Theory of Language Modeling Part I: A Markov Categorical Framework for Language Modeling

TOLM

Released: March 29, 2025
Revised V2: March 31, 2025

Abstract

Auto-regressive (AR) language models underpin state-of-the-art natural language processing, factorizing sequence probabilities as $P_\theta(\mathbf{w}) = \prod_t P_\theta(w_t|\mathbf{w}_{<t})$. While empirically powerful, their internal mechanisms remain partially understood. This work introduces a rigorous analytical framework using Markov Categories (MCs), specifically the category $\mathbf{Stoch}$ of standard Borel spaces and Markov kernels. We model the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$ as a composite kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$. Leveraging the enrichment of $\mathbf{Stoch}$ with statistical divergences D and associated categorical information measures (entropy $\mathcal{H}_D$, mutual information I_D), we define principled metrics: Representation Divergence $D(p_{H_t|s_1} \| p_{H_t|s_2})$, State-Prediction Information $I_D(H_t; W_t)$, Temporal Coherence $I_D(H_t; H_{t+1})$, LM Head Stochasticity $\mathcal{H}_D(k_{\text{head}})$, and Information Flow Bounds via the Data Processing Inequality (e.g., $I_D(S; H_t) \geq I_D(S; W_t)$). Beyond providing metrics, this framework analyzes the negative log-likelihood (NLL) objective itself. We argue NLL minimization equates to optimal compression and learning the data's intrinsic stochasticity ($\overline{\mathcal{H}}_D$). We employ information geometry, analyzing the pullback Fisher-Rao metric g^* on the representation space $\mathcal{H}$, to understand learned sensitivities. Furthermore, we formalize the concept that NLL acts as implicit structure learning, demonstrating how minimizing NLL forces representations of predictively dissimilar contexts apart, establishing connections to spectral graph theory principles based on predictive similarity kernels. This compositional, probabilistic, and information-geometric perspective offers a unified approach to dissecting information processing, representation learning, and the underlying principles driving the effectiveness of large language models.

1 Introduction

Auto-regressive language models (AR LMs), particularly those based on the Transformer architecture [23, 19, 5], 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 $\mathcal{V}$, the model learns a parameterized probability distribution P_θ that factorizes as:

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

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 $\mathcal{V}$ for the next token w_t.

Despite their empirical triumphs, a deep theoretical understanding of the internal information processing pathways and representation learning mechanisms within these large-scale models remains largely incomplete [14, 11, 7]. Current analytical approaches often rely on empirical probes of neural activations [10], correlation studies with linguistic features, or detailed analyses of specific architectural components like attention heads [16]. While insightful, these methods often lack a unified, mathematically principled framework to capture the compositional nature of the computation and the inherent stochasticity involved in the generation process. Furthermore, understanding why the simple NLL objective leads to representations capturing complex linguistic and world knowledge remains a fundamental question.

Recently, category theory has emerged as a powerful tool for analyzing the structural properties and capabilities of machine learning models, particularly foundation models [26]. Yuan [26] utilizes category theory to investigate the fundamental limits of foundation models trained on pretext tasks, characterizing the solvability of downstream tasks based on functor representability and exploring generalization through functors between categories.

This paper introduces a complementary, rigorous analytical framework focused on the internal mechanics of the AR generation step $\mathbf{w}_{<t} \mapsto P_\theta(\cdot|\mathbf{w}_{<t})$, rooted in the theory of Markov Categories (MCs) [6, 8]. MCs provide an abstract algebraic setting tailored for reasoning about systems involving probability, causality, conditioning, and information flow using the language of category theory. We specifically leverage the category **Stoch**, a canonical MC whose objects are standard Borel spaces (e.g., sequence spaces $\mathcal{V}^*$, representation spaces $\mathcal{H} \cong \mathbb{R}^{d_{\text{model}}}$, finite vocabularies $\mathcal{V}$) and whose morphisms are Markov kernels, representing conditional probability distributions [12, 8].

The strength of the MC framework lies in its inherent compositionality, mirroring the layered structure of deep neural networks, its native handling of probability and stochastic transformations, and its capacity for defining fundamental information-theoretic quantities in a principled manner. We model the complex computation within an AR LM as a sequence of morphisms (Markov kernels) composed in **Stoch**. Specifically, the generation process is factored as:

$$k_{\text{gen},\theta} \coloneqq k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V})). \tag{1}$$

Here k_{emb} and k_{bb} represent typically deterministic context embedding and backbone transformations yielding 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 the enrichment of $\mathbf{Stoch}$ with a statistical divergence D (e.g., D_{KL}, d_{TV}, Rényi α-divergence) [2, 18, 17]. Divergences quantify the dissimilarity between probability distributions and satisfy the Data Processing Inequality (DPI): processing through any Markov kernel cannot increase divergence. Building on this, Perrone [18] introduced intrinsic, categorical definitions of entropy $\mathcal{H}_D$ and mutual information I_D associated with D.

Leveraging this divergence-enriched Markov Category $(\mathbf{Stoch}, D)$, this paper provides a multifaceted analysis of AR LMs:

1. **Formal MC Model:** We provide a precise formulation of the AR LM's single-step generation process as a composite Markov kernel (Equation (1)) in $\mathbf{Stoch}$.
2. **Categorical Information Metrics:** We propose principled metrics based on categorical entropy and mutual information to quantitatively analyze information processing (Section 4):
   - **Representation Divergence** (RepDiv_D): $D(p_{H_t|s_1} \| p_{H_t|s_2})$ quantifies how well hidden states distinguish context properties s.
   - **Categorical Mutual Information**: Measures statistical dependencies like state-prediction relevance $I_D(H_t; W_t)$ and temporal state coherence $I_D(H_t; H_{t+1})$.
   - **LM Head Categorical Entropy** ($\mathcal{H}_D(k_{\text{head}})$): Measures the intrinsic stochasticity of the final prediction kernel.
   - **Information Flow Bounds**: Using the inherent DPI, we establish bounds like $I_D(S; H_t) \geq I_D(S; W_t)$, quantifying information preservation/loss.
3. **NLL Objective Interpretation:** We analyze the NLL pretraining objective within this framework (Section 5), arguing that minimizing average KL divergence ($\mathcal{L}_{\text{KL}}(\theta)$) corresponds to optimal compression and forces the model to learn the intrinsic stochasticity of the data, quantifiable via $\bar{\mathcal{H}}_D$.
4. **Information Geometry of Representations:** We connect the framework to information geometry (Section 6), analyzing the pullback Fisher-Rao metric g^* induced on the representation space $\mathcal{H}$ by the LM head. This metric reveals the local sensitivity of predictions to representational changes and highlights the functional geometry learned by the model.
5. **Implicit Structure Learning:** We formalize the idea that NLL optimization implicitly structures the representation space (Section 7). We show that minimizing NLL forces representations of contexts with dissimilar predictive distributions ($P_{\text{data}}(\cdot|x)$ vs $P_{\text{data}}(\cdot|x')$) to be separated along dimensions relevant to the predictor. This establishes rigorous connections, via propositions based on Dirichlet energy and predictive similarity operators, between NLL and the principles underlying spectral clustering methods.

This comprehensive framework offers a principled, mathematically grounded, and compositional alternative to heuristic or purely empirical analysis methods. By providing tools rooted in category theory, probability, information theory, and information geometry, it aims to facilitate a deeper quantitative understanding of information flow, compression, representation geometry, and the implicit structural biases induced by the standard NLL objective, thereby shedding light on the mechanisms underpinning the success of modern AR language models.

The paper is organized as follows: Section 2 provides the necessary mathematical background on Markov Categories, **Stoch**, divergences, and categorical information measures. Section 3 introduces the AR LM model within this framework. Section 4 formally define our proposed information-theoretic metrics. Section 5 connects the NLL objective to compression and learning intrinsic stochasticity. Section 6 explores the information geometry of the representation and prediction spaces. Section 7 develops the argument for NLL as implicit structure learning, linking it to spectral methods. Section 8 discusses related theoretical work, followed by the conclusion in Section 9.

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 [8].

Definition 2.1 (Markov Category [8])**.** *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 or channels transforming systems of type X to type Y. Composition $h \circ k$ denotes sequential processing, $k \otimes h$ parallel processing. The comonoid maps Δ_X (copy) and $!_X$ (discard) model the duplication and deletion of information. States (probability distributions) on X are morphisms $p : I \to X$.

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

Definition 2.2 (Category **Stoch** [8, 18])**.** *The Markov category* **Stoch** *is defined by:*

- ***Objects**: Standard Borel spaces $(X, \mathscr{B}(X))$. 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 $(h \circ k)(x, C) := \int_Y h(y, C)\, k(x, \mathrm{d}y)$ (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 $(k \otimes h)((x, y), \cdot) := k(x, \cdot) \otimes h(y, \cdot)$ (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 $D_X(p\|q)$ [18]. Examples include KL divergence (D_{KL}), Total Variation (d_{TV}), Rényi divergences (D_α), and the broad class of f-divergences (D_f) [1, 15].

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 $p, q : I \to X$:*

$$D_Y(k \circ p\|k \circ q) \leq D_X(p\|q) \tag{2}$$

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

Based on this, Perrone [18] introduced categorical definitions of entropy and mutual information intrinsically tied to the divergence D and the MC structure.

Definition 2.5 (Categorical Entropy and Mutual Information [18])**.** *Let $(\mathbf{Stoch}, D)$ be enriched with a DPI-satisfying divergence D.*

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

$$\mathcal{H}_D(k) := D_{Y\otimes Y}\left(\Delta_Y \circ k \quad \| \quad (k \otimes k) \circ \Delta_X\right) \tag{3}$$

   It compares two processes producing pairs in $Y \otimes Y$. The first ($\Delta_Y \circ k$) applies k once ($x \mapsto y \sim k(x,\cdot)$) and deterministically copies the output (y, y). The second ($(k \otimes k) \circ \Delta_X$) deterministically copies the input (x, x) and applies k independently to each component (y_1, y_2) where $y_1, y_2 \sim k(x,\cdot)$ are i.i.d. The divergence measures how different these two resulting joint distributions are, quantifying how far k is from being deterministic. If k is deterministic, $k = k_f$, both sides yield the same state (corresponding to $\delta_{(f(x),f(x))}$) and $\mathcal{H}_D(k_f) = 0$.

2. *The* ***Categorical Mutual Information*** *of a joint state $p : I \to X \otimes Y$ measures the statistical dependence between X and Y:*

$$I_D(p) := D_{X\otimes Y}\left(p \quad \| \quad p_X \otimes p_Y\right) \tag{4}$$

where $p_X = (\mathrm{id}_X \otimes !_Y) \circ p$ and $p_Y = (!_X \otimes \mathrm{id}_Y) \circ p$ are the marginal states. $I_D(p)$ measures how far the joint state p is from the product of its marginals (representing independence), according to the geometry induced by D.

Remark 2.6 (Properties and Connections)**.** *When $D = D_{\mathrm{KL}}$, $I_{D_{\mathrm{KL}}}(p)$ recovers the standard Shannon mutual information $I(X;Y)$ for the joint distribution p. $\mathcal{H}_{D_{\mathrm{KL}}}(k)$ provides an intrinsic measure of the kernel's stochasticity, related to but distinct from average conditional Shannon entropy [18]. Crucially, these categorical definitions automatically satisfy the DPI. For instance, consider a state $p_{XY} : I \to X \otimes Y$ and a kernel $h : Y \to Z$. Let p_{XZ} be the state obtained by applying $\mathrm{id}_X \otimes h$ to p_{XY}. The DPI for D applied to the states involved in the definition of I_D implies $I_D(p_{XY}) \geq I_D(p_{XZ})$ [18, Prop. 4.8]. This reflects the principle that processing ($Y \to Z$) cannot increase information about X. Furthermore, information geometry [1] arises naturally: the Fisher-Rao metric is induced by the local quadratic approximation of the KL divergence, linking the divergence D to the underlying geometric structure of the space of probability measures.*

3 Auto-Regressive Language Models as Composed Kernels

We now apply the Markov Category framework established in Section 2 to model auto-regressive 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 $\mathbf{Stoch}$.

The relevant measurable spaces (objects in $\mathbf{Stoch}$) are:

- Input context space: $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) = (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where $\mathcal{V}^*$ is the set of finite sequences over the vocabulary $\mathcal{V}$, equipped with a suitable σ-algebra making it standard Borel (e.g., considering it as a disjoint union of finite products $\mathcal{V}^n$).
- Initial sequence representation space: $(\mathcal{H}_{\mathrm{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\mathrm{seq\_emb}})) = (\mathcal{H}_{\mathrm{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\mathrm{seq\_emb}}))$, the space of initial vector sequences (e.g., $\bigcup_n (\mathbb{R}^{d_{\mathrm{model}}})^n$), also equipped with a standard Borel structure.
- Final hidden state space: $(\mathcal{H}, \mathscr{B}(\mathcal{H})) = (\mathcal{H}, \mathscr{B}(\mathcal{H}))$, typically $(\mathbb{R}^{d_{\mathrm{model}}}, \mathscr{B}(\mathbb{R}^{d_{\mathrm{model}}}))$.
- Output vocabulary space: $(\mathcal{V}, \mathcal{P}(\mathcal{V})) = (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, 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 [12].

The generation process decomposes into three kernels (morphisms in $\mathbf{Stoch}$):

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

$$k_{\mathrm{emb}}(\mathbf{w}_{<t}, A) := \delta_{f_{\mathrm{emb}}(\mathbf{w}_{<t})}(A) = \mathbf{1}_A(f_{\mathrm{emb}}(\mathbf{w}_{<t})), \quad \text{for } A \in \mathscr{B}(\mathcal{H}_{\mathrm{seq\_emb}}). \tag{5}$$

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) [20], 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:

$$k_{\text{bb}}(E_{<t}, B) := \delta_{f_{\text{bb}}(E_{<t})}(B) = \mathbf{1}_B(f_{\text{bb}}(E_{<t})), \quad \text{for } B \in \mathscr{B}(\mathcal{H}). \tag{6}$$

This is also a morphism in **Stoch**.

3. **LM Head Kernel ($k_{\text{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$):** This final kernel maps the summary hidden state $h_t \in \mathcal{H}$ to a probability distribution over the finite vocabulary $\mathcal{V}$. Typically, h_t is passed through a linear layer ($f_{\text{head}} : \mathcal{H} \to \mathbb{R}^{|\mathcal{V}|}$) 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:

$$k_{\text{head}}(h, A) := \sum_{w \in A} [\text{softmax}(f_{\text{head}}(h))]_w \quad \text{for } h \in \mathcal{H}, A \subseteq \mathcal{V}. \tag{7}$$

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

The overall single-step generation kernel $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ 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 $P_\theta(\cdot|\mathbf{w}_{<t})$. The subsequent sections will use this representation to define and analyze information-theoretic metrics.

4 Markov Categorical Metrics

We now apply the Markov category framework (**Stoch**, D) to analyze the AR generation kernel $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ (Equation (1)). We select a suitable statistical divergence D satisfying the Data Processing Inequality (DPI) (e.g., D_{KL}, d_{TV}, or more generally an f-divergence [1, 15]) and utilize the corresponding categorical information measures $\mathcal{H}_D$ and I_D (Equations (3) and (4)) to probe the information flow and transformations within the generation step. A particular focus is placed on the final hidden state $H_t \in \mathcal{H}$ and the stochastic prediction kernel k_{head}.

We operate within the probabilistic setting induced by a distribution over input contexts. Let P_{ctx} be a probability measure on the context space $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) = (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$. This corresponds to an initial *state* in the Markov category **Stoch**, represented by a morphism $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, where I is the monoidal unit (a singleton measurable space) and $p_{W_{<t}}(\star, A) = P_{\text{ctx}}(A)$ for any $A \in \mathscr{B}(\mathcal{V}^*)$. Processing this initial state through the sequence of deterministic kernels k_{emb} and k_{bb}, and the stochastic kernel k_{head}, induces distributions (states) at subsequent stages:

- **Initial Sequence Embedding State**: Given $p_{W_{<t}} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, the distribution of the initial vector sequence representation $E_{<t} \in \mathcal{H}_{\text{seq\_emb}}$ is given by the state $p_{E_{<t}} : I \to$

$(\mathcal{H}_{\text{seq\_emb}}, \mathscr{B}(\mathcal{H}_{\text{seq\_emb}}))$, defined as:

$$p_{E_{<t}} \coloneqq k_{\text{emb}} \circ p_{W_{<t}}. \tag{8}$$

Since k_{emb} corresponds to the deterministic function f_{emb}, the measure associated with $p_{E_{<t}}$ is the pushforward measure $(P_{\text{ctx}}) \circ f_{\text{emb}}^{-1}$.

- **Final Hidden State**: The distribution of the final hidden state $H_t \in \mathcal{H}$ is given by the state $p_{H_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H}))$:

$$p_{H_t} \coloneqq k_{\text{bb}} \circ p_{E_{<t}} = (k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}}. \tag{9}$$

  As k_{bb} is also deterministic (representing f_{bb}), p_{H_t} corresponds to the pushforward measure $(P_{\text{ctx}}) \circ (f_{\text{bb}} \circ f_{\text{emb}})^{-1}$.

- **Predicted Next Token State**: The marginal distribution of the predicted next token $W_t \in \mathcal{V}$, averaged over all contexts according to P_{ctx}, is given by the state $p_{W_t} : I \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$:

$$p_{W_t} \coloneqq k_{\text{head}} \circ p_{H_t} = (k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}} = k_{\text{gen},\theta} \circ p_{W_{<t}}. \tag{10}$$

  Let μ_{H_t} be the measure on $\mathcal{H}$ associated with p_{H_t}. Then the measure associated with p_{W_t} on $\mathcal{V}$ is given by $\mu_{W_t}(A) = \int_{\mathcal{H}} k_{\text{head}}(h, A)\, \mu_{H_t}(\mathrm{d}h)$ for $A \subseteq \mathcal{V}$.

The random variables corresponding to these stages are denoted $W_{<t}$ (context sequence), $E_{<t}$ (initial embedding sequence), H_t (final hidden state), and W_t (predicted next token). Their respective distributions (states in **Stoch**) are denoted $p_{W_{<t}}$, $p_{E_{<t}}$, p_{H_t}, and p_{W_t}. Using these rigorously defined states and the categorical information measures, we propose the following metrics.

4.1 Metric 1: Representation Divergence (Context Encoding Fidelity)

To quantify how effectively the distribution of the final hidden state H_t distinguishes between different underlying properties S of the input context $\mathbf{w}_{<t}$.

Consider a random variable S, defined on the probability space underlying P_{ctx}, representing a specific property of the context $\mathbf{w}_{<t}$ (e.g., topic membership $s \in \{s_1, s_2, \dots\}$, presence/absence of a feature). Assume we can condition the context distribution P_{ctx} on the value of S. Let $P_{\text{ctx}}(\cdot|S = s)$ denote the conditional probability measure on $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$. This corresponds to a conditional input state $p_{W_{<t}|s} : I \to (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$ for each value s of S. The conditional distribution of the hidden state H_t given $S = s$ is then represented by the state:

$$p_{H_t|s} \coloneqq (k_{\text{bb}} \circ k_{\text{emb}}) \circ p_{W_{<t}|s} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})). \tag{11}$$

Let $\mu_{H_t|s}$ be the probability measure on $\mathcal{H}$ associated with the state $p_{H_t|s}$.

Definition 4.1 (Representation Divergence). *The Representation Divergence between contexts exhibiting properties s_1 and s_2 is defined as the statistical divergence D between the corresponding conditional hidden state measures:*

$$\textit{RepDiv}_D(s_1\|s_2) \coloneqq D_{\mathcal{H}}(\mu_{H_t|s_1}\|\mu_{H_t|s_2}) \equiv D_{\mathcal{H}}(p_{H_t|s_1}\|p_{H_t|s_2}). \tag{12}$$

Here, $D_{\mathcal{H}}$ denotes the application of the divergence functional D to probability measures (states) on the measurable space $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$.

Interpretation. A large value of $\mathrm{RepDiv}_D(s_1 \| s_2)$ indicates that the measures $\mu_{H_t|s_1}$ and $\mu_{H_t|s_2}$ are highly distinguishable according to the chosen divergence D. This implies that the transformation $(k_{\mathrm{bb}} \circ k_{\mathrm{emb}})$ maps contexts with properties s_1 and s_2 to significantly different distributions in the representation space $\mathcal{H}$. The hidden state H_t thus serves as an effective statistical signature for distinguishing between properties s_1 and s_2. Conversely, a small divergence suggests that the representations generated from contexts with properties s_1 and s_2 are statistically similar, implying that the model either does not encode this specific distinction strongly in H_t or represents them in overlapping regions of $\mathcal{H}$. The choice of D influences the notion of distinguishability (e.g., D_{KL} emphasizes differences in likelihood ratios, while d_{TV} focuses on the maximal difference in probability assigned to any event).

Estimation. Estimating $D_{\mathcal{H}}(\mu_{H_t|s_1} \| \mu_{H_t|s_2})$ can use following techniques:

- *Non-parametric methods*: Estimators based on k-nearest neighbor distances can approximate certain divergences like D_{KL} [24] or Rényi divergences. However, their statistical efficiency degrades rapidly with increasing dimension (curse of dimensionality), requiring a large number of samples h_t drawn from each conditional distribution.
- *Variational methods*: Techniques utilizing neural networks to estimate density ratios or variational bounds offer potential alternatives that may scale better with dimension. Examples include methods for f-divergences [15] and KL divergence (e.g., MINE [3] using the Donsker-Varadhan representation: $D_{\mathrm{KL}}(\mu \| \nu) = \sup_T(\mathbb{E}_\mu[T] - \log \mathbb{E}_\nu[e^\top])$, where $T : \mathcal{H} \to \mathbb{R}$ is a critic function). These methods introduce optimization complexity and potential biases from the critic network's capacity.

Estimation requires sampling contexts $\mathbf{w}_{<t}^{(i)} \sim P_{\mathrm{ctx}}(\cdot|s_1)$ and $\mathbf{w}_{<t}^{(j)} \sim P_{\mathrm{ctx}}(\cdot|s_2)$, computing the corresponding hidden states $h_t^{(i)} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{<t}^{(i)})$ and $h_t^{(j)} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{<t}^{(j)})$, and feeding these samples $\{h_t^{(i)}\}$ and $\{h_t^{(j)}\}$ into the chosen divergence estimator.

4.2 Metric 2: Categorical Mutual Information (Statistical Dependencies)

To measure the strength of statistical dependence between key random variables involved in the generation step, using the intrinsic definition of mutual information within the Markov Category framework.

We use the categorical mutual information I_D (Equation (4)), which measures the D-divergence between a joint state and the product of its marginals.

Definition 4.2 (State-Prediction Dependence ($I_D(H_t; W_t)$))**.** *We aim to quantify the statistical dependence between the final hidden state H_t and the predicted next token W_t. This requires defining the joint state $p_{H_t,W_t} : I \to (\mathcal{H}, \mathscr{B}(\mathcal{H})) \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, representing the joint distribution of (H_t, W_t) induced by the process $p_{W_{<t}} \to p_{H_t} \to p_{W_t}$. This state is obtained categorically by taking the state p_{H_t}, copying the $\mathcal{H}$ component using $\Delta_{\mathcal{H}} : \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$, and then applying the LM head kernel k_{head} only to the second component using the tensored identity* $\mathrm{id}_{\mathcal{H}} \otimes k_{\mathrm{head}}$*:*

$$p_{H_t,W_t} := (\mathrm{id}_{\mathcal{H}} \otimes k_{\mathrm{head}}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \tag{13}$$

The marginal states p_{H_t} and p_{W_t} can be recovered from p_{H_t,W_t} by discarding the other component using the unique maps $!_h : \mathcal{V} \to I$ and $!_{\mathcal{H}} : \mathcal{H} \to I$: $p_{H_t} = (\mathrm{id}_{\mathcal{H}} \otimes !_h) \circ p_{H_t,W_t}$ and $p_{W_t} = (!_{\mathcal{H}} \otimes \mathrm{id}_{\mathcal{V}}) \circ p_{H_t,W_t}$. The *categorical mutual information between H_t and W_t is then defined as:*

$$I_D(H_t; W_t) \coloneqq I_D(p_{H_t,W_t}) \equiv D_{\mathcal{H}\otimes\mathcal{V}}(p_{H_t,W_t} \quad \| \quad p_{H_t} \otimes p_{W_t}). \tag{14}$$

This measures the deviation of the joint distribution from independence, according to the geometry induced by D. If $D = D_{\mathrm{KL}}$, this recovers the standard Shannon mutual information $I(H_t; W_t)$.

Definition 4.3 (Temporal State Dependence ($I_D(H_t; H_{t+1})$))**.** *To analyze the coherence between consecutive hidden states, we need to model the transition from H_t to H_{t+1}. This involves generating $W_t \sim k_{\mathrm{head}}(H_t, \cdot)$, updating the context $\mathbf{w}_{\le t} = \mathbf{w}_{<t}W_t$, and then computing $H_{t+1} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{\le t})$. This defines a complex transition kernel $k_{step} : \mathcal{H} \to \mathcal{H}$ which implicitly depends on the full history through H_t and the generated W_t. Assuming we average over the generation of W_t and the distribution p_{H_t}, we can define an effective transition kernel $\bar{k}_{step} : \mathcal{H} \to \mathcal{H}$. The joint state $p_{H_t,H_{t+1}} : I \to \mathcal{H} \otimes \mathcal{H}$ is constructed similarly to Equation (13):*

$$p_{H_t,H_{t+1}} \coloneqq (\mathrm{id}_{\mathcal{H}} \otimes \bar{k}_{step}) \circ \Delta_{\mathcal{H}} \circ p_{H_t}. \tag{15}$$

The temporal statistical dependence is measured by:

$$I_D(H_t; H_{t+1}) \coloneqq I_D(p_{H_t,H_{t+1}}) \equiv D_{\mathcal{H}\otimes\mathcal{H}}(p_{H_t,H_{t+1}} \quad \| \quad p_{H_t} \otimes p_{H_{t+1}}), \tag{16}$$

where $p_{H_{t+1}} = (!_{\mathcal{H}} \otimes \mathrm{id}_{\mathcal{H}}) \circ p_{H_t,H_{t+1}} = \bar{k}_{step} \circ p_{H_t}$ is the marginal state at time $t + 1$.

Interpretation. $I_D(H_t; W_t)$ quantifies the average amount of information (relative to divergence D) that the hidden state H_t provides about the immediately following token W_t. A high value suggests H_t strongly constrains the distribution over W_t, indicating high predictive relevance. $I_D(H_t; H_{t+1})$ measures the average statistical dependency between consecutive hidden states. A high value implies that the state H_{t+1} is highly predictable from H_t, suggesting the model maintains and evolves contextual information coherently over time. Low values might indicate information loss or abrupt changes in representation between time steps.

Estimation Challenges. Estimating I_D involving the high-dimensional continuous variable H_t (and potentially H_{t+1}) is difficult.

- For $I_D(H_t; W_t)$: W_t is discrete (finite vocabulary $\mathcal{V}$), while H_t is high-dimensional continuous. Mutual information estimation in such mixed settings is non-trivial. One could adapt general high-dimensional MI estimators like kNN-based methods (e.g., Kraskov-Stögbauer-Grassberger [13]) or variational methods (e.g., MINE [3]) by treating W_t appropriately (e.g., conditioning or embedding).
- For $I_D(H_t; H_{t+1})$: Both variables are high-dimensional and continuous. This poses the most significant estimation challenge, requiring robust high-dimensional MI estimators (kNN or variational) and potentially large sample sizes.

Estimation requires generating trajectories: sample $\mathbf{w}_{<t} \sim P_{\mathrm{ctx}}$, compute $h_t = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{<t})$, sample $w_t \sim k_{\mathrm{head}}(h_t, \cdot)$, form $\mathbf{w}_{\le t} = \mathbf{w}_{<t}w_t$, compute $h_{t+1} = (f_{\mathrm{bb}} \circ f_{\mathrm{emb}})(\mathbf{w}_{\le t})$, and collect pairs (h_t, w_t) and (h_t, h_{t+1}) for input into the chosen MI estimator.

4.3 Metric 3: 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 $k_{\mathrm{head}} : (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$.

We focus on the properties of the kernel k_{head} itself. The definition of categorical entropy (Equation (3)) involves comparing two processes that generate pairs of outputs in $Y \otimes Y$ from inputs in X, where $k : X \to Y$.

Definition 4.4 (Categorical Entropy of k_{head})**.** *The Categorical Entropy of k_{head} is defined using Equation* (3) *with $X = \mathcal{H}$, $Y = \mathcal{V}$, and $k = k_{\text{head}}$:*

$$\mathcal{H}_D(k_{\text{head}}) \coloneqq D_{\mathcal{V}\otimes\mathcal{V}}\left(\Delta_{\mathcal{V}} \circ k_{\text{head}} \quad \| \quad (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}\right). \tag{17}$$

Let's analyze the two morphisms inside the divergence $D_{\mathcal{V}\otimes\mathcal{V}}(\cdot\|\cdot)$, which map $\mathcal{H} \to \mathcal{V} \otimes \mathcal{V}$:

- $k_1 = \Delta_{\mathcal{V}} \circ k_{\text{head}}$: For an input $h \in \mathcal{H}$, this first applies k_{head} to get a distribution $p_h(\cdot) = k_{\text{head}}(h, \cdot)$ on $\mathcal{V}$. Then, it applies the deterministic copy map $\Delta_{\mathcal{V}}(w, \cdot) = \delta_{(w,w)}$. The resulting kernel $k_1(h, \cdot)$ corresponds to sampling $w \sim p_h(\cdot)$ and then outputting the pair (w, w). The measure is $\sum_{w\in\mathcal{V}} p_h(w)\delta_{(w,w)}$ on $\mathcal{V} \otimes \mathcal{V}$.
- $k_2 = (k_{\text{head}} \otimes k_{\text{head}}) \circ \Delta_{\mathcal{H}}$: For an input $h \in \mathcal{H}$, this first applies the deterministic copy map $\Delta_{\mathcal{H}}(h, \cdot) = \delta_{(h,h)}$, producing the pair (h, h). Then, it applies k_{head} independently to each component via the tensor product kernel $(k_{\text{head}}\otimes k_{\text{head}})((h, h), \cdot) = k_{\text{head}}(h, \cdot)\otimes k_{\text{head}}(h, \cdot)$. The resulting kernel $k_2(h, \cdot)$ corresponds to sampling $w_1 \sim p_h(\cdot)$ and $w_2 \sim p_h(\cdot)$ independently and outputting the pair (w_1, w_2). The measure is $p_h \otimes p_h$ (the product measure) on $\mathcal{V} \otimes \mathcal{V}$.

The categorical entropy $\mathcal{H}_D(k_{\text{head}})$ thus measures the divergence between generating (W, W) where $W \sim p_h(\cdot)$ and generating (W_1, W_2) where $W_1, W_2 \overset{\text{i.i.d.}}{\sim} p_h(\cdot)$.

This divergence $D_{\mathcal{V}\otimes\mathcal{V}}(\sum_w p_h(w)\delta_{(w,w)}\|p_h \otimes p_h)$ quantifies how far p_h is from a point mass (Dirac measure), averaged appropriately over the input space $\mathcal{H}$. Note that $\mathcal{H}_D(k)$ as defined in [18] can be interpreted as a state on X (specifically $X = \mathcal{H}$ here) whose value at h relates to this divergence. A common practical approach is to consider the average divergence with respect to the input distribution p_{H_t}:

$$\bar{\mathcal{H}}_D(k_{\text{head}}; p_{H_t}) \coloneqq \mathbb{E}_{h\sim p_{H_t}}\left[D_{\mathcal{V}\otimes\mathcal{V}}\left(\sum_{w\in\mathcal{V}} k_{\text{head}}(h, \{w\})\delta_{(w,w)} \quad \| \quad k_{\text{head}}(h, \cdot) \otimes k_{\text{head}}(h, \cdot)\right)\right]. \tag{18}$$

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 $\mathcal{H}_D(k_{\text{head}})$ 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 $\mathcal{V} \otimes \mathcal{V}$ induced by D.

For the specific case $D = D_{\text{KL}}$, the inner divergence relates closely to the Shannon entropy of p_h. The average categorical entropy $\bar{\mathcal{H}}_{D_{\text{KL}}}(k_{\text{head}}; p_{H_t})$ provides a measure akin to the average conditional Shannon entropy:

$$\mathbb{E}_{h\sim p_{H_t}}[H(k_{\text{head}}(h, \cdot))] = \mathbb{E}_{h\sim p_{H_t}}\left[-\sum_{w\in\mathcal{V}} k_{\text{head}}(h, \{w\}) \log k_{\text{head}}(h, \{w\})\right]. \tag{19}$$

While not identical, both measures capture the average uncertainty in the next-token prediction given the hidden state. As discussed in Section 5, minimizing NLL implicitly drives the model to match this intrinsic stochasticity present in the data.

Estimation. Estimating the average categorical entropy (Equation (18)) involves averaging over samples $h_t \sim p_{H_t}$. For each sampled h_t:

1. Compute the output probability vector $p_{h_t} = [k_{\text{head}}(h_t, \{w\})]_{w \in \mathcal{V}}$.

2. Construct the two required probability measures on the finite space $\mathcal{V} \times \mathcal{V}$: $\mu_1 = \sum_w p_{h_t}(w)\delta_{(w,w)}$ and $\mu_2 = p_{h_t} \otimes p_{h_t}$.

3. Compute the divergence $D_{\mathcal{V}\otimes\mathcal{V}}(\mu_1 \| \mu_2)$. Since the space is finite, this calculation is often straightforward (e.g., for KL divergence: $\sum_{(w_1,w_2)} \mu_1(w_1, w_2) \log(\mu_1(w_1, w_2)/\mu_2(w_1, w_2))$).

4. Average these divergence values over many samples of h_t obtained by sampling contexts $\mathbf{w}_{<t} \sim P_{\text{ctx}}$ and applying $k_{\text{emb}}, k_{\text{bb}}$. Estimating the average Shannon entropy (Equation (19)) follows a similar Monte Carlo approach, calculating $H(p_{h_t})$ in step 3.

4.4 Metric 4: Information Flow Bounds via Data Processing Inequality

To leverage the fundamental Data Processing Inequality (DPI), inherent in the Markov category **Stoch** and satisfied by I_D, to establish bounds on how much information about a context property S can propagate through the processing chain to the final output token W_t.

Let S be a property of the context $\mathbf{w}_{<t}$ as in Metric 1. The sequence of transformations $S \to \mathbf{w}_{<t} \to E_{<t} \to H_t \to W_t$ forms a Markov chain, provided we consider the joint distribution $P(s, \mathbf{w}_{<t}, e_{<t}, h_t, w_t)$ induced by the process. Since k_{emb} and k_{bb} are deterministic functions of $\mathbf{w}_{<t}$, H_t is a function of $\mathbf{w}_{<t}$. Furthermore, k_{head} generates W_t based only on H_t. Thus, we have the Markov chain structure $S \to H_t \to W_t$. This means the conditional distribution of W_t given H_t and S depends only on H_t: $P(W_t|H_t, S) = P(W_t|H_t)$.

Consider the joint distribution of (S, H_t), represented by the state $p_{S,H_t} : I \to S \otimes (\mathcal{H}, \mathscr{B}(\mathcal{H}))$ (assuming S takes values in a measurable space, also denoted S). Similarly, let $p_{S,W_t} : I \to S \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ be the joint state of (S, W_t). Crucially, the state p_{S,W_t} can be obtained from p_{S,H_t} by applying the kernel $\text{id}_S \otimes k_{\text{head}} : S \otimes (\mathcal{H}, \mathscr{B}(\mathcal{H})) \to S \otimes (\mathcal{V}, \mathcal{P}(\mathcal{V}))$, which acts as k_{head} on the $\mathcal{H}$ component while leaving the S component unchanged:

$$p_{S,W_t} = (\text{id}_S \otimes k_{\text{head}}) \circ p_{S,H_t}. \tag{20}$$

Theorem 4.5 (Categorical Information Flow Bound). *Let $I_D(S; X) := D_{S\otimes X}(p_{S,X} \| p_S \otimes p_X)$ denote the categorical mutual information between S and $X \in \{H_t, W_t\}$, where $p_{S,X}$ is the joint state and p_S, p_X are the corresponding marginal states. The Data Processing Inequality for the divergence D, when applied to the definition of I_D and the Markov kernel $\text{id}_S \otimes k_{\text{head}}$ processing p_{S,H_t} to p_{S,W_t}, implies:*

$$I_D(S; H_t) \geq I_D(S; W_t). \tag{21}$$

Proof sketch: The DPI states that for any kernel $k : A \to B$ and states $p, q : I \to A$, we have $D_B(k \circ p \| k \circ q) \leq D_A(p\|q)$. Apply this with $A = S \otimes \mathcal{H}$, $B = S \otimes \mathcal{V}$, $k = \text{id}_S \otimes k_{\text{head}}$, $p = p_{S,H_t}$, and $q =$

$p_S \otimes p_{H_t}$. Then $k \circ p = p_{S,W_t}$ and $k \circ q = (\mathrm{id}_S \otimes k_{\mathrm{head}}) \circ (p_S \otimes p_{H_t}) = p_S \otimes (k_{\mathrm{head}} \circ p_{H_t}) = p_S \otimes p_{W_t}$. The inequality $D_{S \otimes \mathcal{V}}(p_{S,W_t} \| p_S \otimes p_{W_t}) \leq D_{S \otimes \mathcal{H}}(p_{S,H_t} \| p_S \otimes p_{H_t})$ directly yields $I_D(S; W_t) \leq I_D(S; H_t)$.

Interpretation. This fundamental inequality asserts that the amount of statistical information (measured by I_D) that the next token W_t carries about the context property S cannot exceed the amount of information about S that is already encoded in the intermediate hidden representation H_t. The final stochastic step $k_{\mathrm{head}} : H_t \to W_t$ can only preserve or lose information about S; it cannot create it. The difference $I_D(S; H_t) - I_D(S; W_t) \geq 0$ quantifies the information about S that is present in the representation H_t but is "lost" or not utilized in the immediate prediction of W_t. This loss could be due to the inherent stochasticity of k_{head} (as measured by $\mathcal{H}_D(k_{\mathrm{head}})$) or because the mapping discards aspects of H_t relevant to S but not relevant for predicting W_t. This unused information might still be crucial for predicting subsequent tokens ($W_{t+1}, \ldots$).

Estimation Challenges. Requires estimating two mutual information quantities:

- $I_D(S; W_t)$: If S is discrete or low-dimensional, this involves MI between S and the discrete variable W_t. This is often the more tractable quantity to estimate, potentially using direct frequency counts or simple estimators if S has few categories.
- $I_D(S; H_t)$: This involves MI between the context property S and the high-dimensional continuous hidden state H_t. This estimation faces the same challenges as $I_D(H_t; W_t)$ and $I_D(H_t; H_{t+1})$, requiring robust high-dimensional MI estimators (kNN or variational) sensitive to the specific structure of S (discrete vs. continuous).

Estimation relies on obtaining samples of (s, h_t) and (s, w_t) pairs. This is done by sampling contexts $\mathbf{w}_{<t}$ associated with property value s (i.e., from $P_{\mathrm{ctx}}(\cdot|s)$), running the AR generation step ($k_{\mathrm{emb}}, k_{\mathrm{bb}}, k_{\mathrm{head}}$) to get h_t and w_t, and collecting these samples for input into the chosen MI estimators. Comparing the estimated values provides an empirical check on the information flow bottleneck at the LM head stage.

5 Pretraining Objective, Compression, and Learning Intrinsic Stochasticity

A central question surrounding large language models is why the seemingly simple auto-regressive 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 $k_{\text{data}} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ 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 $\mathcal{V}$. Let $p_{W_{<t}}$ denote the marginal probability measure on the context space $(\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*))$, 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:

$$L_{\text{CE}}(\theta) = -\mathbb{E}_{(\mathbf{w}_{<t}, w_t) \sim P_{\text{data}}}[\log P_\theta(w_t | \mathbf{w}_{<t})] \tag{22}$$

where $P_\theta(w_t|\mathbf{w}_{<t})$ is the model's predicted probability. Let $k_{\text{gen},\theta} : (\mathcal{V}^*, \mathscr{B}(\mathcal{V}^*)) \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ 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)**.** *Minimizing the cross-entropy loss $L_{\text{CE}}(\theta)$ (Equation (22)) with respect to parameters θ is equivalent to minimizing the average KL divergence between the true conditional data distribution and the model's conditional distribution, averaged over the context distribution $p_{W_{<t}}$ derived from P_{data}:*

$$\arg\min_\theta L_{\text{CE}}(\theta) = \arg\min_\theta \mathcal{L}_{\text{KL}}(\theta) := \arg\min_\theta \mathbb{E}_{\mathbf{w}_{<t} \sim p_{W_{<t}}}[D_{\text{KL}}(k_{data}(\mathbf{w}_{<t}, \cdot) \,\|\, k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot))] \tag{23}$$

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., $k_{data} = k_{\text{gen},\theta_{true}}$ for some $\theta_{true} \in \Theta$), then the minimum value is 0, achieved if and only if $k_{\text{gen},\theta^*}(\mathbf{w}_{<t}, \cdot) = k_{data}(\mathbf{w}_{<t}, \cdot)$ 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. (Full proof in Appendix A.1). □

This theorem formally states that the NLL training objective directly drives the model's conditional predictions $k_{\text{gen},\theta}(\mathbf{w}_{<t}, \cdot)$ to match the true data conditionals $k_{\text{data}}(\mathbf{w}_{<t}, \cdot)$ in the sense of average KL divergence. This perspective is fundamental to understanding language modeling as density estimation [4]. 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 (18)). 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:

$$\bar{\mathcal{H}}_D(k; p_X) \coloneqq \mathbb{E}_{x\sim p_X}\left[D_{Y\otimes Y}\left(\sum_{y\in Y} k(x,\{y\})\delta_{(y,y)} \quad \| \quad k(x,\cdot)\otimes k(x,\cdot)\right)\right]. \tag{24}$$

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 $k_{\text{head},\theta} : \mathcal{H} \to (\mathcal{V}, \mathcal{P}(\mathcal{V}))$ 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 statistical divergence (e.g., an f-divergence with f differentiable at 1, $f(1) = 0$, or d_{TV}) defined on $\mathcal{P}(\mathcal{V}\times\mathcal{V})$. Assume a sequence of model parameters θ_n achieves convergence in the training objective, such that the average KL divergence $\mathcal{L}_{\text{KL}}(\theta_n) \to \inf_\theta \mathcal{L}_{\text{KL}}(\theta)$ as $n \to \infty$. Let $k_n = k_{\text{gen},\theta_n}$ be the corresponding sequence of model kernels, and $k_{head,n} = k_{head,\theta_n}$ be the LM head kernels. Assume this convergence implies:*

- *(i) Pointwise convergence of the head kernel's output distributions: $k_{head,n}(h,\cdot) \to k_{head,\theta^*}(h,\cdot)$ in a suitable topology on $\mathcal{P}(\mathcal{V})$ (e.g., in total variation, which is strong for finite $\mathcal{V}$) for p_{H_t,θ^*}-almost every h, where θ^* minimizes $\mathcal{L}_{\text{KL}}(\theta)$.*
- *(ii) Weak convergence of the induced hidden state distributions: $p_{H_t,\theta_n} \Rightarrow p_{H_t,\theta^*}$.*
- *(iii) The function $\Psi_D(h,p) \coloneqq D_{\mathcal{V}\otimes\mathcal{V}}(\sum_w p(w)\delta_{(w,w)} \| p \otimes p)$ is continuous and bounded as a function of $p = k_{head}(h,\cdot)$ for h in the support of p_{H_t,θ^*}, with respect to the convergence topology in (i). (This holds for standard divergences like KL and TV on finite $\mathcal{V}$).*

Then, the average categorical entropy of the learned LM head converges to that of the optimal LM head:

$$\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{head,n}; p_{H_t,\theta_n}) = \bar{\mathcal{H}}_D(k_{head,\theta^*}; p_{H_t,\theta^*}). \tag{25}$$

Furthermore, if the model class is sufficiently expressive such that the optimal model kernel k_{gen,θ^*} matches the data kernel k_{data} (i.e., $\mathcal{L}_{\text{KL}}(\theta^*) = 0$), then the learned average categorical entropy approximates that of the implicit final stochastic step of the true data generating process. Assuming k_{data} admits a similar factorization with a final stochastic kernel $k_{head,\ data}$ acting on some "true" state h_{data}, then:

$$\bar{\mathcal{H}}_D(k_{head,\theta^*}; p_{H_t,\theta^*}) \approx \bar{\mathcal{H}}_D(k_{head,\ data}; p_{H_t,data}). \tag{26}$$

Proof Sketch. The average categorical entropy is an expectation:

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

The assumptions ensure that the random variable inside the expectation converges in distribution. Specifically, weak convergence of p_{H_t,θ_n} (ii) combined with the pointwise convergence of $k_{\text{head},n}(h,\cdot)$ (i) and the continuity of Ψ_D (iii) allow us to apply variants of the continuous mapping theorem or dominated convergence theorem (leveraging the assumed boundedness in (iii) and the fact that probability measures are bounded). This yields the convergence of the expectation to $\mathbb{E}_{h\sim p_{H_t,\theta^*}}[\Psi_D(h, k_{\text{head},\theta^*}(h,\cdot))] = \bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*})$. The final approximation holds if $k_{\text{gen},\theta^*} = k_{\text{data}}$, which implies $k_{\text{head},\theta^*} \approx k_{\text{head, data}}$ and $p_{H_t,\theta^*} \approx p_{H_t,\text{data}}$ under reasonable assumptions about the factorization. (Full proof in Appendix A.2). □

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 $(\mathbf{Stoch}, D)$ enriched with a divergence like D_{KL}, provides a natural bridge to Information Geometry [1, 17]. 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 $\mathcal{P}(\mathcal{V})$.

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

$$g_{ij}^{\text{FR}}(\xi) = \sum_{w\in\mathcal{V}} p_\xi(w) \frac{\partial \log p_\xi(w)}{\partial \xi_i} \frac{\partial \log p_\xi(w)}{\partial \xi_j} = \mathbb{E}_{W\sim p_\xi}\left[\frac{\partial \log p_\xi(W)}{\partial \xi_i} \frac{\partial \log p_\xi(W)}{\partial \xi_j}\right]. \tag{27}$$

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 $\mathcal{P}(\mathcal{V})$ also includes dual affine connections ($\pm\alpha$-connections) related to the KL divergence, providing a richer dually flat structure [1].

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

Specifically, the Fisher-Rao metric g^{FR} on $\mathcal{P}(\mathcal{V})$ 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:

$$g_{ab}^*(h) = \sum_{i,j} g_{ij}^{\mathrm{FR}}(g_{\mathrm{head}}(h)) \frac{\partial (g_{\mathrm{head}}(h))_i}{\partial h_a} \frac{\partial (g_{\mathrm{head}}(h))_j}{\partial h_b}, \quad a, b \in \{1, \ldots, d_{\mathrm{model}}\}, \tag{28}$$

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 $p_h \in \mathcal{P}(\mathcal{V})$ (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 ($|\mathcal{V}|-1 \times d_{\mathrm{model}}$ or $|\mathcal{V}| \times d_{\mathrm{model}}$ 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 $g_{head} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ 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 $p_h = g_{head}(h)$ and $p_{h+\epsilon v} = g_{head}(h + \epsilon v)$ 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 $g^*(h)$:*

$$D_{\mathrm{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 g^*(h)(v, v) + O(\epsilon^3) \tag{29}$$

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 $\mathcal{P}(\mathcal{V})$ around p_h. The KL divergence between two nearby distributions p_ξ and $p_{\xi'}$ can be expanded around p_ξ as [1]:

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

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

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

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:

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

The term $J(h)^\top g^{\mathrm{FR}}(\xi(h))J(h)$ 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 (28). Thus, $D_{\mathrm{KL}}(p_{h+\epsilon v} \,\|\, p_h) = \frac{1}{2}\epsilon^2 g^*(h)(v,v) + O(\epsilon^3)$. 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 $\mathcal{P}(\mathcal{V})$.

A key property of the pullback metric is its potential degeneracy, arising from the dimensionality difference between $\mathcal{H}$ and $\mathcal{P}(\mathcal{V})$.

Remark 6.2 (Connection to Score Function). *The Fisher-Rao metric g^{FR} is intrinsically linked to the score function $\nabla_\xi \log p_\xi(w)$. The pullback metric g^* can also be expressed directly in terms of the score function with respect to the hidden state h. Let $p_h(w) = k_{\text{head}}(h, \{w\})$. The score vector for token w is $\nabla_h \log p_h(w) \in \mathcal{H}$. The pullback metric tensor components are:*

$$g^*_{ab}(h) = \sum_{w \in \mathcal{V}} p_h(w)(\nabla_{h_a} \log p_h(w))(\nabla_{h_b} \log p_h(w))^\top = \mathbb{E}_{W \sim p_h}[(\nabla_{h_a} \log p_h(W))(\nabla_{h_b} \log p_h(W))^\top]. \tag{30}$$

This shows $g^*(h)$ is the expected outer product of the score function $\nabla_h \log p_h(W)$, directly connecting the information geometry to the sensitivity of log-probabilities to changes in the representation h. This score vector $\nabla_h \log p_h(W)$ is analogous to the score used in score-based generative models, but here taken with respect to the conditioning variable h.

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

$$\operatorname{rank}(g^*(h)) \le \min(d_{\text{model}}, |\mathcal{V}| - 1). \tag{31}$$

Proof. The pullback metric $g^*(h)$ is defined as $g^*(h) = J(h)^\top g^{\mathrm{FR}}(g_{\text{head}}(h))J(h)$, where $J(h)$ is the Jacobian of the map $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ (represented in appropriate local coordinates). The dimension of $\mathcal{H}$ is d_{model}, and the dimension of $\mathcal{P}(\mathcal{V})$ is $d_{\text{prob}} = |\mathcal{V}| - 1$. 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}). The rank of a product of matrices satisfies $\operatorname{rank}(ABC) \le \min(\operatorname{rank}(A), \operatorname{rank}(B), \operatorname{rank}(C))$. Therefore,

$$\operatorname{rank}(g^*(h)) = \operatorname{rank}(J(h)^\top g^{\mathrm{FR}}(g_{\text{head}}(h))J(h)) \le \min(\operatorname{rank}(J(h)^\top), \operatorname{rank}(g^{\mathrm{FR}}), \operatorname{rank}(J(h))).$$

Since $\operatorname{rank}(J(h)) \le \min(d_{\text{prob}}, d_{\text{model}})$ and $\operatorname{rank}(g^{\mathrm{FR}}) = d_{\text{prob}}$, we have:

$$\operatorname{rank}(g^*(h)) \le \min(d_{\text{model}}, |\mathcal{V}| - 1).$$

□

Remark 6.4 (Typical Rank and Degeneracy). *In modern large language models, it is typically the case that the model dimension is much smaller than the vocabulary size, $d_{\text{model}} \ll |\mathcal{V}|$. For instance, d_{model}*

might be 8192 while $|\mathcal{V}|$ is 65,536. In this common scenario, the rank bound becomes $\mathrm{rank}(g^*(h)) \leq d_{\mathrm{model}}$. *If the Jacobian $J(h)$ of the map g_{head} (e.g.,* $\mathrm{softmax}(Wh)$*) has full rank d_{model}, which is plausible if the weight matrix W has rank d_{model} and the softmax function does not induce rank collapse, then the pullback metric $g^*(h) = J(h)^\top g^{FR} J(h)$ will also have rank d_{model}. Since $g^*(h)$ is a $d_{\mathrm{model}} \times d_{\mathrm{model}}$ matrix, having rank d_{model} means it is non-degenerate (i.e., positive definite) as a metric on the representation space $\mathcal{H}$. Therefore, contrary to scenarios where the latent space dimension might exceed the output space dimension, in typical LLMs, the pullback metric g^* is expected to be full rank and non-degenerate. The geometry it induces is still highly relevant, but the interpretation should focus on anisotropy (different sensitivities in different directions) rather than degeneracy.*

6.1 Interpretation and Implications

The information-geometric perspective, formalized by the pullback metric g^* and its properties, provides several insights:

- **Sensitivity Analysis:** The quadratic form associated with $g^*(h)$, $g^*(h)(v, v)$, directly quantifies the local distinguishability (via KL divergence, Equation (29)) between the output distributions p_h and $p_{h+\epsilon v}$. It measures the sensitivity of the model's prediction p_h to perturbations of the hidden state h in the direction v, viewed through the geometric lens of $\mathcal{P}(\mathcal{V})$. Directions v with large $g^*(h)(v, v)$ correspond to changes in h that significantly alter the output distribution's geometry according to the Fisher-Rao metric.
- **Rank and Anisotropy (Proposition 6.3, Remark 6.4):** Typically, $d_{\mathrm{model}} \ll |\mathcal{V}|$, and the pullback metric $g^*(h)$ is expected to have full rank d_{model} and be non-degenerate. This means that infinitesimal changes to the hidden state h in *any* direction $v \neq 0$ will induce a non-zero local change in the output distribution p_h (measured by KL divergence, Equation (29)). However, the metric is generally anisotropic: the magnitude of $g^*(h)(v, v)$ will vary depending on the direction v. This reflects that the model's predictions are more sensitive to changes in h along certain directions than others. There is no significant "null space" of directions irrelevant to prediction in this typical scenario.
- **Spectrum:** The eigenvalues and eigenvectors of $g^*(h)$ (restricted to its support, the orthogonal complement of the null space) reveal the principal directions of sensitivity in the representation space $\mathcal{H}$ with respect to the prediction task. Directions corresponding to large eigenvalues are those where small changes in h induce large changes (geometrically measured by g^{FR}, corresponding to large KL divergence) in the predicted distribution p_h.

This geometric perspective provides a rigorous way to analyze the functional geometry of the representation space $\mathcal{H}$ as shaped by the downstream prediction task defined by k_{head}. Investigating how g^* varies across regions of $\mathcal{H}$ populated by the hidden state distribution p_{H_t} (Section 4) could reveal how the model allocates representational sensitivity. For example, we could compute the average metric $\bar{g}^* = \mathbb{E}_{h \sim p_{H_t}}[g^*(h)]$ or study the properties of the manifold $(\mathcal{H}, g^*)$ near typical representations h. The spectrum of $\bar{g}^*$ would indicate the average principal directions of predictive sensitivity.

Furthermore, the pullback metric g^* connects directly to the Representation Divergence metric (Equation (12)). 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 $\mathcal{P}(\mathcal{V})$ via g_{head}, the distance between these regions in $\mathcal{P}(\mathcal{V})$ (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 that underlies the global Representation Divergence. 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 $\mathcal{P}(\mathcal{V})$, implicitly structuring the manifold $(\mathcal{H}, g^*)$.

7 Implicit Spectral Structuring via NLL Optimization

A central question in the study of large language models is why the objective of minimizing the negative log-likelihood (NLL) of the next token (Equation (23)) induces internal representations $h_t \in \mathcal{H}$ that appear to capture rich semantic, syntactic, and contextual information. While auto-regressive models lack the explicit positive/negative pairing structure typical of contrastive learning methods, we argue that the NLL objective itself implicitly imposes significant structural constraints on the learned representations. Specifically, this constraint encourages the geometry of the representation space $\mathcal{H}$, particularly as modulated by the prediction head (Section 6), to align with the underlying predictive structure inherent in the data $P_{\text{data}}(\cdot|x)$. This perspective shares conceptual similarities with spectral methods in self-supervised representation learning [9, 21].

Let $f_{\text{enc}} : \mathcal{V}^* \to \mathcal{H}$ 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 $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ 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 $\mu_{ctx} = p_{W_{<t}}$:

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim \mu_{ctx}}[D_{\text{KL}}(P_{\text{data}}(\cdot|x) \,\|\, g_{\text{head}}(f_{\text{enc}}(x)))] \tag{32}$$

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 (32)) 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 $\mathcal{L}(\theta) = \mathbb{E}_{x \sim \mu_{ctx}}[D_{\text{KL}}(P_{data}(\cdot|x) \| p_\theta(\cdot|x))]$, where $p_\theta(\cdot|x) = g_{head}(f_{enc}(x))$. Let d_{out} be a metric on $\mathcal{P}(\mathcal{V})$ satisfying a Pinsker-type inequality, such as $d_{out}(p,q)^k \leq C \cdot D_{\text{KL}}(p\|q)$ 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:*

$$\mathbb{E}_{x \sim \mu_{ctx}}[d_{out}(P_{data}(\cdot|x), p_\theta(\cdot|x))^k] \leq C \cdot \mathcal{L}(\theta). \tag{33}$$

Consequently, by the triangle inequality for the metric d_{out}, for any pair of contexts (x, x'), we have:

$$\begin{aligned} d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x')) \leq d_{out}(P_{data}(\cdot|x), p_\theta(\cdot|x)) + d_{out}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \\ + d_{out}(p_\theta(\cdot|x'), P_{data}(\cdot|x')). \end{aligned} \tag{34}$$

If the model fits the data well, such that $\mathcal{L}(\theta)$ is small, the first and third terms on the right-hand side (RHS) are small on average (by Equation (33)) and thus typically small for individual contexts x, x'. Therefore, the distance between the model's output distributions, $d_{out}(p_\theta(\cdot|x), p_\theta(\cdot|x'))$, must necessarily approximate the distance between the true target distributions, $d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$. Specifically, contexts with predictively dissimilar target distributions (large LHS) must yield model output distributions that are also separated (large $d_{out}(p_\theta(\cdot|x), p_\theta(\cdot|x'))$).

Proof Sketch. The inequality Equation (33) arises directly from the assumed Pinsker-type inequality $d_{\text{out}}(p, q)^k \leq C \cdot D_{\mathrm{KL}}(p\|q)$. Letting $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$, we have $d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))^k \leq C \cdot D_{\mathrm{KL}}(P_{\text{data}}(\cdot|x)\|p_\theta(\cdot|x))$. Taking the expectation over $x \sim \mu_{ctx}$ on both sides yields the result. The triangle inequality (Equation (34)) is a standard property of any metric d_{out}. Let $\epsilon_x = d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))$ and $\epsilon_{x'} = d_{\text{out}}(p_\theta(\cdot|x'), P_{\text{data}}(\cdot|x'))$. If $\mathcal{L}(\theta)$ is small, then $\mathbb{E}_x[\epsilon_x^k]$ is small. By Markov's inequality, $\mathbb{P}(\epsilon_x \geq \delta) \leq \mathbb{E}[\epsilon_x^k]/\delta^k$, implying that ϵ_x is small with high probability for typical x. Therefore, for typical pairs (x, x'), both ϵ_x and $\epsilon_{x'}$ are small. Rearranging the triangle inequality gives $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \geq d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')) - \epsilon_x - \epsilon_{x'}$. This shows that $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x'))$ lower bounds the target distance $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ up to small error terms. Similarly, $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \leq d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')) + \epsilon_x + \epsilon_{x'}$, showing the distance between model outputs approximates the distance between target distributions when errors are small. (Full proof in Appendix A.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}.

Before stating the corollary, we formally define the sensitivity of the head mapping. The information geometry perspective (Section 6) provides the necessary tools. Recall the pullback Fisher-Rao metric $g^*(h)$ on $\mathcal{H}$ induced by the head mapping $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$.

Definition 7.2 (Sensitive Directions and Head Sensitivity). *Let $h \in \mathcal{H}$ and $v \in T_h\mathcal{H} \cong \mathcal{H}$.*

1. *A direction v is* ***sensitive for*** g_{head} ***at*** h *if $g^*(h)(v,v) > 0$. Equivalently, v is sensitive if it is not in the null space of the pullback metric $g^*(h)$, meaning an infinitesimal displacement of h along v induces a non-zero change in the output distribution $p_h = g_{head}(h)$ as measured locally by KL divergence (Equation (29)). The subspace spanned by sensitive directions at h is the support of $g^*(h)$.*

2. *The head mapping g_{head} is* ***sufficiently sensitive in a region*** $R \subseteq \mathcal{H}$ *if for any two distinct representations $h_1, h_2 \in R$ such that $g_{head}(h_1) \neq g_{head}(h_2)$, the difference vector $(h_1 - h_2)$ must have a non-zero component along a sensitive direction v (as defined in item 1, evaluated e.g., at h_1 or h_2 or along the path between them). More practically for the corollary's assumption: if $d_{out}(g_{head}(h_1), g_{head}(h_2)) > \delta > 0$, then $(h_1 - h_2)$ must have a significant projection onto the subspace spanned by sensitive directions.*

Intuitively, sufficient sensitivity means that if the head mapping produces noticeably different outputs, the inputs must differ in ways that the head mapping can detect, as measured by the induced information geometry.

Corollary 7.3 (Implicit Representation Separation)**.** *Suppose that:*

(i) The model parameters θ achieve a sufficiently small average NLL loss $\mathcal{L}(\theta)$ such that, for typical contexts x, x', the error terms $\epsilon_x = d_{out}(P_{data}(\cdot|x), p_\theta(\cdot|x))$ and $\epsilon_{x'} = d_{out}(p_\theta(\cdot|x'), P_{data}(\cdot|x'))$ are negligible compared to the distance $d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$. (This follows from Theorem 7.1.)

(ii) The head mapping $g_{head} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ exhibits sufficient local sensitivity in the region of $\mathcal{H}$ populated by representations $\{h_x = f_{enc}(x) \mid x \sim \mu_{ctx}\}$. This means that if two representations $h_x, h_{x'}$ within this region produce output distributions separated by a non-negligible distance $d_{out}(g_{head}(h_x), g_{head}(h_{x'})) > \delta > 0$, then h_x and $h_{x'}$ must differ along directions relevant to the head mapping (i.e., $h_x - h_{x'}$ must have non-zero components along ***sensitive directions for*** g_{head} *according to Definition 7.2).*

Then, if two contexts x, x' have predictively dissimilar target distributions, meaning $d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$ is significantly greater than zero, their learned representations $h_x = f_{enc}(x)$ and $h_{x'} = f_{enc}(x')$ must also differ in $\mathcal{H}$ along dimensions relevant to the head mapping:

$$d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x')) \text{ large} \implies h_x \text{ and } h_{x'} \text{ differ along sensitive directions for } g_{head}. \tag{35}$$

Conversely, if contexts x, x' are predictively similar ($d_{out}(P_{data}(\cdot|x), P_{data}(\cdot|x'))$ is small), the NLL objective does not strongly constrain the distance between h_x and $h_{x'}$ along relevant dimensions; they might be mapped closely together without incurring significant loss penalty.

Proof Sketch. From Theorem 7.1, under assumption (i), we have $d_{\text{out}}(p_\theta(\cdot|x), p_\theta(\cdot|x')) \approx d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$. If the target distributions are dissimilar, $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, implying $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be large (significantly greater than δ from assumption (ii)). Assumption (ii), using Definition 7.2, states that achieving this separation in the output space requires that the inputs h_x and $h_{x'}$ must differ along sensitive directions for g_{head}. Therefore, large predictive dissimilarity forces separation in the representation space along these sensitive dimensions. The sensitivity assumption (ii) is crucial; it ensures that the necessity of separating the outputs $p_\theta(\cdot|x)$ and $p_\theta(\cdot|x')$ (forced by the NLL objective via Theorem 7.1) translates into a requirement for separation of the corresponding inputs h_x and $h_{x'}$ along directions detectable by g_{head} (i.e., those where $g^*(h)(v,v) > 0$). Note that h_x and $h_{x'}$ could still be close or identical along directions that are not sensitive (i.e., in the null space of $g^*(h)$). (Full proof in Appendix A.4.) □

As discussed in Remark 6.4, g^* is likely non-degenerate, so the "null space" is trivial. Assumption (ii) essentially requires g_{head} to be locally injective or at least structured such that significantly different outputs require significantly different inputs. The conclusion remains that large predictive dissimilarity forces separation of representations $h_x, h_{x'}$ along directions v where $g^*(h)(v, v)$ is large enough to effect the required output separation.

This corollary establishes that NLL minimization implicitly forces representations $h_x, h_{x'}$ apart along relevant dimensions if their corresponding contexts x, x' are predictively dissimilar, provided the LM head g_{head} is sufficiently sensitive. This differential pressure based on predictive similarity/dissimilarity forms the basis for connections to spectral methods.

7.3 Predictive Similarity Kernels

The preceding analysis (Theorem 7.1, Corollary 7.3) suggests that the NLL objective implicitly shapes the geometry of the representation space $\mathcal{H}$ based on the *dissimilarity* between the true next-token distributions $P_{\text{data}}(\cdot|x)$ for different contexts x, x'. To facilitate connections with methods that operate on similarity structures, such as spectral graph methods, it is useful to formalize the complementary notion of *predictive similarity*. We can define kernels that quantify the degree to which two contexts share similar predictive futures, effectively measuring the inverse of predictive dissimilarity.

Definition 7.4 (Predictive Similarity Kernel). *Let $p_x := P_{data}(\cdot|x)$ denote the true conditional distribution for context x. A predictive similarity kernel is a function $K : \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}_{\geq 0}$ quantifying the similarity between p_x and $p_{x'}$. Potential examples include:*

- ***Bhattacharyya Coefficient Kernel**: $K_{BC}(x, x') := BC(p_x, p_{x'}) = \sum_{w \in \mathcal{V}} \sqrt{p_x(w) p_{x'}(w)}$. This measures the cosine similarity between the square-root vectors $(\sqrt{p_x(w)})_w$. It relates directly to the Hellinger distance $d_H^2(p_x, p_{x'}) = 2(1 - K_{BC}(x, x'))$ and defines a positive semidefinite kernel. High K_{BC} corresponds to low d_H.*
- ***Hellinger-based Kernel (Gaussian Kernel on $\sqrt{p}$)**: $K_H(x, x') := \exp(-\beta d_H^2(p_x, p_{x'}))$ 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)**: $K_{Lin}(x, x') := \langle p_x, p_{x'} \rangle = \sum_{w \in \mathcal{V}} p_x(w) p_{x'}(w)$. 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**: $K_{KL}(x, x') := \exp(-\beta S_{KL}(p_x, p_{x'}))$ 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 $d_{out}(p_x, p_{x'})$). It is this similarity structure, derived from the data's conditional probabilities, that we hypothesize the NLL objective implicitly captures within the geometry of $\mathcal{H}$.

7.4 Connection to Graph Laplacian and Dirichlet Energy Minimization

Consider an undirected graph where contexts $x \in \mathcal{V}^*$ are nodes distributed according to μ_{ctx}, and edge weights are given by a symmetric predictive similarity kernel $K(x, x')$ (e.g., K_{BC}, K_{H} from Def. 7.4). The graph Laplacian operator associated with this kernel captures the structure of this similarity graph.

Definition 7.5 (Graph Laplacian Operator)**.** *Let $K : \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}_{\geq 0}$ be a symmetric predictive similarity kernel and let μ_{ctx} be the measure on $\mathcal{V}^*$. Define the degree function $d(x) = \int K(x, x')\, \mu_{ctx}(\mathrm{d}x')$. The (unnormalized) graph Laplacian Δ_K is an operator acting on functions $\phi \in L^2(\mathcal{V}^*, \mu_{ctx})$ defined by:*

$$(\Delta_K \phi)(x) = d(x)\phi(x) - \int K(x, x')\phi(x')\, \mu_{ctx}(\mathrm{d}x'). \tag{36}$$

A key property is that the quadratic form of the Laplacian corresponds to the Dirichlet energy, measuring function smoothness over the graph.

$$\mathcal{E}_K(\phi) \coloneqq \frac{1}{2} \iint K(x, x')(\phi(x) - \phi(x'))^2\, \mu_{ctx}(\mathrm{d}x)\mu_{ctx}(\mathrm{d}x') = \langle \phi, \Delta_K \phi \rangle_{L^2(\mu_{ctx})}. \tag{37}$$

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 7.3, exerts a related pressure.

Proposition 7.6 (NLL Objective and Implicit Dirichlet Energy Minimization)**.** *Let $h_x = f_{enc}(x)$ be the representation mapping and assume the conditions of Corollary 7.3 hold, including the sufficient sensitivity of g_{head} (Definition 7.2). Let v be a direction in $\mathcal{H}$ that is sensitive for g_{head} (i.e., $g^*(h)(v, v) > 0$ in the relevant region). Let $\phi_v(x) = \langle h_x, v \rangle$ be the projection of the representation onto this sensitive direction. Minimizing the NLL loss $\mathcal{L}(\theta)$ encourages configurations where the representations $h_x, h_{x'}$ are close along such sensitive directions v when the predictive similarity $K(x, x')$ is high. This implicitly favors representations h_x such that the projected components $\phi_v(x)$ tend to have lower Dirichlet energy $\mathcal{E}_K(\phi_v)$ with respect to the predictive similarity kernel $K(x, x')$:*

$$\mathcal{L}(\theta) \textit{ small} \implies \mathcal{E}_K(\phi_v) \textit{ tends to be small for sensitive directions } v. \tag{38}$$

Proof Sketch. Corollary 7.3 implies that when $\mathcal{L}(\theta)$ is small, if $K(x, x')$ is large (meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small), then the NLL objective encourages h_x and $h_{x'}$ to be close along sensitive directions v (as defined in Def. 7.2) to ensure $p_\theta(\cdot|x) \approx p_\theta(\cdot|x')$. That is, $\phi_v(x) = \langle h_x, v \rangle$ should be close to $\phi_v(x') = \langle h_{x'}, v \rangle$ when $K(x, x')$ is large, for sensitive v. (Full proof in Appendix A.5.) □

This proposition formalizes the link: NLL minimization pushes representations towards configurations favored by spectral clustering on the predictive similarity graph, specifically along dimensions relevant for the prediction head.

7.5 Connection to Predictive Similarity Operator

An alternative viewpoint, closer to the analysis in [21], considers an operator derived from the predictive similarity kernel acting on the representation space.

Definition 7.7 (Predictive Similarity Operator). *Let $f_{enc} : \mathcal{V}^* \to \mathcal{H}$ be a fixed encoder, inducing a distribution $\mu = p_{H_t}$ on $\mathcal{H}$ via μ_{ctx}. Let $K(x, x')$ be the predictive similarity kernel. The predictive similarity operator $M_K : L^2(\mathcal{H}, \mu) \to L^2(\mathcal{H}, \mu)$ is defined via its action on functions $\psi : \mathcal{H} \to \mathbb{R}$:*

$$(M_K\psi)(h_x) \triangleq \mathbb{E}_{x' \sim \mu_{ctx}}[K(x, x')\psi(h_{x'})] = \int_{\mathcal{V}^*} K(x, x')\psi(f_{enc}(x'))\, \mu_{ctx}(\mathrm{d}x'), \tag{39}$$

where $h_x = f_{enc}(x)$. This operator averages the function ψ over representations $h_{x'}$ weighted by the predictive similarity $K(x, x')$ between their originating contexts x, x' and the reference context x.

If K is symmetric and induces a suitable integral kernel on $\mathcal{H}$, M_K is typically a compact self-adjoint operator with a discrete spectrum. Its eigenfunctions capture dominant patterns of predictive similarity as reflected in the representation space.

Hypothesis 7.8 (NLL Objective and Alignment with Operator Eigenspace). *Let f_{enc}, g_{head}, K, and M_K be as defined above. Assume the conditions of Corollary 7.3 hold. Let $\{\phi_i\}$ be the eigenfunctions of M_K with eigenvalues $\{\lambda_i\}$. The NLL objective $\mathcal{L}(\theta)$, by implicitly encouraging representations $h_x, h_{x'}$ to be close along directions v where the pullback metric $g^*(h)(v, v)$ is large when $K(x, x')$ is high (similar predictions), favors configurations that align with the eigenspace of M_K. Specifically, it might encourage compression (reduced variance) of representations along directions associated with large eigenvalues λ_i (high predictive similarity clusters), but primarily along components v where $g^*(h)(v, v)$ is small (less predictively sensitive directions) while preserving variance along components where $g^*(h)(v, v)$ is large, which are needed to distinguish dissimilar contexts (low $K(x, x')$).*

Argument. This hypothesis is more heuristic and draws parallels with spectral contrastive learning [9]. The operator M_K averages functions $\psi(h_{x'})$ weighted by similarity $K(x, x')$. Eigenfunctions ϕ_i with large λ_i represent stable patterns under this averaging, indicating clusters or directions of high predictive similarity. The NLL objective requires representations h_x to encode enough information for g_{head} to approximate $P_{\text{data}}(\cdot|x)$. Corollary 7.3 shows this necessitates separating h_x and $h_{x'}$ along sensitive directions (Def. 7.2) if $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$ are dissimilar (low $K(x, x')$). Conversely, if $K(x, x')$ is high, h_x and $h_{x'}$ can be close along sensitive directions.

Consider an eigenfunction ϕ_i with a large eigenvalue λ_i, corresponding to a set of contexts $\{x\}$ with high pairwise predictive similarity $K(x, x')$. NLL allows their representations $\{h_x\}$ to be close along sensitive dimensions. However, to minimize encoding cost or promote generalization (as argued in Section 5), the model might compress these representations along dimensions not needed by the head g_{head} (i.e., directions that are not sensitive according to Def. 7.2, lying in the null space of g^*). This aligns with the idea that contrastive/predictive coding collapses representations along dimensions capturing common information (high similarity, large λ_i) while preserving dimensions needed for downstream tasks (distinguishing dissimilar contexts, sensitive directions for g_{head}).

Invoking a compression principle (Section 5), the model might reduce representational variance for such high-similarity clusters. However, this compression must be selective. It should preserve the variance needed to encode the information for g_{head} (i.e., along directions v where $g^*(h)(v, v)$ is large enough to matter for prediction), especially if there are subtle but non-zero predictive differences within the cluster. Variance reduction might preferentially occur along directions v where $g^*(h)(v, v)$ is small (less predictively sensitive directions), as these components contribute less to the NLL loss.

While NLL does not explicitly optimize an objective related to M_K, the pressure to accurately predict while potentially minimizing representational complexity could lead to an implicit alignment where directions of high predictive similarity (captured by leading eigenfunctions of M_K) are compressed along dimensions irrelevant to the immediate prediction task. A rigorous proof would require formalizing this complexity/compression trade-off and connecting it to the spectrum of M_K during optimization. (Full argument in Appendix A.6.) □

In summary, while NLL optimization does not explicitly perform spectral clustering or eigenspace alignment, the core objective of matching conditional probabilities $P_{\text{data}}(\cdot|x)$ imposes geometric constraints on the representation space $\mathcal{H}$. These constraints push representations of predictively similar contexts closer together (along dimensions relevant to the predictor), mirroring the objectives of spectral methods applied to the graph defined by predictive similarity. This provides a more rigorous foundation for understanding why NLL training leads to structurally organized representations that capture predictive relationships.

8 Related Work

Our work applies the specific formalism of divergence-enriched Markov Categories [18, 17] to analyze the internal information processing of AR language models. This provides quantitative, information-theoretic metrics based on categorical entropy and mutual information.

Broader applications of category theory to understand machine learning and foundation models have also been explored. Notably, Yuan [26] uses general category theory, including concepts like functors, natural transformations, and the Yoneda Lemma, to establish theoretical limits on the capabilities of foundation models. Yuan's work focuses on the relationship between pretext tasks (defining a category), downstream tasks (viewed as functors), and the solvability of these tasks via prompt-tuning (linked to representable functors) or fine-tuning (linked to functor extension). It also addresses multimodal learning through functors between different categories (e.g., text and image).

While both approaches leverage category theory, our work differs in its focus and methodology. We concentrate on the probabilistic and information-geometric structure of the single-step AR generation process using the specialized tools of Markov Categories and divergence-based information measures. In contrast, Yuan [26] adopts a more abstract categorical perspective to analyze the *overall functional capabilities* of foundation models concerning downstream task solvability and cross-modal generalization, without delving into the specific probabilistic mechanics or internal information flow metrics that are central to our analysis. Our framework provides fine-grained tools to dissect the AR step, whereas Yuan's framework offers high-level theorems about what tasks a foundation model, viewed as embodying a category learned from a pretext task, can fundamentally achieve. The two perspectives are complementary, offering different levels of abstraction for understanding large generative models.

Other related theoretical directions include information bottleneck theory [22] which studies optimal compression under relevance constraints, information-theoretic generalization bounds [25], analyses connecting contrastive learning to spectral methods [9, 21], and studies of information geometry in statistical models [1]. Our work integrates ideas from these areas within the unifying language of Markov Categories.

9 Conclusion

In this work, we introduced a rigorous mathematical framework for analyzing the core autoregressive generation step in language models, leveraging the expressive power of Markov Categories (MCs). By explicitly modeling the context-to-prediction mapping as a composition of Markov kernels $k_{\text{gen},\theta} = k_{\text{head}} \circ k_{\text{bb}} \circ k_{\text{emb}}$ within the canonical category $\mathbf{Stoch}$, we provided a foundation for compositional analysis that respects the probabilistic nature of the process. The enrichment of $\mathbf{Stoch}$ with a statistical divergence D, and the subsequent use of categorical entropy $\mathcal{H}_D$ and mutual information I_D, enabled the development of novel metrics for quantifying representation fidelity (RepDiv_D), statistical dependencies ($I_D(H_t; W_t)$, $I_D(H_t; H_{t+1})$), prediction stochasticity ($\mathcal{H}_D(k_{\text{head}})$), and information flow bottlenecks ($I_D(S; H_t) \geq I_D(S; W_t)$).

Beyond providing these analytical tools, our framework offers deeper insights into the functioning of AR LMs trained via negative log-likelihood (NLL) minimization. We argued that the NLL objective is fundamentally linked to data compression (Section 5), forcing the model to not only predict accurately but also to capture the intrinsic stochasticity of the language generation process, as measured by categorical entropy. Furthermore, we explored the geometric consequences of this objective. The information geometry perspective (Section 6) revealed how the prediction task induces a specific functional geometry on the representation space $\mathcal{H}$ via the pullback Fisher-Rao metric g^*, characterizing the learned predictive sensitivities. Critically, we formalized the notion that NLL training performs implicit structure learning (Section 7). By minimizing the KL divergence between the model and data conditionals, NLL implicitly enforces geometric constraints, separating representations based on predictive dissimilarity. We established formal connections between this implicit pressure and the principles of spectral methods operating on graphs defined by predictive similarity, suggesting NLL sculpts representations aligned with the underlying predictive structure of the data.

This Markov Categorical approach provides a unified lens, integrating concepts from category theory, probability theory, information theory, and information geometry, to move beyond purely empirical or heuristic analyses of AR LMs. It offers formal language and quantitative tools for investigating information flow, representation structure, compression, and the mechanisms underlying the capabilities of these powerful models. While acknowledging limitations, particularly in the estimation of high-dimensional information quantities and the assumptions required for some theoretical results, this work lays the theoretical groundwork. Future research directions include refining the spectral connections, extending the analysis to multi-step dynamics, empirically validating the metrics, and exploring how these categorical insights might inform the design of more interpretable, controllable, or efficient language model architectures. Ultimately, this framework contributes to the development of a more principled theoretical understanding of deep generative models.

References

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

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

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 2018.

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

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

[13] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

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

[23] 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.

[24] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

[25] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in neural information processing systems*, volume 30, 2017.

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

A Full Proofs of Theorems and Arguments of Hypotheses

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

Let $p_x(\cdot) := k_{\text{data}}(x, \cdot)$ 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:

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

The average KL divergence is defined as:

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

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 $\mathcal{L}_{\text{KL}}(\theta) = 0$ is achieved if and only if the integrand is zero $p_{W_{<t}}$-almost everywhere. That is, $D_{\text{KL}}(p_x(\cdot)\,\|\,q_{x,\theta^*}(\cdot)) = 0$ for $p_{W_{<t}}$-almost every x. This occurs if and only if $p_x(\cdot) = q_{x,\theta^*}(\cdot)$ for $p_{W_{<t}}$-almost every x. In terms of kernels, this means $k_{\text{data}}(x, \cdot) = k_{\text{gen},\theta^*}(x, \cdot)$ for $p_{W_{<t}}$-almost every x.

If the model class $\{k_{\text{gen},\theta}\}$ contains k_{data}, say $k_{\text{data}} = k_{\text{gen},\theta_{\text{true}}}$, then choosing $\theta^* = \theta_{\text{true}}$ achieves $\mathcal{L}_{\text{KL}}(\theta^*) = 0$, which is the minimum possible value. □

A.2 Proof of Theorem 5.2 (Convergence of Average Categorical Entropy)

We want to show that $\lim_{n\to\infty} \bar{\mathcal{H}}_D(k_{\text{head},n}; p_{H_t,\theta_n}) = \bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*})$.

Recall the definition:

$$\bar{\mathcal{H}}_D(k_{\text{head},\theta}; p_{H_t,\theta}) = \mathbb{E}_{h\sim p_{H_t,\theta}}[\Psi_D(h, k_{\text{head},\theta}(h,\cdot))],$$

where $\Psi_D(h,p) := D_{\mathcal{V}\otimes\mathcal{V}}(\sum_{w\in\mathcal{V}} p(w)\delta_{(w,w)} \parallel p\otimes p)$, and $p = k_{\text{head},\theta}(h,\cdot)$.

Let X_n be the random variable $\Psi_D(H_n, k_{\text{head},n}(H_n,\cdot))$ where $H_n \sim p_{H_t,\theta_n}$. We want to show $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X^*]$, where $X^* = \Psi_D(H^*, k_{\text{head},\theta^*}(H^*,\cdot))$ with $H^* \sim p_{H_t,\theta^*}$.

We are given: (i) $k_{\text{head},n}(h,\cdot) \to k_{\text{head},\theta^*}(h,\cdot)$ in a suitable topology (e.g., total variation) for p_{H_t,θ^*}-almost every h. Let's denote this $p_n(h) \to p^*(h)$.

(ii) $p_{H_t,\theta_n} \Rightarrow p_{H_t,\theta^*}$ (weak convergence). This means $\int g(h)p_{H_t,\theta_n}(dh) \to \int g(h)p_{H_t,\theta^*}(dh)$ for all bounded continuous functions $g : \mathcal{H} \to \mathbb{R}$.

(iii) The function $\Psi_D(h,p)$ is continuous and bounded in p (with respect to the topology in (i)) for relevant h. Since $\mathcal{V}$ is finite, standard divergences like KL and TV are continuous functions of the probability vectors $p \in \Delta^{|\mathcal{V}|-1}$. The map $p \mapsto \sum p(w)\delta_{(w,w)}$ and $p \mapsto p\otimes p$ are also continuous. Thus, $p \mapsto \Psi_D(h,p)$ is continuous for fixed h. Boundedness also holds for typical divergences on finite spaces. Let M be an upper bound: $|\Psi_D(h,p)| \leq M$.

Let $\Phi_n(h) = \Psi_D(h, k_{\text{head},n}(h,\cdot))$ and $\Phi^*(h) = \Psi_D(h, k_{\text{head},\theta^*}(h,\cdot))$. +From (i) and the continuity part of (iii), we have $\Phi_n(h) \to \Phi^*(h)$ for p_{H_t,θ^*}-almost every h.

We want to show $\lim_{n\to\infty} \int \Phi_n(h)p_{H_t,\theta_n}(dh) = \int \Phi^*(h)p_{H_t,\theta^*}(dh)$.

Consider the difference:

$$\begin{aligned}
|\mathbb{E}[X_n] - \mathbb{E}[X^*]| &= \left|\int \Phi_n(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right| \\
&\leq \left|\int \Phi_n(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta_n}(dh)\right| \\
&+ \left|\int \Phi^*(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right| \\
&= \left|\int (\Phi_n(h) - \Phi^*(h))p_{H_t,\theta_n}(dh)\right| + \left|\int \Phi^*(h)p_{H_t,\theta_n}(dh) - \int \Phi^*(h)p_{H_t,\theta^*}(dh)\right|
\end{aligned}$$

The second term converges to 0 as $n\to\infty$ due to the weak convergence (ii), provided $\Phi^*(h)$ is bounded and continuous. While $\Phi^*(h)$ might not be continuous in h, if it is bounded and continuous p_{H_t,θ^*}-almost everywhere, weak convergence is often sufficient. Let's assume $\Phi^*(h)$ behaves well enough (e.g., is bounded and continuous almost everywhere w.r.t. the limiting measure p_{H_t,θ^*}) for $\int \Phi^*(h)p_{H_t,\theta_n}(dh) \to \int \Phi^*(h)p_{H_t,\theta^*}(dh)$. (This is sometimes known as the Generalized Continuous Mapping Theorem or Portmanteau Theorem).

For the first term, we have $\Phi_n(h) \to \Phi^*(h)$ for p_{H_t,θ^*}-almost every h. We also have the bound $|\Phi_n(h) - \Phi^*(h)| \leq |\Phi_n(h)| + |\Phi^*(h)| \leq 2M$ from the boundedness assumption (iii). We can use a variant of the Dominated Convergence Theorem adapted for converging measures. Since $p_{H_t,\theta_n} \Rightarrow$

p_{H_t,θ^*} and $\Phi_n \to \Phi^*$ pointwise a.e. (w.r.t. p_{H_t,θ^*}), and the sequence Φ_n is uniformly bounded, we can conclude that $\int(\Phi_n(h) - \Phi^*(h))p_{H_t,\theta_n}(dh) \to 0$. Combining these, we get $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X^*]$.

For the final part: if the model class is expressive such that $k_{\text{gen},\theta^*} = k_{\text{data}}$ (meaning $\mathcal{L}_{\text{KL}}(\theta^*) = 0$), then the model perfectly matches the data generating process almost everywhere. If we assume the data process can be similarly factorized $k_{\text{data}} = k_{\text{head, data}} \circ k_{\text{enc, data}}$, then matching $k_{\text{gen},\theta^*} = k_{\text{data}}$ implies that the components must match (up to potential identifiability issues, e.g., transformations between the encoder output and head input that cancel out). Under reasonable assumptions (e.g., the factorization is unique in the relevant sense), we would have $k_{\text{head},\theta^*} \approx k_{\text{head, data}}$ and the distribution induced by the encoder $k_{\text{bb}} \circ k_{\text{emb}}$ would approximate the distribution of the ”true” internal state feeding into $k_{\text{head, data}}$, i.e., $p_{H_t,\theta^*} \approx p_{H_t,\text{data}}$. Therefore, $\bar{\mathcal{H}}_D(k_{\text{head},\theta^*}; p_{H_t,\theta^*}) \approx \bar{\mathcal{H}}_D(k_{\text{head, data}}; p_{H_t,\text{data}})$. □

A.3 Proof of Theorem 7.1 (Output Distribution Approximation Constraint)

We are given the average KL divergence loss:

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

where $p_\theta(\cdot|x) = g_{\text{head}}(f_{\text{enc}}(x))$ and μ_{ctx} is the distribution over contexts. We are also given a metric d_{out} on $\mathcal{P}(\mathcal{V})$ satisfying a Pinsker-type inequality:

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

for some constants $k, C > 0$. Examples include Hellinger distance d_H ($k = 2, C = 1/2$) and Total Variation distance d_{TV} ($k = 2, C = 1$).

Let $p = P_{\text{data}}(\cdot|x)$ and $q = p_\theta(\cdot|x)$ for a specific context x. Applying the inequality yields:

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

Now, we take the expectation of both sides with respect to the context distribution $x \sim \mu_{ctx}$. Since expectation is linear and the inequality holds pointwise for each x, we get:

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

This establishes the first part of the theorem, Equation (33).

For the second part, consider any two contexts x, x'. Let $p_x^{\text{data}} = P_{\text{data}}(\cdot|x)$, $p_{x'}^{\text{data}} = P_{\text{data}}(\cdot|x')$, $p_x^\theta = p_\theta(\cdot|x)$, and $p_{x'}^\theta = p_\theta(\cdot|x')$. The triangle inequality for the metric d_{out} states:

$$d_{\text{out}}(A, C) \leq d_{\text{out}}(A, B) + d_{\text{out}}(B, C)$$

Applying this twice:

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

This is Equation (34). Let $\epsilon_x = d_{\text{out}}(p_x^{\text{data}}, p_x^{\theta})$ and $\epsilon_{x'} = d_{\text{out}}(p_{x'}^{\theta}, p_{x'}^{\text{data}})$. If the model fits the data well, $\mathcal{L}(\theta)$ is small. From Equation (33), $\mathbb{E}_{x\sim\mu_{ctx}}[\epsilon_x^k] \leq C\mathcal{L}(\theta)$, meaning the expected error (to the power k) is small. By Markov's inequality, for any $\delta > 0$,

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

Thus, $\mathbb{P}(\epsilon_x \geq \delta)$ is small if $\mathcal{L}(\theta)$ is small, implying that for a vast majority of contexts x drawn from μ_{ctx}, the individual error ϵ_x is small. Therefore, for typical pairs (x, x'), both ϵ_x and $\epsilon_{x'}$ are small.

Rearranging the triangle inequality gives:

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

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

When ϵ_x and $\epsilon_{x'}$ are small, these inequalities show that the distance between the model's output distributions, $d_{\text{out}}(p_x^{\theta}, p_{x'}^{\theta})$, must be close to the distance between the true data distributions, $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$. In particular, if $d_{\text{out}}(p_x^{\text{data}}, p_{x'}^{\text{data}})$ is large (predictively dissimilar contexts), then $d_{\text{out}}(p_x^{\theta}, p_{x'}^{\theta})$ must also be large, as the difference is bounded by small error terms. □

A.4 Proof of Corollary 7.3 (Implicit Representation Separation)

We assume the two conditions hold: (i) $\mathcal{L}(\theta)$ is sufficiently small such that for typical x, x', the errors $\epsilon_x = d_{\text{out}}(P_{\text{data}}(\cdot|x), p_\theta(\cdot|x))$ and $\epsilon_{x'} = d_{\text{out}}(p_\theta(\cdot|x'), P_{\text{data}}(\cdot|x'))$ are negligible compared to $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$. (ii) The head mapping $g_{\text{head}} : \mathcal{H} \to \mathcal{P}(\mathcal{V})$ is sufficiently sensitive: if $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) > \delta > 0$ for $h_x, h_{x'}$ in the populated region, then h_x and $h_{x'}$ must differ along directions sensitive to g_{head}. These sensitive directions span the subspace orthogonal to the null space of the Jacobian $J_{g_{\text{head}}}(h)$, which corresponds to the support of the pullback metric $g^*(h)$ (Section 6).

From the proof of Theorem 7.1, under assumption (i), we have:

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

Substitute $p_\theta(\cdot|y) = g_{\text{head}}(h_y)$ where $h_y = f_{\text{enc}}(y)$:

$$d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) \approx d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x')).$$

Now, consider the case where contexts x, x' are predictively dissimilar, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large. Specifically, assume it is significantly larger than the approximation error $(\epsilon_x + \epsilon_{x'})$ and also larger than the sensitivity threshold δ from assumption (ii). Then, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be large, and in particular, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'})) > \delta$.

By assumption (ii), if the distance between the outputs $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ exceeds the threshold δ, then the inputs h_x and $h_{x'}$ must differ along directions to which g_{head} is sensitive. Therefore, we conclude that if $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is large, then h_x and $h_{x'}$ must differ along the sensitive directions for g_{head} (as defined in Definition 7.2).

Conversely, consider the case where x, x' are predictively similar, i.e., $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small (e.g., close to zero). Then, $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ must also be small. This condition

($g_{\text{head}}(h_x)$ close to $g_{\text{head}}(h_{x'})$) can potentially be satisfied even if h_x and $h_{x'}$ differ significantly, provided their difference lies primarily within the null space of the Jacobian of g_{head} (directions insensitive to the head). However, the condition does not require h_x and $h_{x'}$ to be far apart along sensitive directions. In fact, mapping them closely together along sensitive dimensions is consistent with achieving a small $d_{\text{out}}(g_{\text{head}}(h_x), g_{\text{head}}(h_{x'}))$ and thus satisfying the NLL objective constraint in this case. Therefore, NLL optimization does not strongly constrain the distance between h_x and $h_{x'}$ along relevant dimensions when contexts are predictively similar. □

A.5 Proof of Proposition 7.6 (NLL Objective and Implicit Dirichlet Energy Minimization)

We are given a sensitive direction v (in the support of $g^*(h)$) and the projection $\phi_v(x) = \langle h_x, v\rangle$. The Dirichlet energy with respect to a predictive similarity kernel $K(x, x')$ is:

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

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

We assume $\mathcal{L}(\theta)$ is small, and the conditions of Corollary 7.3 hold.

Consider a pair of contexts (x, x') where the predictive similarity $K(x, x')$ is high. By Definition 7.4, high $K(x, x')$ implies that the true conditional distributions $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$ are similar, meaning $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small. From the converse part of Corollary 7.3, when $d_{\text{out}}(P_{\text{data}}(\cdot|x), P_{\text{data}}(\cdot|x'))$ is small, the NLL objective does not force h_x and $h_{x'}$ apart along sensitive directions v. In fact, to ensure that $p_\theta(\cdot|x) = g_{\text{head}}(h_x)$ is close to $p_\theta(\cdot|x') = g_{\text{head}}(h_{x'})$, which is required to approximate the small distance between $P_{\text{data}}(\cdot|x)$ and $P_{\text{data}}(\cdot|x')$, the representations h_x and $h_{x'}$ are encouraged to be close along these sensitive directions v. That is, if $K(x, x')$ is large, minimizing NLL encourages $\langle h_x - h_{x'}, v\rangle$ to be small for sensitive v.

Now examine the integral defining $\mathcal{E}_K(\phi_v)$. The integrand is $K(x, x')(\langle h_x - h_{x'}, v\rangle)^2$. This term makes a significant contribution only when $K(x, x')$ is large (otherwise the factor $K(x, x')$ makes it small) and $(\langle h_x - h_{x'}, v\rangle)^2$ is large. However, we just argued that minimizing NLL exerts pressure such that when $K(x, x')$ is large, the term $(\langle h_x - h_{x'}, v\rangle)^2$ tends to be small for sensitive directions v.

Therefore, NLL minimization actively discourages configurations where the integrand is large for the pairs (x, x') that contribute most due to high $K(x, x')$. This means the optimization process implicitly favors representations h_x such that the overall integral $\mathcal{E}_K(\phi_v)$ is small for projections ϕ_v onto directions v sensitive to the prediction head. While this is not a direct minimization of $\mathcal{E}_K(\phi_v)$, the pressure exerted by NLL aligns with reducing the terms that dominate the Dirichlet energy, thus implicitly favoring lower energy configurations along predictively relevant dimensions. □

A.6 Argument of Hypothesis 7.8 (NLL Objective and Alignment with Operator Eigenspace)

This argument remains more interpretative, formalizing the sketch. We assume the setup: encoder f_{enc}, head g_{head}, predictive similarity kernel K, similarity operator M_K (Equation (39)) with eigenfunctions $\{\phi_i\}$ and eigenvalues $\{\lambda_i\}$. Assume Corollary 7.3 holds.

The operator M_K acts on functions ψ defined on the representation space $\mathcal{H}$. Its eigenfunctions ϕ_i represent directions or patterns in $\mathcal{H}$ that are stable under averaging weighted by predictive similarity. A large eigenvalue λ_i signifies that the corresponding eigenfunction ϕ_i captures a dominant structure of predictive similarity: contexts x whose representations h_x have high values of $\phi_i(h_x)$ tend to be predictively similar to other contexts x' whose representations $h_{x'}$ also have high values of $\phi_i(h_{x'})$.

From Corollary 7.3, minimizing NLL requires:

- If $K(x, x')$ is low (dissimilar predictions), then h_x and $h_{x'}$ must differ along sensitive directions for g_{head}.
- If $K(x, x')$ is high (similar predictions), then h_x and $h_{x'}$ are allowed (and encouraged) to be close along sensitive directions for g_{head}.

Consider directions u in $\mathcal{H}$ that are strongly correlated with eigenfunctions ϕ_i having large eigenvalues λ_i. These directions capture clusters or variations associated with high predictive similarity. For contexts x, x' within such a cluster (high $K(x, x')$), NLL allows their representations $h_x, h_{x'}$ to be close along the sensitive components of u.

Now, invoke a principle of representational efficiency or compression (Section 5). A model minimizing prediction error (NLL) might also implicitly seek compact representations, discarding information not necessary for the immediate task. Dimensions in $\mathcal{H}$ that are insensitive to the head g_{head} (i.e., directions w in the null space of the Jacobian $J_{g_{\text{head}}}$, or where $g^*(h)(w, w) \approx 0$) carry information not used for the next-token prediction $p_\theta(\cdot|x)$. These directions are precisely those that are not sensitive according to Definition 7.2. As g^* is likely non-degenerate (Remark 6.4), the insensitive subspace is trivial. The relevant distinction is between directions v where $g^*(h)(v, v)$ is large (highly sensitive) vs small (weakly sensitive).

Consider the variation of representations $\{h_x\}$ along a direction u associated with a large eigenvalue λ_i. This variation reflects differences among contexts that are generally predictively similar. NLL requires $g_{\text{head}}(h_x)$ to be close for these contexts. This allows h_x to be close along highly sensitive directions (large $g^*(h)(v, v)$). An efficient model might compress representations by reducing variance along directions where variation is less critical for NLL. This could include directions v where $g^*(h)(v, v)$ is small (weakly sensitive directions), as variation along these directions has less impact on the output distribution $p_\theta(\cdot|x)$ and thus the NLL loss.

This leads to an implicit alignment: directions u associated with high predictive similarity (large λ_i) may exhibit reduced variance along components v where $g^*(h)(v, v)$ is small. Conversely, directions needed to distinguish predictively dissimilar contexts (low $K(x, x')$) must maintain variance along components v where $g^*(h)(v, v)$ is large, as this is necessary to separate their output distributions $g_{\text{head}}(h_x)$ and $g_{\text{head}}(h_{x'})$ as required by NLL.

This behavior mirrors the outcome of spectral contrastive learning, where representations are collapsed along directions of assumed similarity (analogous to large λ_i) while preserving discriminative information. While NLL optimization doesn't explicitly target the spectrum of M_K, the combined pressure of accurate prediction and potential representational compression leads to a structure where the geometry of $\mathcal{H}$ implicitly reflects the eigenspectrum of the predictive similarity operator, particularly in how variance is distributed between head-sensitive and head-insensitive dimensions. A fully rigorous proof connecting NLL optimization dynamics directly to the spectrum of M_K would require stronger assumptions about the optimization process and the model's implicit biases towards compression.