequently, the sequence of random variables $Y_n = \Psi_D(H_t^{(n)}, P^{(n)})$ is uniformly bounded by M. Any uniformly bounded sequence of random variables is uniformly integrable.

Since we have both convergence in distribution and uniform integrability, we can conclude that the expectations converge:

$$\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{\text{head},n}; p_{H_t,\theta_n}) = \mathbb{E}[\Psi_D(H_t^{(n)}, P^{(n)})] \to \mathbb{E}[\Psi_D(H_t^\star, P^\star)] = \bar{\mathcal{H}}_D(k_{\text{head},\theta^\star}; p_{H_t,\theta^\star}).$$

If the model class is expressive and $k_{\text{gen},\theta^\star} = k_{\text{data}}$ almost surely, the limiting joint law of $(H_t^\star, P^\star)$ must be identical to that of the true data generating process. This implies the equality of the limiting average categorical entropy with that of the data. □

A.5 Proof of Theorem 7.1 (Output Distribution Approximation Constraint)

Bounding the expected output distance by the NLL loss. The theorem starts with the NLL objective, which is equivalent to minimizing the average KL divergence:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, p_\theta(\cdot|x))].$$

We are given a metric d_{out} on the space of probability distributions $\mathcal{P}(\mathbb{V})$ that satisfies a Pinsker-type inequality of the form:

$$d_{\text{out}}(p, q)^k \leq C \cdot D_{\text{KL}}(p\|q),$$

for some constants $k, C > 0$. This inequality holds for any pair of distributions $p, q \in \mathcal{P}(\mathbb{V})$.

We apply this inequality pointwise for each context $x \in \mathbb{V}^*$. Let $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$. This gives:

$$d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k \leq C \cdot D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x)).$$

This inequality relates the distance between the true and predicted distributions to the KL divergence for a single, fixed context x. To obtain a statement about the average behavior, we take the expectation of both sides with respect to the context distribution $x \sim \mu_{ctx}$. Since expectation is a positive linear operator, it preserves the inequality:

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq \mathbb{E}_{x\sim\mu_{ctx}}[C \cdot D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))].$$

Using the linearity of expectation on the right-hand side, we can pull the constant C out:

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathbb{E}_{x\sim\mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))].$$

The term on the right is precisely the definition of the average KL loss, $C \cdot \mathcal{L}(\theta)$. This proves the first part of the theorem:

$$\mathbb{E}_{x\sim\mu_{ctx}}[d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathcal{L}(\theta).$$

Approximation of pairwise distances. Let $p_x^{\text{data}} = P_{\text{data}}(\cdot|x)$ and $p_x^\theta = p_\theta(\cdot|x)$. We want to show that for any two contexts x, x', the distance $d_{\text{out}}(p_x^\theta, p_{x'}^\theta)$ approximates $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$.

The triangle inequality for the metric d_{out} states that for any three points A, B, C, we have $d_{\text{out}}(A, C) \leq d_{\text{out}}(A, B) + d_{\text{out}}(B, C)$. We apply this twice to relate the distance between the true distributions to the distance between the model's distributions:

$$\begin{aligned} d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) &\leq d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + d_{\text{out}}(p_x^\theta, p_{x'}^{\text{data}}) \\ &\leq d_{\text{out}}(p_x^{\text{data}}, p_x^\theta) + \left(d_{\text{out}}(p_x^\theta, p_{x'}^\theta) + d_{\text{out}}(p_{x'}^\theta, p_{x'}^{\text{data}})\right). \end{aligned}$$

Let $\epsilon_x = d_{\text{out}}(p_x^{\text{data}}, p_x^\theta)$ be the pointwise approximation error for context x. Rearranging the inequality above gives a lower bound on the model's output distance:

$$d_{\text{out}}(p_x^\theta, p_{x'}^\theta) \geq d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) - (\epsilon_x + \epsilon_{x'}).$$

Similarly, we can establish an upper bound:

$$d_{\text{out}}(p_x^\theta, p_{x'}^\theta) \leq d_{\text{out}}(p_x^\theta, p_x^{\text{data}}) + d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) + d_{\text{out}}(p_{x'}^{\text{data}}, p_{x'}^\theta) = d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}}) + (\epsilon_x + \epsilon_{x'}).$$

Combining these, we have:

$$|d_{\text{out}}(p_x^\theta, p_{x'}^\theta) - d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})| \leq \epsilon_x + \epsilon_{x'}.$$

If the model is well-trained, then $\mathcal{L}(\theta)$ is small. From previous derivations, this implies that $\mathbb{E}_{x\sim\mu_{ctx}}[\epsilon_x^k]$ is small. By Markov's inequality, for any $\delta > 0$:

$$\mathbb{P}(\epsilon_x \geq \delta) = \mathbb{P}(\epsilon_x^k \geq \delta^k) \leq \frac{\mathbb{E}[\epsilon_x^k]}{\delta^k} \leq \frac{C\mathcal{L}(\theta)}{\delta^k}.$$

Since the right-hand side can be made arbitrarily small by choosing a small $\mathcal{L}(\theta)$, this means that for a randomly drawn context x, the approximation error ϵ_x is small with high probability. For typical pairs of contexts (x, x'), both ϵ_x and $\epsilon_{x'}$ will be small. Therefore, the distance between the model's predictions $d_{\text{out}}(p_x^\theta, p_{x'}^\theta)$ will be a close approximation of the distance between the true data distributions $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$. □

A.6 Proof of Corollary 7.2 (Implicit Representation Separation)

From Theorem 7.1, if the model fits the data well, then for predictively dissimilar contexts x, x', the distance between their model output distributions, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$, must be large.

The space of output distributions $\mathcal{P}(\mathbb{V})$ is a Riemannian manifold endowed with the Fisher-Rao metric g^{FR}. The distance between two points on this manifold, p_1 and p_2, is the infimum of the lengths of all smooth paths connecting them. The length of a path $\gamma : [0,1] \to \mathcal{P}(\mathbb{V})$ is given by the integral:

$$L(\gamma) = \int_0^1 \sqrt{g^{\mathrm{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t))}\, dt.$$

The mapping $g_{\mathrm{head}} : \mathcal{H} \to \mathcal{P}(\mathbb{V})$ allows us to map paths from the representation space to the output space. Consider the straight-line path in representation space connecting $h_{x'}$ to h_x:

$$h(t) = (1-t)h_{x'} + t h_x, \quad \text{for } t \in [0,1].$$

The tangent vector to this path is constant: $h'(t) = h_x - h_{x'}$. This path in $\mathcal{H}$ induces a corresponding path in $\mathcal{P}(\mathbb{V})$ given by $\gamma(t) = g_{\mathrm{head}}(h(t))$. The tangent vector to this induced path is found using the chain rule:

$$\gamma'(t) = J_{g_{\mathrm{head}}}(h(t)) \cdot h'(t) = J_{g_{\mathrm{head}}}(h(t)) \cdot (h_x - h_{x'}),$$

where $J_{g_{\mathrm{head}}}(h)$ is the Jacobian of the map g_{head} evaluated at h.

The squared length of this tangent vector at point $\gamma(t)$ is given by the quadratic form of the metric g^{FR}:

$$\begin{aligned} g^{\mathrm{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t)) &= g^{\mathrm{FR}}_{g_{\mathrm{head}}(h(t))}(J_{g_{\mathrm{head}}}(h(t))(h_x - h_{x'}), J_{g_{\mathrm{head}}}(h(t))(h_x - h_{x'})) \\ &= (h_x - h_{x'})^\top \left[J_{g_{\mathrm{head}}}(h(t))^\top g^{\mathrm{FR}}_{g_{\mathrm{head}}(h(t))} J_{g_{\mathrm{head}}}(h(t)) \right] (h_x - h_{x'}). \end{aligned}$$

The term in the square brackets is precisely the definition of the pullback metric tensor g^* evaluated at the point $h(t) \in \mathcal{H}$. Thus, the squared length of the tangent vector is:

$$g^{\mathrm{FR}}_{\gamma(t)}(\gamma'(t), \gamma'(t)) = g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'}).$$

The total length of this specific path in $\mathcal{P}(\mathbb{V})$ is therefore:

$$L(\gamma) = \int_0^1 \sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}\, dt.$$

The Riemannian distance $d_{\mathrm{FR}}(g_{\mathrm{head}}(h_x), g_{\mathrm{head}}(h_{x'}))$ is the length of the shortest path (the geodesic), so it is bounded above by the length of our chosen path:

$$d_{\mathrm{FR}}(g_{\mathrm{head}}(h_x), g_{\mathrm{head}}(h_{x'})) \leq \int_0^1 \sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}\, dt.$$

If the contexts are predictively dissimilar, the left-hand side must be large (as standard distances like d_{out} are topologically equivalent to d_{FR} on the simplex). For the integral on the right-hand side to be large, the integrand $\sqrt{g^*_{h(t)}(h_x - h_{x'}, h_x - h_{x'})}$ must be large over a substantial portion of the integration interval $t \in [0,1]$. This can only happen if the difference vector $v = h_x - h_{x'}$ has significant components along the directions of high predictive sensitivity—that is, the directions where the quadratic form defined by the pullback metric g^* is large.

Conversely, if contexts are predictively similar, the LHS must be small. This encourages the model to find representations $h_x, h_{x'}$ such that the path integral is small, which is achieved by making $h_x - h_{x'}$ small along the directions where g^* is large. □

A.7 Proof of Proposition 7.4 (NLL Objective and Implicit Dirichlet Energy Minimization)

The Dirichlet energy of a function $\phi_v(x) = \langle h_x, v \rangle$ over the predictive similarity graph is given by:

$$\mathcal{E}_K(\phi_v) = \frac{1}{2} \iint K(x, x')(\phi_v(x) - \phi_v(x'))^2 \, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x').$$

Substituting the definition of ϕ_v:

$$\mathcal{E}_K(\phi_v) = \frac{1}{2} \iint K(x, x')(\langle h_x - h_{x'}, v \rangle)^2 \, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x').$$

We analyze the integrand $I(x, x') = K(x, x')(\langle h_x - h_{x'}, v \rangle)^2$ for a direction v of high predictive sensitivity. The NLL objective shapes the representations h_x learned by the encoder. We examine how this shaping affects the integrand for pairs of contexts (x, x').

Case 1: High Predictive Similarity. If two contexts x, x' are predictively similar, the similarity kernel $K(x, x')$ has a large value. By definition of the kernel, this implies that the true conditional distributions $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$ are close, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small. From the converse part of Corollary 7.2, when the target distributions are close, NLL training encourages their representations h_x and $h_{x'}$ to be close along directions of high predictive sensitivity. Therefore, for our chosen vector v, the term $(\langle h_x - h_{x'}, v \rangle)^2$ is encouraged to be small. In this case, the integrand $I(x, x')$ is the product of a large term $(K(x, x'))$ and a small term $((\langle h_x - h_{x'}, v \rangle)^2)$, resulting in a small value.

Case 2: Low Predictive Similarity. If two contexts x, x' are predictively dissimilar, the similarity kernel $K(x, x')$ has a small (near-zero) value. This implies that $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large. From Corollary 7.2, when the target distributions are far apart, NLL training forces their representations h_x and $h_{x'}$ to be separated along directions of high predictive sensitivity. This means the term $(\langle h_x - h_{x'}, v \rangle)^2$ is encouraged to be large. In this case, the integrand $I(x, x')$ is the product of a small term $(K(x, x'))$ and a large term. The small value of $K(x, x')$ acts as a weight, suppressing the contribution of this pair to the overall integral, resulting in a small value for the integrand.

In both cases, the NLL training objective pushes the learned representations towards a configuration where the integrand $I(x, x')$ is small for nearly all pairs (x, x'). By discouraging configurations that lead to a large integrand, the overall integral—the Dirichlet energy $\mathcal{E}_K(\phi_v)$—is implicitly encouraged to be small. This demonstrates that NLL training implicitly performs an optimization related to Dirichlet energy minimization on the graph defined by predictive similarity. □

A.8 Well-posedness of the Predictive Similarity Operator $M_{\widetilde{K}}$

The operator $M_{\widetilde{K}} : L^2(\mathcal{H}, \mu) \to L^2(\mathcal{H}, \mu)$ is defined by the integral $(M_{\widetilde{K}}\psi)(h) = \int_{\mathcal{H}} \widetilde{K}(h, h')\psi(h')\mu(\mathrm{d}h')$. For $M_{\widetilde{K}}$ to be a compact self-adjoint operator, its kernel $\widetilde{K}(h, h')$ must satisfy certain conditions.

1. **Symmetry:** Since the original kernel $K(x, x')$ is assumed to be symmetric, the disintegrated kernel $\widetilde{K}(h, h') = \mathbb{E}[K(X, X') \mid f(X) = h, f(X') = h']$ is also symmetric.

2. **Square-Integrability:** The space $(\mathcal{H}, \mu)$ is a probability space. A sufficient condition for compactness on a probability space is that the kernel is square-integrable, i.e., $\iint |\widetilde{K}(h, h')|^2 \mu(\mathrm{d}h)\mu(\mathrm{d}h') < \infty$. As we assumed K is a bounded function, its conditional expectation $\widetilde{K}$ is also bounded. A bounded measurable function on a finite measure space is always square-integrable.

Since $\widetilde{K}$ is symmetric and square-integrable with respect to the probability measure μ, it is a Hilbert-Schmidt kernel. Every Hilbert-Schmidt integral operator is compact. Because the kernel is also real and symmetric, the operator is self-adjoint. By the spectral theorem for compact self-adjoint operators, $M_{\widetilde{K}}$ has a discrete, real spectrum accumulating only at zero, and its eigenfunctions form an orthonormal basis for $L^2(\mathcal{H}, \mu)$. □

A.9 Proof of Theorem 7.6 (NLL as Implicit Spectral Alignment)

We provide a more detailed derivation for the claim that NLL optimization acts as a spectral objective under the conditions stated in Theorem 7.6.

Step 1: Decomposing the NLL Loss. We begin with the expected NLL loss under the log-linear head assumption (i):

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mu_{ctx}} \left[\mathbb{E}_{w\sim p_x}[-\log p_\theta(w|x)]\right],$$

where $p_x = P_{\text{data}}(\cdot|x)$ and $p_\theta(w|x) = \frac{\exp(\langle g(w), h_x\rangle)}{Z(h_x)}$ with $Z(h_x) = \sum_{w'} \exp(\langle g(w'), h_x\rangle)$.

Expanding the inner expectation over $w \sim p_x$:

$$\begin{aligned}
\mathcal{L}(\theta) &= \mathbb{E}_{x\sim\mu_{ctx}} \left[\sum_{w\in\mathbb{V}} p_x(w) \left(-\langle g(w), h_x\rangle + \log Z(h_x)\right)\right] \\
&= \mathbb{E}_{x\sim\mu_{ctx}} \left[-\left\langle \sum_{w\in\mathbb{V}} p_x(w) g(w), h_x \right\rangle + \left(\sum_{w\in\mathbb{V}} p_x(w)\right) \log Z(h_x)\right] \\
&= \underbrace{-\mathbb{E}_{x\sim\mu_{ctx}}[\langle \bar{g}_x, h_x\rangle]}_{\mathcal{L}_{\text{align}}} + \underbrace{\mathbb{E}_{x\sim\mu_{ctx}}[\log Z(h_x)]}_{\mathcal{L}_{\text{unif}}},
\end{aligned}$$

where we define the context-specific expected prototype as $\bar{g}_x := \mathbb{E}_{w\sim p_x}[g(w)]$.

The NLL loss consists of two competing terms:

- $\mathcal{L}_{\text{align}}$: To minimize this term (i.e., maximize its negative), the model must make the inner product $\langle \bar{g}_x, h_x\rangle$ as large as possible on average. By the Cauchy-Schwarz inequality, this encourages the representation vector h_x to be aligned with (point in the same direction as) its corresponding expected prototype vector $\bar{g}_x$.
- $\mathcal{L}_{\text{unif}}$: This term, the log-partition function, acts as a regularizer. If all h_x and $g(w)$ vectors were to align in a single direction, the inner products would be large, causing $Z(h_x)$ to grow exponentially and increasing this loss term. Minimizing this term encourages the representations h_x to be spread out, such that they are not simultaneously aligned with all prototype vectors $g(w)$. This is analogous to the role of negative samples in contrastive learning.

The theorem's assumptions are crucial for formalizing the link. Assumption (ii) states that representations are whitened: $\mathbb{E}[h_x] = 0$ and $\mathbb{E}[h_x h_x^\top] = I$. This is a strong form of regularization that forces the representations to be decorrelated and isotropic on average. Assumption (iii) states that the partition function is bounded. Together, these assumptions imply that the uniformity term $\mathcal{L}_{\text{unif}}$ is well-behaved and does not dominate the optimization. The primary driver for shaping the relative geometry of representations is therefore the alignment term, $\mathcal{L}_{\text{align}}$.

The goal of the optimization becomes maximizing $\mathbb{E}_x[\langle \bar{g}_x, h_x \rangle]$ subject to the whitening constraint. This encourages the distribution of $\{h_x\}$ to mimic the distribution of $\{\bar{g}_x\}$. This implies that the geometric relationships (inner products) between representations should reflect the geometric relationships between their corresponding expected prototypes:

$$\langle h_x, h_{x'} \rangle \quad \text{should be correlated with} \quad \langle \bar{g}_x, \bar{g}_{x'} \rangle.$$

We define the predictive similarity kernel as precisely this geometric relationship: $K(x, x') = \langle \bar{g}_x, \bar{g}_{x'} \rangle$.

Consider the objective of maximizing the total alignment between the representation Gram matrix and the similarity kernel matrix:

$$\max \mathcal{J} = \mathbb{E}_{x,x' \sim \mu_{ctx}}[\langle h_x, h_{x'} \rangle K(x, x')].$$

This can be expressed using the disintegrated kernel $\widetilde{K}$ on $\mathcal{H}$ and the operator $M_{\widetilde{K}}$. Let the function $\Phi_v(h) = \langle h, v \rangle$ represent projection onto a vector v. The objective seeks to find the directions v that are "most aligned" with the similarity structure. This leads to maximizing a Rayleigh quotient of the form:

$$R(\Phi) = \frac{\langle \Phi, M_{\widetilde{K}} \Phi \rangle_{L^2(\mu)}}{\langle \Phi, \Phi \rangle_{L^2(\mu)}}.$$

The maximizers of this quotient are the leading eigenfunctions of the operator $M_{\widetilde{K}}$. The optimization of the NLL objective, by forcing the representations $\{h_x\}$ to align with the structure defined by $\{\bar{g}_x\}$, effectively seeks a solution where the principal components of the representation distribution are aligned with the leading eigenfunctions of $M_{\widetilde{K}}$. The whitening constraint ensures the denominator of the Rayleigh quotient is well-behaved (equal to 1 for normalized directions), making the maximization direct.

Thus, under these specific modeling and regularization assumptions, NLL minimization is equivalent to solving a spectral problem on the graph of predictive similarity. □